UCL
Université
catholique
de Louvain

Fluency and Disfluency

A Corpus Study of Non-native and Native Speaker (Dis)fluency Profiles

Amandine DUMONT

Dissertation submitted for the Degree of
Doctor of Philosophy and Letters
Université catholique de Louvain

Members of the jury

Prof. L. Isebaert (Université catholique de Louvain), president
Prof. G. Gilquin (Université catholique de Louvain), supervisor
Prof. S. Granger (Université catholique de Louvain), supervisor
Prof. Y. Bestgen (Université catholique de Louvain)
Prof. S. De Cock (Université catholique de Louvain)
Dr. S. Götz (Justus-Liebig-Universität Giessen)
Dr. T. Gráf (Univerzita Karlova)

2017-2018

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would not have been able to complete the long PhD journey on my own. I met many people, some of whom I've had the opportunity to learn to know better and become close friends with, and without them, the journey would have probably been a one-day trip only.

I've had the honour to be supervised by two Ladies who each have deep qualities and an impressive professional mind-set. Gaëtanelle has the eye for the small details, the textual ornaments, as well as a calm demeanour that could reassure me in times of need. Sylviane has a sufficient amount of enthusiasm to convince you that you could cross not only one, but two (or perhaps even three) mountains in one go. Never could I have had a supervisor as responsive, clear-sighted, and resourceful as her. To the two of you, a huge and sincere thank you for having nurtured in me the addiction to learner corpus research and for having trusted me with this project five years ago. I have been sincerely honoured.

I would like to take the opportunity to extend my sincere thanks to Prof. Yves Bestgen, Prof. Sylvie de Cock, and Dr. Sandra Götz (Justus-Liebig University), my committee members, as well as to my external examiner, Dr. Tomáš Gráf (Univerzita Karlova), for their strong support, the time spent discussing, and their expert advice on linguistic and statistical matters. Thank you also to Prof. Isebaert, the president of my jury.

During these five years, I was very happy to be surrounded by a team of exceptional people – colleagues, fellow PhD students, and dear friends. I would like to warmly thank all of them for their unfailing support and patience, their interest, as well as for taking my mind off my thesis at times. In alphabetical order: Caroline, Fanny, Julie, Magali, Maïté (and her funny anecdotes), Myriam (who supported me in many different ways), Natassia (and her earnestness), Samantha (who was my confident and advisor in various matters), Tove, and Verena. Thank you also to the ARC members and to all the others I did not cite, but who left an impression on me and contributed to make me who I am today.

Finally, I want to extend my thanks to my closest supporters, who believed in me and gave me unfailing support and continuous encouragements throughout the years. More particularly, I would like to give a big thank you to those who were always there for me, from the very beginning. Dad. Mum. My sister Céline. My dear family. And, of course, my beloved other half, without whose delicious meals I would definitely have run out of energy and motivation before the end of this journey.

Well, er… I'll .. you know, I'll never be able to say enough *thank you* er *thank you*'s to all of you!

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

ANOVA	Analysis of variance
BASE	British Academic Spoken English Corpus
BNC	British National Corpus
C(s)	Conjunctions (*and, so, but*)
CA	Cluster Analysis
CANCODE	Cambridge and Nottingham Corpus of Discourse in English
CEFR	Common European Framework of Reference
CHILDES	Child Language Data Exchange System
CI	Confidence interval
CIA	Contrastive Interlanguage Analysis
COBUILD	Collins Birmingham University International Language Database
COLT	Bergen Corpus Of London Teenage Language
CYLIL	Corpus of Young Learner Interlanguage
DM(s)	Discourse marker(s)
ESF	European Science Foundation Second Language Database
EVA	Evaluation of English in Norwegian schools corpus
FA	Factor Analysis
FP(s)	Filled pauses(s)
FS(s)	False start(s)
InterFra	Interlangue Française
L(s)	Lengthening(s)
L1	First language
L2	Second of foreign language
LCR	Learner Corpus Research
LCW	Learner Corpora around the World list
LEAD	Louvain EAP Dictionary
LeaP	Learning Prosody in a Foreign Language corpus
LINDSEI	Louvain International Database of Spoken English Interlanguage
LINDSEI-FR	The French component of the Louvain International Database of Spoken English Interlanguage
LINDSEI-FR+	The time aligned and annotated version of the French component of the Louvain International Database of Spoken English Interlanguage
LOCNEC	Louvain Corpus of Native English Conversation
LOCNEC+	The time aligned and annotated version of the Louvain Corpus of Native English Conversation
LONGDALE	Longitudinal Database of Learner English
MAELC	Multimedia Adult ESL Learner Corpus
MARSEC	Machine Readable Spoken English Corpus
MICASE	Michigan Corpus of Academic Spoken English
MLR(s)	Mean length of runs
MLUP	Mean length of unfilled pauses

NICT JLE	NICT Japanese Learner English corpus
NNS(s)	Non-native speaker(s)
NS(s)	Native speaker(s)
p	*p* value
PAROLE	Corpus PARallèle Oral en Langue Etrangère
PCA	Principal Components Analysis
phw	Per hundred words
PTR	Phonation time ratio
R1 / R2 / R3	Rater 1 / rater 2 / rater 3
Rep(s)	Repetition(s)
RLD	Reference Level Description
RPD	University of Toronto Romance Phonetics Database
RS(s)	Restart(s)
SCoSE	Saarbrücken Corpus of Spoken English
sd	Standard deviation
SLA	Second Language Acquisition
SPLLOC	Spanish Learner Language Oral Corpora
SR(s)	Speech rate(s)
SWB	Switchboard Corpus
SWECCL	Spoken and Written English Corpus of Chinese Learners
T(s)	Truncation(s)
UP(s)	Unfilled pause(s)
VOICE	Vienna-Oxford International Corpus of English
W(s)	Foreign word(s)

INTRODUCTION

> [H]esitation phenomena [...] provide good evidence that speaking is not a matter of regurgitating material already stored in the mind in linguistic form, but that it is a creative art, relating two media, thought and language, which are not isomorphic but require adjustments and readjustments to each other. A speaker does not follow a clear, well traveled path, but must find his way through territory not traversed before, where pauses, changes of direction, and retracing of steps are quite to be expected. The fundamental reason for hesitating is that speech production is an act of creation.
>
> (Chafe 1980:170)

TUNING IN

Spoken language

Speech is a truly mesmerising thing. It is probably one of the most important distinguishing features of the human species, and, as such, it has exerted, and still exerts, perennial fascination. Yet, there is no consensus on exactly why, when, how, or where spoken language has emerged – some theories of language posit its evolution contemporary with *Homo sapiens* or even earlier (Whishaw *et al.* 2010) –, but there have been many accounts pertaining to its origins in early civilisations and their mythologies. For example, the Judeo-Christian tradition attributes the origins of language to Adam being assigned by God to name the world's creatures. Similarly, in Norse, Greek, Indian, and Chinese mythology, speech is a gift from god(s).

In the history of humankind, the development of speech occurred well before the development of writing. At the level of each human's life too, speech is mastered years before other language skills such as writing are fully developed. However, despite this historical and acquisitional precedence, our current knowledge of speech is still fragmentary, which is arguably attributable to two main factors: first, the volatile nature of speech, and second, its depreciation due to an alleged non-polished aspect.

Speech is, by essence, non-visual, non-tangible, and elusive as it flies away as soon as it has been uttered. Writing, by contrast, is visual, tangible, and permanent. *Verba volant, scripta manent*. In other words, speech exists but in time (Carter & McCarthy 2006:193). Whilst the permanency of writing enabled its dissection and thorough analysis throughout the years, the volatility of speech made it much more difficult to examine, especially when it was not technically possible to faithfully capture and engrave speech onto some material. It is only quite recently that tools began to be developed that capture and represent the acoustics of

speech, which consequently gave impetus to many new research possibilities, for example, the analysis of prosody, pronunciation, the grammar of speech, the structure of discourse, or fluency and disfluency phenomena. Technological evolutions have thus made speech come to the fore, and enabled it to be analysed to an extent that was barely imaginable before.

The second factor explaining why we still have only a fragmentary knowledge of speech mechanisms is the fact that speech has long been **depreciated** as an object of research. For a long time, the general view has been that spoken language is structurally, syntactically and lexically simple, unsystematic and not representative of "true" language, contrary to writing, which is structurally elaborated, syntactically and lexically complex, abstract and formal[1] (Biber 1988). These differences mainly come from the fact that, compared to speaking, writing is normally not (or less) constrained by time, and writers have the possibility of pondering about the specific wording of their sentences, and to edit leisurely, without the final piece of writing containing traces of the writing process. Historically, thus, a number of researchers have regarded writing as a "purer" or "truer" form of language and speech as "degenerate and not worthy of study" (Biber 1988:5; see also Allwood, Nivre & Ahlsén 1990).

Homing in on the main topic of this thesis, one of the main differences between spoken and written language is that the former contains elements such as *eh, erm, I mean*, *you know,* restarts, or blanks. This phenomenon so typical of speech has commonly been referred to as *fluency* or *disfluency* and is the main object of study of this thesis. This study on fluency and disfluency takes two distinct types of speakers into its scope, viz. French-speaking learners of English and British English native speakers.

Fluency and disfluency

Spoken language is replete with so-called disfluencies, i.e. elements such as pauses, reformulations or repetitions. In fact, it is estimated that about six out of 100 words are affected by disfluencies in spoken language (Fox Tree 1995). Consider, for example, the following excerpt, which is the transcription of the spontaneous answer of a native speaker during an interview[2] (the disfluencies are shown in bold font).

0-1: A transcription of spontaneous speech (dots represent unfilled pauses, "=" truncations, and ":" lengthenings)

well y= y= you you know it is **.** time **. I= I= I** tell people **..** that **..** **uh ..** this has been a great run **..** I have loved this job **I I I'**m not going to pretend that there haven't been moments of great frustration but it is a singular privilege **..**

[1] Incidentally, although speech and writing are often assumed to stand in contrast, speech can, quite paradoxically, only be analysed through a *written* transcription.

[2] https://www.youtube.com/watch?v=xXH5agV7skw (last accessed 14/12/2018).

uh I think I'm **.** as good a president now as I've ever been because **you:** learn stuff over eight years **you've you've you've .. sort of ..** been around the track a bunch of times **erm ...** but I also **.** see now the wisdom of **.** the founders that **uh .** at a certain point **.** you have **to: .** let go **er .** for the democracy to work that there has to be fresh legs there have to be new people **uh ..** and you have to **uh ..** have the humility to recognize that **... you know** you're a citizen **and and and uh .** you go back to being a citizen **after uh after** this office **is uh . is** over **so:** we're just trying to run through the tape **and uh** the great thing is I've got an unbelievable team around me of **. uh** people **who: .** have done extraordinary work

Example 0-1 is actually a transcription of President Barack Obama answering a simple question about the end of his term of office. Although a trained and fluent orator used to dealing with and answering complex questions, Obama fills his utterances with many *uh*s, blanks, repetitions, truncations, lengthenings, *well*s and *you know*s.

Is Obama disfluent because his speech is interspersed with disfluencies? I believe that the answer is "no", and this example perfectly illustrates the great paradox in fluency research: all speakers produce disfluencies, but these disfluencies do not intrinsically make a speaker disfluent. In fact, increasingly more researchers adhere to the view that disfluencies have a **dual role** in speech. While disfluencies may at times be simple traces of the act of producing language in real time (i.e. they are symptomatic of cognitive load and may be used to gain time to think about what to say next or to retrieve a specific word), they may also be functional and meaningful cues for the listener. For example, the two unfilled pauses surrounding the word *time* in Example 0-1 should not be interpreted as indicators that Obama had difficulties finding the word *time*: they are arguably used purposefully to add emphasis to this word. Likewise, the abundance of rather long unfilled pauses throughout the excerpt might be interpreted as a way of accentuating the seriousness of the topic, or as a way of demonstrating judiciousness, rather than as the President thinking about what to say every other word. Similarly, it does not seem very likely that the two discourse markers (*well* and *you know*) at the very beginning of his answer are markers of "unclear thinking, lack of confidence, [or] inadequate social skills" (Crystal 1988:47): in this case, it seems more probable that they are used functionally to create an interpersonal relationship between the speaker and his audience (Erman 2001), or, potentially, to soften the upcoming words (Crystal 1988).

Granted that the presence of so-called disfluencies (e.g. pauses or discourse markers) does not necessarily imply disfluency, one might wonder why some speakers are indeed perceived as less fluent, or even as disfluent. Research indicates that temporal variables such as speech rate or length of runs are the primary factors affecting the perception of fluency, and that the over- or misuse of some fluency features may also be detrimental to our perception of a speaker's fluency. Caroline Kennedy's nearly compulsive use of the discourse marker *you know* (Example 0-2) has, for instance, largely been criticised in the press. Moreover, recent studies have revealed that speakers may also be differentiated based on different "fluency profiles" (or combinations of fluency characteristics), and that these profiles might be

differentially perceived by listeners. In native speech, several linguists (e.g. Liberman 2015a; 2017; Tian 2016) have commented on the fluency profiles of Donald Trump, Hillary Clinton and Barack Obama. Donald Trump's fluency, for example, features many false starts and repetitions, but a rapid rate of speech and very few (if any) filled pauses. By contrast, Hillary Clinton and Barack Obama are characterised by a rare use of repetitions and repairs, but a slower rate of speech and an abundance of filled pauses. In a blog post, Tian (*ibid*.) concludes that, while both Clinton and Trump produce about the same number of disfluencies, Clinton's preference for filled pauses and Trump's preference for abandoned utterances, repetitions and repairs have affected the way they were perceived: Clinton is perceived as more fluent because she plans her utterances during her numerous filled pauses, and Trump is generally perceived as less fluent because he lacks discourse coherence due to his manifold false starts and restarts.

0-2: Caroline Kennedy's (over-)use of 'you know'[3]

> so I think in many ways **you know** we want to have all kinds of different voices **you know** representing us and I think what I bring to it is **you know** my experience as a mother as a woman as a lawyer **you know** I've been an education activist for the last six years here and **you know** I've written seven books two on the Constitution two on American politics so obviously **you know** we have different strengths and weaknesses

In sum, the literature suggests that, far from being the ability to fill time with talk or to produce speech uninterrupted by pauses or disfluencies (Fillmore 1979), native fluency might rather be found in the skilful use of disfluencies, and in the delicate alchemy of the combinations of disfluencies.

Moving away from native speech (and presidents and politics) and turning to foreign language **learner** fluency and disfluency, the same observations apply as those described above. Like native speech, learner speech is replete with disfluencies. As in native speech, disfluencies may be used functionally. And as in native speech, several learner fluency profiles have been shown to coexist (Götz 2013a). However, a core difference between learner and native fluency is that the former is assumed – and has been shown – to be generally more disfluent (i.e. to contain more disfluencies) than the latter: L2 speech is produced at a considerably lower speed, and there is a higher incidence of all kinds of disfluencies (Cucchiarini, Strik & Boves 2000; Deschamps 1980). This gap between learner and native fluency has usually been ascribed to differences in L2 knowledge and processing, and, more specifically, to a deficient knowledge of the lexis, syntax, morphology, and/or phonology of the L2, as well as to limited attentional resources, and greater demands on self-monitoring (Bosker 2014; de Bot 1992; Guz 2015; Kormos 2006). Granted, the degree of language

[3] Transcription adapted from http://www.nytimes.com/2008/12/28/nyregion/28kennedytranscript.html (last accessed 15/02/2018).

mastery greatly affects learner fluency, with more proficient learners producing more fluent speech, especially in terms of temporal fluency measures (Freed 2000; Lennon 1990; Towell, Hawkins & Bazergui 1996). To add yet another piece to this complex puzzle, evidence is accumulating that many fluency characteristics in a speaker's L2 are, in fact, also related to the speaker's speech characteristics in his or her mother tongue. In other words, some aspects of L2 fluency are in fact attributable to each person's individual and idiosyncratic speech characteristics (Cox & Baker-Smemoe 2013; Derwing *et al.* 2009; Hincks 2010).

Before concluding this section, it is important to take a step backwards and to underline that fluency and disfluency lie at the crossroads between many varied disciplines. They have not only been investigated from the perspective of first and foreign language research, but also from the perspective of speech pathology, psychology, cognition, neurology, sociology, gender studies, and computational linguistics, to name but a few. The scope of this study is obviously – and unfortunately – limited. Suffice it to say that the buzzing activity around fluency and disfluency has brought forth a wealth of fascinating and very insightful studies (*cf.* e.g. Eklund 2004 for an overview).

A FIRST GLIMPSE INTO THE THESIS

Objectives and scope of the thesis

Starting from the assumption that fluency features have a dual function (hence my use of the terms "(dis)fluency" and "(dis)fluency features"), this thesis explores the (dis)fluency of French-speaking learners of English as compared to British English native speakers. The overarching aim of this thesis is to fine-tune the understanding of this complex phenomenon in learner and native speech. More specifically, the thesis is articulated around **four main objectives**.

At a **theoretical** level, I will delve into the various approaches to fluency and disfluency in first and foreign language, attempt to circumscribe the scope of these notions, operationalise them into a set of (dis)fluency features, and review what has been uncovered so far in the literature with respect to those features. The second goal is to provide a detailed **description** of fluency and disfluency in L1 and L2 spoken English in informal dialogic interviews in a contrastive interlanguage analysis perspective (Granger 1996, 2015). The focus will be on the description of individual (dis)fluency features, their interrelationships, and the individual variation between speakers. A related objective is to examine the link between learners' productive fluency and their perceived **fluency level** as assessed by the *Common European Framework of Reference for Languages* (CEFR, Council of Europe 2001). Part of this thesis will also be devoted to **methodological** issues in connection with the analysis of fluency and

disfluency in spoken corpora, in particular the transcription of spoken language, the time alignment of the spoken corpora (i.e. the mapping between the audio recordings and the transcriptions) and the corpus annotation of (dis)fluency phenomena.

In terms of possible applications and implications, it is hoped that the methodology and findings reported in this thesis will contribute some insights to the field of foreign language teaching, learning and assessment, to spoken corpus research, and, possibly, to spoken dialogue systems.

In the following section, I will present the research questions and briefly outline the data and methodology adopted for the research.

Research questions and a bird's eye view into the methodology

The thesis aims to answer four main research questions.

- **RQ 1: Learner vs. native speaker (dis)fluency**

  How can the speech of French-speaking learners of English be characterised in terms of (dis)fluency and how does it compare to British English native speakers' (dis)fluency?

- **RQ 2: (Dis)fluency profiles**

  What is the importance of idiolects in the measurement of learner and native (dis)fluency and what (dis)fluency profiles can be identified among French-speaking learners of English and native speakers of English?

- **RQ 3: (Dis)fluency dimensions**

  What is the nature of the relationship between (dis)fluency variables in learner and in native English speech?

- **RQ 4: Assessed CEFR fluency levels**

  How does the learners' assessed CEFR fluency level relate with empirical measurements of (dis)fluency features?

To answer these research questions, a methodology that combines corpus-driven and corpus-based methods is adopted that makes use of two spoken corpora. The French component of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; Gilquin, De Cock & Granger 2010) provides the data for the French-speaking learners of English. This component consists in 50 interviews of high intermediate to advanced French-speaking learners of English. The native counterpart of LINDSEI, the *Louvain Corpus of Native*

English Conversation (LOCNEC; De Cock 2004), also includes 50 interviews of British English native speakers and will provide the data for the native speakers.

The initial challenge in using corpora in (dis)fluency research is the (un)availability of accurate and reliable temporal data and frequency counts of the range of (dis)fluency phenomena. To obtain these data, the two corpora have been time aligned and annotated according to a specifically-designed (dis)fluency annotation scheme.

The learners' (dis)fluency measurements have also been related to their fluency level as assessed by the *Common European Framework of Reference*. These CEFR fluency levels were obtained from three professionally-trained native speaker raters on the basis of the CEFR grids and descriptors for spoken language skills.

Lastly, an innovative aspect of this thesis is the combination of different statistical techniques to better understand the complexities of the phenomena at hand, and to highlight similarities and discrepancies between learner and native speech.

Fluency and disfluency in a multimodal perspective

Before we delve into the structure of this thesis, some project-related acknowledgements need to be made. This thesis is inscribed in the frame of a large-scale Concerted Action Research project (ARC) entitled "Fluency and disfluency markers. A multimodal contrastive perspective" (2012-2017)[4], which involves the University of Louvain (UCL, Belgium) and the University of Namur (UNamur, Belgium). This Concerted Action Research project brings together a team of researchers who investigate fluency and disfluency markers in a multimodal contrastive perspective. In the project, two languages are studied (French and English) and four modalities are examined (spoken and sign language; native and learner language).

The project gave rise to four theses, namely Crible (2017a), Notarrigo (2017), Grosman (forthcoming), and the present one, as well as several publications, including Crible *et al.* (2015b; 2017) and Dumont (2017a). Moreover, several collaborative projects involving the standardisation of variables have been carried out, such as a categorisation of "situational features" (Appendix 9.2), which aims to ensure the comparability of different communicative

[4] https://uclouvain.be/fr/instituts-recherche/ilc/fluency-and-disfluency-markers-a-multimodal-contrastive-perspective.html (last accessed 09/03/2018).

situations across theses, and a general annotation framework to ensure the comparability of the main (dis)fluency features across theses[5].

THE ORGANISATION OF THIS STUDY

The main body of the thesis is structured around seven chapters, grouped into three parts, as shown in Table 0-1.

The first part of the thesis provides the **theoretical background** and consists in two chapters. **Chapter 1** sheds light on the various definitions and conceptualisations of fluency and disfluency in learner and native speaker research. It also delineates the operationalisation of (dis)fluency in the frame of this thesis and offers a review of what has been uncovered so far with respect to fourteen (dis)fluency features. **Chapter 2** then deals with the contributions of spoken corpora to L1/L2 (dis)fluency research. First, the range and specificities of spoken corpora are examined. Then, issues with the representation of speech are considered before concentrating on the use of spoken corpora for fluency assessment.

Part 1
Chapter 1 - Fluency and disfluency in non-native and native speech Chapter 2 - The contribution of spoken (learner) corpora to (dis)fluency analysis
Part 2
Chapter 3 - Methodology Chapter 4 - Aligning and annotating LINDSEI-FR and LOCNEC
Part 3
Chapter 5 - A quantitative sketch of learner and native speaker (dis)fluency Chapter 6 - A multivariate approach to learner and native speaker (dis)fluency Chapter 7 - Linking up learners' productive (dis)fluency, the CEFR fluency scale and assessed CEFR fluency ratings
Conclusion and prospects
Appendices

Table 0-1: Overview of the structure of the thesis

[5] This annotation framework is available on the ARC website (https://uclouvain.be/fr/instituts-recherche/ilc/towards-a-shared-multi-linear-annotation-scheme-corpus-design-and-annotation.html; last accessed 09/03/2018).

Two **methodological chapters** form the second part of this thesis. **Chapter 3** is devoted to a thorough description of the corpus data. It also includes an overview of the variables under analysis, and a brief description of the main statistical procedures. **Chapter 4** offers a detailed account of the alignment and annotation procedures developed for the French component of LINDSEI (LINDSEI-FR) and LOCNEC.

The third main part of the thesis includes three chapters. **Chapter 5** aims to provide a descriptive account of the fourteen (dis)fluency measures under investigation in learner vs. native speech, with special attention devoted to individual variation and illustration of the phenomena. **Chapter 6** delves into multifactorial analyses, and seeks to highlight the interrelationships between (dis)fluency measures. First, underlying dimensions of learner and native speaker fluency are uncovered. Second, individual L1 and L2 (dis)fluency profiles are delineated. Finally, the last chapter, **Chapter 7**, delves into the relationship between empirical measurements of learner (dis)fluency and their CEFR fluency level.

The thesis is rounded off by a general conclusion, which summarises and discusses the main findings.

Finally, for the reader's convenience, appendices are included at the end of this thesis, following the bibliography section. They provide additional material and some more detailed statistical results.

PART I

Chapter 1 FLUENCY AND DISFLUENCY IN NON-NATIVE AND NATIVE SPEECH

Footprints? you ask.
Well, I wonder whose those could be.

The Book Thief

For we all stumble in many ways.
And if anyone does not stumble in what he says,
he is a perfect man.

James 3:2

Fluency is a commonly used term in foreign language teaching and assessment, where it is frequently contrasted with accuracy. For the layperson, fluency is used as a synonym for (first or foreign language) oral proficiency. In language testing, fluency is one of the descriptors of oral performance. The veritable cornucopia of different fields and research angles within which fluency and disfluency have been the object of study has given rise to a kaleidoscope of definitions, which this introductory chapter seeks to review.

This chapter provides the **theoretical background** on the notions of fluency and disfluency. In Section 1.1, the object of study is briefly defined in a broad and cross-disciplinary perspective. Section 1.2 examines more closely the notions of learner and native speaker fluency and disfluency. Section 1.3 then discusses some aspects specific to learner vs. native fluency. Finally, the last part of the chapter (Section 1.4) outlines the **operationalisation** of L1 and L2 fluency and disfluency in the frame of this thesis, and provides a systematic review of findings on the fourteen (dis)fluency variables under investigation.

1.1 BUT *ER …* WHAT DO WE MEAN BY *FLUENCY*?

It seems to have become something of a routine to start a thesis by saying how difficult it is to define the precise object of study (and I will conform to this tradition): fluency is indeed hard to define. Part of this difficulty comes from the fact that fluency has bearings on an impressive number of fields (see e.g. Eklund 2004 for an overview). So, before delving into theoretical considerations, it is useful to delineate the object of this study from a larger perspective, especially by briefly considering (1) the modality and (2) the (non-)pathological aspect of disfluency.

Fluency is a far-reaching concept that relates not only to speech, but also to writing (e.g. Abdel Latif 2013; Chenoweth & Hayes 2001; Ellis & Yuan 2004; van Gelderen & Oostdam 2002; Oh 2006; Taylor 1947). **Written fluency**, which may be defined as the "efficient access to linguistic knowledge and retrieval of linguistic form" (Miller, Lindgren & Sullivan 2008:438), can, for example, be investigated through keystroke logging based on the rationale that it reveals traces of underlying cognitive processes. The focus of studies into written fluency often lies on pauses (their length, number, distribution, location etc.), writing speed, and the average length of strings of words between pauses, which are taken as indexes of cognitive effort. Additionally, it also often lies on revisions (their number, type, location etc.), which are seen as indicators of a discrepancy between the writer's intentions and the text produced (Leijten & Van Waes 2013:360–361). Although the questions of what constitutes fluency and disfluency in writing are truly fascinating, I will not delve deeper into those aspects here given that the focus of this thesis lies on **fluency in the oral modality**.

A second important element to delineate the object of this study can be found in the inseparable companion of *fluency*, namely *dysfluency* or *disfluency*. Whereas the former spelling is generally used in a clinical sense, that is, in connection with speech disorders where speech is characterised by "an abnormally high frequency and/or duration of stoppages in the forward flow of speech" (Peters & Guitar 1991), the latter (*disfluency*) usually refers to non-pathological and inherent elements in speech such as pauses or reformulations. An enormous amount of research has been devoted to stuttering and speech pathologies, especially in children. Summarising the recent research in the field of speech disorders is, unfortunately, out of the scope of this thesis given that its focus lies on "normal", **non-pathological disfluency**. Nonetheless, it is worth citing the work by Johnson and colleagues, who were the first to produce a full-fledged set of categories of pathological dysfluencies. What has become known as *Johnson's eight categories* has become used widely and became a standard for researchers from different disciplines. It includes the following: interjections of sounds, syllables, words or phrases[6]; part-word repetitions; word repetitions; phrase repetitions; revisions[7]; incomplete phrases; broken words; and prolonged sounds (Johnson 1961:3–4; see also Johnson *et al.* 1948; Johnson 1959). For a more in-depth overview of the research in this field, see e.g. Eklund (2004:55–77).

[6] This category includes sounds such as *uh* and *hmmm*, i.e. filled pauses.
[7] This category includes changes in content, or at the level of grammar and pronunciation.

1.2 PERSPECTIVES ON FLUENCY

Spoken, non-pathological fluency can be, and has been, discussed from various perspectives and with many different objectives. It has, for example, been approached from the angle of language acquisition, psycholinguistics, cognitive linguistics, testing and assessment, natural language processing, sociolinguistics or discourse analysis. It thus comes as no surprise that there is some degree of **fuzziness surrounding what is precisely meant by *fluency*** (*fluency variables*, *fluency measures*, *disfluency*, *disfluencies* etc.). In spite of this, what the definitions available from previous literature have in common is that:

- fluency not only includes an **objective and quantifiable dimension**, but also a **subjective** aspect, which has to do with the **impression made on the hearer** (fluency "in the ears of the beholder", as Freed (2000) beautifully phrases it);
- **disfluencies**, far from being the scoria of spoken performance, reveal something about the state of the speaker's cognitive processes, and are **functional and useful to the listener**.

The ensuing summary is a synoptic overview of the main definitions of fluency. It is intended to provide a sense of how the notions of fluency and disfluency have been conceptualised from a historical perspective and to point out similarities across definitions.

1.2.1 Language in motion

As rightly pointed out by Koponen and Riggenbach (2000), a powerful conceptual metaphor underlies the meaning of *fluency*, not only in English, but also in other languages. In English, the term *fluency* is used; in French, it is *fluidité*; in Dutch, *vlotheid*; in German, *Flüssigkeit*; in Spanish, *fluidez* etc. All these terms suggest the idea of **language in motion**, and this is perhaps the common thread underlying most of the scientific descriptions of fluency.

1.2.2 The pioneers – or how it all started with pausing

In the 1950s, studies of speech phenomena took off and what had previously been considered as trivial performance aspects, especially pausing, began to be seen as objects of study in their own right.

The **pioneering work by Goldman-Eisler** (1954a; 1954b; 1956; 1958a; 1958b; 1961a) on pauses and speech rate set the stage for subsequent research on fluency and disfluency. For example, she demonstrated that short and long pauses "tend to be constant within limits and

characteristic of individuals independent of changing partners and topics" (Goldman-Eisler 1951:355). Another of her main findings is that what **listeners perceive as changes in the rate of speech is primarily a function of changes in pausing** by the speaker, and not in time spent articulating. She writes that (Goldman-Eisler 1961b:171):

> The speed of the actual articulation movements producing speech sounds occupies a very small range of variation [...] while the range of pause time in relation to speech time was five times that of the rate of articulation.

Goldman-Eisler also explored differences in speech rhythm between fluent and less fluent stretches, and pointed to the **central role played by cognitive mechanisms**, organised into a dynamic system, **in shaping temporal fluency**.

Following the pioneering work by Goldman-Eisler, several researchers set out to investigate pauses and pausing (e.g. Boomer & Dittmann 1962; Henderson, Goldman-Eisler & Skarbek 1966; Lounsbury 1969; Hawkins 1971). Taking a slightly wider perspective, Maclay and Osgood (1959) quantitatively investigated four hesitation phenomena, namely filled and unfilled pauses, repeats, and false starts. Among their main findings is the observation that, not only are there consistent differences between speakers, but also that **speakers seem to have a relative preference for hesitation phenomena of different types**, with, for example, some speakers being characterised by a relatively large number of filled pauses and repeats, while others might show more unfilled pauses and false starts in their speech. They conclude their article by claiming that, despite the fact that "hesitations" occur non-randomly in speech, they are but "auxiliary events" that are not on the same footing as the "raw data" and only "help to identify and circumscribe linguistic units" (*ibid*.:39).

Building both on previous pausological work and on studies that considered a larger panel of "hesitation phenomena", Grosjean (1972; 1980a; 1980b; 1980c) suggested a two-tier categorisation of fluency variables. He differentiated *primary variables* of fluency from *secondary variables*:

- **primary variables**, which are always present in language output, include the rate of speech, the phonation time ratio, the mean length of runs, the number of unfilled pauses, the duration of unfilled pauses per minute, the articulation rate, and the mean length of unfilled pauses. Primary variables correspond to the **temporal aspect** of fluency;

- **secondary variables** are related to **hesitation phenomena**; their presence is not required. They include the frequency of filled pauses and of "disfluencies" (i.e. drawls, repeats and false starts).

Note that Grosjean (1980a; 1980c; Grosjean & Deschamps 1975) was also among the first researchers who adopted a **cross-linguistic** approach to fluency: he compared and contrasted fluency variables in French, English, and sign language.

1.2.3 Fillmore – a four-headed chimaera

Unlike many researchers of his time, Fillmore (1979; 2000) approached the conceptualisation of fluency as a complex set of skills attained by "the maximally gifted wielder of language" (Fillmore 1979:93), that is, an idealised native speaker. His conceptualisation of fluency goes far beyond the idea of pausing and rhythm to include semantics, appropriateness, and creativity. He distinguished **four different types of fluency** that a maximally fluent speaker has (*ibid*.:93):

1. the ability to **fill time with talk**, that is, the ability to talk at length with few pauses[8];
2. the ability to **talk in coherent, reasoned and "semantically dense" sentences**, that is, the ability to package the message into "semantically dense" sentences without too many "semantically empty" material;
3. the ability to **have appropriate things to say in a wide range of contexts**, that is, the ability to meet the communicative demands of different contexts and situations;
4. the ability to **be creative and imaginative in language use**, that is, the ability to express ideas creatively, for example by using humour or metaphors.

Note, incidentally, that the importance of the third type of fluency has also been emphasised by Meisel (1987) and Sajavaara (1987), who argue that fluency indeed refers to the communicative acceptability of the speech act, in other words, its fit according to what is appropriate in a specific communicative context.

Fillmore's four-tier definition of fluency is very extensive, but some aspects seem very **difficult to operationalise** (e.g. "appropriateness", "creativity", "semantically dense sentences") and it is unclear how this conceptualisation differs from global oral proficiency. Moreover, it is not very clear whether even "maximally gifted" speakers could demonstrate these four abilities together.

1.2.4 The aetiology of disfluency – disfluency as a communicative tool

Many of the approaches and definitions above make the more or less tacit assumption that disfluencies are flaws, or noise, in the speech signal, in other words, evidence of problems in the linguistic production that present obstacles to comprehension (Brennan & Schober 2001:275). An alternative way to view disfluency phenomena that slowly came to the

[8] Fillmore does not specify whether he refers to filled or unfilled pauses, or both.

foreground is to regard them as **containing meaningful information** for the communicative activity.

Starting from the 1980s, and drawing from speech act theory (Austin 1965; Searle 1965), fluency and disfluency started to be regarded from the perspective of communication strategies[9]. Initially, communication strategies were intimately related to problem-solving activities and disfluency was seen as an important cue for the participants in an interaction as to the relationship between the speaker and his/her utterance. Good and Butterworth (1980), for example, demonstrated that, whilst disfluency is an indicator of cognitive load for the speaker, speakers may also use hesitations in their speech to achieve some interactional goal such as signalling to the listener that they are experiencing production difficulties. In the same vein, Clark and Wasow (1998) pointed out that disfluency can be either seen as the **outcome of a process** that cannot be controlled by the speaker[10], or as the **result of strategies *under the control of the speaker***. With respect to the latter view, they argued that speakers, when confronted to a problem, deploy disfluencies in a strategic manner to signal ongoing difficulty in producing the utterance. For example, speakers tend to choose filled pauses when they expect a longer delay, and an unfilled pause when they expect only a brief interruption. Also, within the category of filled pauses, speakers seem to draw a distinction between the non-nasal *uh* and the nasal *um*: the former is used to **signal short delays**, and the latter to signal long delays (Clark & Fox Tree 2002). Besides, further studies found evidence that disfluencies allow listeners to **predict the likely upcoming word(s)** (Arnold, Fagnano & Tanenhaus 2003; Arnold *et al.* 2004; Arnold *et al.* 2007; Lowder & Ferreira 2016). An alternative account also suggests that **disfluencies heighten listeners' attention to upcoming speech** (Fox Tree 2001).

To sum up, bearing in mind that "it is hard to determine the reason that a speaker is disfluent, especially if the investigation is carried out after the fact from a corpus of recorded speech" (Corley & Stewart 2008:595), evidence suggests that disfluencies, and filled and unfilled pauses more particularly, might in fact contribute to a smoother understanding on the part of the listener because they are used both as symptoms of heavy cognitive processes, and as functional signals, both of which may be decoded and usefully interpreted by the listener (Clark & Wasow 1998).

---

[9] Communication strategies can briefly be defined as "strategies which a language user employs in order to achieve his intended meaning on becoming aware of problems arising during the planning phase of an utterance due to his own linguistic shortcomings" (Poulisse, Bongaerts & Kellerman 1984:72; in Kasper & Kellerman 1997:2). Tarone (1983) further emphasised that the interactional function of communication strategies must not be overlooked, which set the foundations to the view of co-constructed fluency, or "confluence" (Götz 2013a; Molenda & Pęzik 2014), that is, the view that the conversational output is the result of the joint contribution of the speaker and the listener.

[10] This corresponds to what Clark and Fox Tree (2002:75) call the "filler-as-symptom" view.

1.2.5 Proficiency or native-like rapidity

In his famous article, Lennon (1990) attempted to clarify the uses of the term *fluency* in EFL contexts. He distinguished two main senses. Fluency in the "**broad sense**" is used as a cover term for **oral proficiency**: it equals the spoken command of a foreign language, for example in statements such as 'She is fluent in Spanish', and refers to the extent of grammatical accuracy, vocabulary range, and production skills. Fillmore's (1979) conceptualisation is one of the examples of fluency in the broad sense.

In its **narrow sense**, fluency is used to refer to one, presumably isolatable, component of oral proficiency, i.e. **native-like rapidity** (*cf*. Grosjean's "primary variables"). This interpretation of fluency is often encountered in speaking tests and oral examinations, where the "flow" or "smoothness" is assessed in a specific rubric. As underlined by Chambers (1997:538), a definition restricting fluency in spoken production to temporal variables "provides a useful anchorage for a concept which is prone to vagueness and multiple interpretations", even if "foreign language teaching results in nativelike fluency in exceptional cases only" (Sajavaara 1987:45).

However, Lennon also underlines that, even in the narrow sense, the concept of fluency is often extended to cover other elements of oral proficiency. The assumption is that **fluent delivery in performance might be the overriding determiner of perceived oral proficiency** and that other features of spoken proficiency (such as accuracy or pronunciation) easily become subsumed under fluency because a fluent delivery in performance directs the listener's attention away from deficiencies in other areas. The proponents of this view claim that native-like performance (i.e. fluency in the broad sense) is not the final goal of foreign language teaching, as fluency refers to "natural language use" resulting from the "maximally effective operation of the language system so far acquired by the student" (Brumfit 1984:56–57) or the ability to engage in "successful communication" (Kennedy & Trofimovich 2008:460) in the target language. Going a step further, Denke (2009) argues that speakers can be evaluated as fluent when they can **cope with stumbles satisfactorily**, which is in line with Sajavaara's (1987:62) conclusion that "[t]he 'good' speaker 'knows' how to hesitate, how to be silent, how to self-correct, how to interrupt, and how to complete expressions or leave them unfinished".

Moving back to Lennon's twin conceptualisation of fluency, the author stresses that fluency differs from the other elements of oral proficiency in one important respect: fluency is **purely a performance phenomenon** for which there is no fluency store. According to him, disfluencies make the listener aware of the production process under strain and, consequently, fluency can be defined as "an impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently" and it also "reflects the **speaker's ability to focus the listener's attention on**

his or her message by presenting a finished product rather than inviting the listener to focus on the working of the production mechanisms" (Lennon 1990:391 my emphasis).

1.2.6 The age of classifications and frameworks – towards an operationalisable multidimensional construct

Two decades later, Segalowitz (2010), in his monograph, distinguished three facets of Lennon's (1990) narrow definition of fluency, namely cognitive fluency, utterance fluency, and perceived fluency. These three facets are briefly considered below.

A. Cognitive fluency

Cognitive fluency refers to the efficiency of operation of the underlying cognitive processes responsible for the production of utterances. It is the ease of mental preparation.

Segalowitz adopted the **model of speech production by Levelt** (1989), which is a **blueprint of the monolingual speaker**. According to this model (see Figure 1-1), speakers plan their utterances in three consecutive stages. In the first phase, the *Conceptualiser*, which also contains sociopragmatic knowledge, creates a preverbal or pre-linguistic message, which is fed into the *Formulator*. This module, which includes a submodule for grammatical encoding and another for phonological encoding, selects the suitable linguistic verbal and prosodic elements to provide the message with an appropriate morpho-phonological form. This linguistic form is then catered for by the *Articulator*, where the phonetic plan is used to produce the actual message (i.e. overt speech).

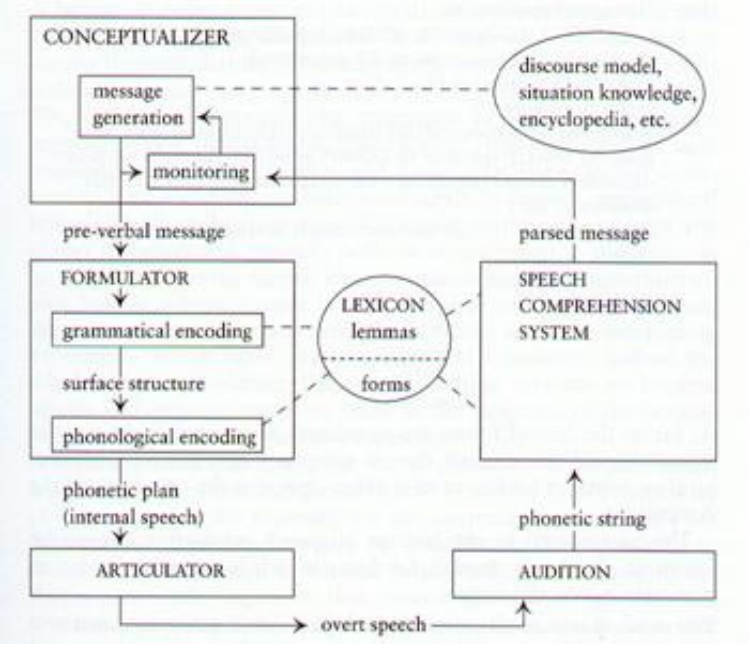


Figure 1-1: Levelt's blueprint of the speaker (Levelt 1989:9)

Besides those three modules, Levelt's model of speech production also includes the *Monitor*, which supervises the overall production process, and detects and repairs potential errors. Although the Monitor is situated in the Conceptualiser, which implies that generation and control are tightly related, it also keeps track of the message throughout its production. In other words, it can change the message not only at the stage of conceptualisation, but also after having made use of the speech comprehension system (i.e. the system that speakers use to understand the speech of others, but also their own).

Segalowitz argues that the different phases – conceptualiser, formulation, and articulation – are three "fluency vulnerability points" that may lead to disfluencies. Disfluency can, for example arise in the conceptualiser when speakers have trouble finding out what to say, in the formulator when they have trouble choosing the correct words, or in the articulator, when they have trouble articulating the phonetic plan. The cognitive processes involved in the conception, formulation, and, to some extent, articulation, are then revealed in the fluency of the utterance (see next sub-section).

Levelt's blueprint of the monolingual speaker has later been adapted by de Bot (1992) and Kormos (2006) for **speech production in an L2**. According to De Bot, language production in an L2 is, overall, very similar to the production of spoken language in an L1 because it also involves the conceptualisation, formulation, and articulation of a message. Hhe assumes that some processes are not language-specific (e.g. the elaboration of the propositional content of the message, i.e. the macroplanning), which, Segalowitz argues, indicates that no L2-specific disfluency can arise at this stage. Microplanning level, where a specific information structure is assigned to the macroplan, is, however, presumed to be language-specific, just like the other stages of speech production. L2-specific disfluency may thus come from the formulator or the articulator, in other words, from an **incomplete lexico-grammatical knowledge of the target language**, or from **insufficient skills with which L2 knowledge is used** (lexical access, articulation etc.) (Bosker 2014:5–6). It is consequently at the stages of formulation and articulation that L2 speech is more vulnerable to disfluency.

B. Utterance fluency

The second facet of fluency in Segalowitz's conceptualisation, **utterance fluency**, consists in the actual temporal, pausing, hesitation, and repair characteristics of utterances that reflect the speakers' cognitive fluency and that can be acoustically and tangibly measured (they are not just impressions a listener might have).

A number of measurements may be associated with this interpretation of fluency, such as speech rate, number of reformulations, of filled and unfilled pauses etc. Skehan (1997; 2003; 2009; Tavakoli & Skehan 2005a; 2005b) suggested, however, that **utterance fluency is itself multidimensional**, and that the cognitive effort invested in speech production affects different aspects of oral performance. On this ground, he made a three-way distinction

between sub-dimensions of utterance fluency, namely *speed fluency*, *breakdown fluency*, and *repair fluency*.

- **Speed fluency** refers to the length and density of linguistic units; it is characterised as the rate of speech delivery;
- **Breakdown fluency** is tapped by measures such as number of pauses and amount of silence;
- **Repair fluency** relates to "the extent to which speed is repeated, reformulated, or left incomplete" (Witton-Davies 2010:119). As this quote makes clear, repair fluency relates to three main phenomena: restarts, false starts, and repetitions.

From a psycholinguistic point of view, while speed fluency is reliant on procedures of storage and recall of linguistic information from memory systems, breakdown and repair fluency are related to "the extent to which the learner is confident that what has been stored is reliable and the extent to which the learner has also created procedures which can be brought into operation to repair the situation when communication breakdown occurs, for whatever reason" (Towell 2002:55–56).

It is worth noting that, in the literature on first and second/foreign language fluency, the most commonly used measures are the temporal variables and the measures related to pauses, in other words, speed and breakdown fluency. Slightly less work has been carried out on the variables associated to **repair fluency**. This is probably related to the fact that early work on fluency tended to use laboratory speech, where speed and breakdown fluency measures could quite easily be analysed, but which is, per definition, less prone to repairs. Only with the availability of transcriptions of more naturally occurring speech could they start being analysed. It is also quite striking that the terminology of repair phenomena displays tremendous variability across studies. A major issue relates to the fact that many studies have confounded (or conflated) restarts, repetitions and false starts (and/or other (dis)fluency features) into a single category. For example, Riggenbach (1991:427), following Maclay and Osgood (1959), subsumes actual restarts, repetitions *and* false starts under the term "(retraced/unretraced) restarts".

Although **Skehan's three-tier typology** is larger in scope than most typologies so far, one of its **drawbacks** is that it does not cover all the (dis)fluency features that are studied in the fluency literature. In particular, truncated words, foreign words, vowel lengthenings and discourse markers, which have all been shown to contribute to fluency, are not included in this typology. Also, the classification of fluency measures under either speed, breakdown or repair fluency is not always very straightforward: the measure for speech rate (calculated as the number of syllables or words per minute, including pausing time) is, for example, dependent on the duration of unfilled pauses and articulation rate. It could thus arguably be seen as a measure of both speed and breakdown fluency.

C. Perceived fluency

The third interpretation of fluency by Segalowitz is **perceived fluency**. Perceived fluency is seen as the inferences listeners make about speakers' cognitive fluency based on their perceptions of their utterance fluency. It is the listeners' judgement made about speakers based on impressions drawn from utterance fluency (Segalowitz 2010:47–48). Perceived fluency is most commonly assessed with subjective judgements and rating scales, but, as stressed by De Jong *et al*. (2012a:896)

> From a methodological perspective, fluency as perceived by listeners, or raters, is dependent on the instructions that the raters receive, and on the definitions and notions the listeners or raters have of the construct of fluency prior to the rating instructions.

The three-way distinction between cognitive, utterance, and perceived fluency has become fairly influential, and has given researchers the impetus to probe into the relationships between the three facets (see De Jong *et al.* 2012a for a review). Research into the relation between **utterance fluency and cognitive fluency** aims to determine which aspects of utterance fluency are **indicators of ease and efficiency of cognitive processes from the speaker's point of view**. De Jong *et al*. (2012a), for example, looked at the link between cognitive fluency and utterance fluency. They found that, although the duration of unfilled pauses is but weakly related to learners' cognitive fluency, L2 speech rates are indeed strongly related with underlying cognitive processes. Research into the relationship between **utterance fluency and perceived fluency** aims to investigate what constitutes **fluency from the listener's point of view** and to assess the **relative contributions of different fluency measures** to the perception of fluency. This field of research is enjoying growing interest from L2 researchers, notably Bosker *et al*. (2013), Rose (2015) or Götz (2013a). These studies generally found that temporal measures can account for a large proportion of the variance in perceived fluency (but see Section 2.4.2 for a more detailed account of the main findings).

1.2.7 A holistic view – productive, perceptive, and non-verbal fluency

Like Segalowitz, Götz (2013a) offers a **holistic, triadic, view of fluency** in the narrow sense, which includes:

- **productive fluency**, which is performance-based, and consists in a combination of temporal variables, formulaic sequences and so-called fluency enhancement strategies (i.e. repeats, filled pauses etc.);
- **perceptive fluency**, which is concerned with the effect speech has on the listener, and includes dimensions such as accuracy, idiomaticity, intonation, accent, pragmatic features, lexical diversity, register, and sentence structure.

These two categories can arguably be largely paralleled with Segalowitz's utterance and perceived fluency. However, Götz goes a step further than previous characterisations of productive (utterance) fluency by also considering formulaic sequences (3- and 4-grams) as well as discourse markers and "smallwords"[11].

On the premises that it is not sufficient to only consider verbal aspects of fluency when fluency is seen as a holistic phenomenon and that "nonverbal acts are a significant aspect of conversational fluency" (Bavelas 2000:90), Götz also, and quite unconventionally, included a third category of fluency, namely:

- **non-verbal fluency**, which has to do with the use of paralinguistic features such as gestures, facial expressions, body language, looks, and emblems.

Although she did not investigate non-verbal fluency *per se*, Götz was the first to recognise the importance of the non-verbal aspect of communication by integrating it in a typology of fluency, on the same footing with other, long-standing, categories.

One of Götz's main findings in analyzing productive fluency is the existence of different "**fluency groups**" (or "fluency profiles") in the speech of both German-speaking learners of English and native speakers of English. Based on a cluster analysis of eight productive fluency variables, namely speech rate, mean length of runs, the frequency of unfilled pauses within clauses, the frequency of filled pauses within clauses, of 3-grams, of repetitions, of discourse markers, and of smallwords, she found **three fluency groups in the native speaker data**:

- The first group of speakers are characterised by an extremely fast speech rate, a high proportion of unfilled pauses and an average mean length of runs, an average use of formulaic language and of filled pauses, but a high rate of discourse markers, smallwords, and repetitions.

- The second profile corresponds to speakers who speak slowly, pause a lot, and have short runs of speech. Speakers belonging to this profile also show a low proportion of formulaic language, discourse markers, and smallwords. By contrast, they use filled pauses and repetitions very frequently, which, the author argues, seems to be this profile's main characteristic.

- The third profile includes a fast speech rate, long speech runs, few unfilled pauses, an extensive use of formulaic language, and few filled pauses and smallwords.

[11] Smallwords are defined by Hasselgren (2002:150) as "small words and phrases, occurring with high frequency in the spoken language, that help to keep our speech flowing, yet do not contribute essentially to the message itself". In her study, she includes markers like *sort of, kind of, and stuff* etc.

Likewise, she found three **fluency groups in the learner data**:

- The first profile corresponds to the temporally most fluent learners. The profile also includes a low use of filled pauses and repetitions, and an average performance with regard to discourse markers and smallwords.
- The second and the third group have an average temporal fluency. While the second group is characterised by a high use of formulaic language and low use of filled pauses, discourse markers, smallwords, and repetitions, the reverse is true for the third group.

1.2.8 A computational perspective – the art of grooming

This section on diverse perspectives on fluency would not be complete without a few words on computational approaches to fluency.

Extensive work in this domain has been carried out, or at least initiated, by Elizabeth Shriberg (esp. 1994; 2001). She noted that the presence of **disfluencies** in spontaneous speech was an **issue for the field of natural language processing**, for example because most speech recognition models at the time were trained on highly constrained data, or because the accuracy of part-of-speech taggers was greatly affected by the "ungroomed" (disfluent) aspect of spontaneous speech. To date, the problem is probably also acutely relevant for speech recognition systems.

To improve the automatic processing of speech, a number of studies set out to develop systems for the **automatic detection of disfluencies** with a view to eliminating them and obtain speech **samples cleaned of disfluencies**. Early work was carried out by O'Shaughnessy (1992; 1993) on the detection of "false starts", which he defines as the initial utterance in a reformulation. Then, extensive work was conducted by Shriberg (Bear *et al.* 1993; Eklund & Shriberg 1998; Liu, Shriberg & Stolcke 2003; Shriberg 1994; 2001; Shriberg, Bates & Stolcke 1997). One of her main contributions is the characterisation of the common underlying structure of disfluency phenomena as comprising three "regions": the *reparandum*, which is separated from the *interregnum* by an *interruption point*, and then followed by the *repair* (see Section 2.3.3 for more details).

Against this backdrop, the next section goes on to clarify some differences between fluency and disfluency in native and in learner speech.

1.3 NATIVE AND NON-NATIVE FLUENCY

1.3.1 Fluency in a native language

Many researchers have underlined the striking paradox of native speech: while native speakers are perceived as **fluent by default, their speech is nevertheless interspersed with disfluencies** (Bosker *et al.* 2014; Davies 2003; Raupach 1983; Riggenbach 1991). Apparently fluent L1 speech is, in fact, anything but fluent in an idealised sense because it abounds in disfluencies – in this regard, it is noteworthy that Fox Tree (1995) estimated that **six in every 100 words are affected by disfluency**. Arguably, these disfluencies may not disturb the overall impression of fluency, because "people around the world fill pauses in their own languages as naturally as watermelons have seeds" (Erard 2004), i.e. most disfluencies are not noticeable/noticed, and only a detailed transcription may reveal their presence.

Various factors have been shown to **influence native speaker fluency**, such as the gender of the speaker (Lickley 1994), his or her age, the setting and topic (Bortfeld *et al.* 2001), anxiety and alcohol consumption (Christenfeld & Creager 1996)[12], the speaker's psychological state (Friedman 1991a; 1991b), the academic discipline (Schachter *et al.* 1991)[13] etc. (see also Eklund 2004). Furthermore, most researchers agree that the realisation of (at least some) disfluency phenomena in native speech is **language-specific**: Shriberg (1994) suggested that there are dialectal differences between British and American English in the use of nasal and non-nasal filled pauses, and Grosjean and Deschamps (1975) observed that French L1 speakers pause less but longer than English native speakers (see also Riazantseva 2001) and de Leeuw (2007) provided evidence that there are cross-linguistic differences between English, German, and Dutch. In addition, native speakers have also been shown to have **highly variable hesitation patterns** (Götz 2011; de Leeuw 2007). All these results suggest that **it is misleading to assume that native speakers are a homogenous group whose fluency is uniform and consistently high**.

[12] In the study, anxiety was found to increase the number of filled pauses, and alcohol consumption was shown to reduce their frequency. This was related to the speaker's differing degree of attention to his/her speech in the two conditions: while anxiety makes speakers more aware of their speech, alcohol makes speakers "care less about what they say" (Christenfeld & Creager 1996:451).

[13] The authors tallied filled pauses among professors giving lectures: they found substantial differences between disciplines, with the humanities professors saying *uh* 4.85 times per minute in their lectures, social scientists 3.84 and natural science professors 1.39 times. They explore various likely explanations to account for these results, including the fact that lectures in the humanities are linguistically more complex and the fact that disciplines may attract "very different sorts of people". In this respect, they hypothesise that "[s]cientists may be people of steel who know and can firmly speak their minds; humanists may be ditherers" (Schachter et al. 1991:364).

In addition to being analysed in their own right, native speaker data are increasingly used as a **baseline in analyses of learner fluency** (see *infra* for further details). As convincingly pointed out by Foster and Tavakoli (2009), despite the fact that there might be some unease in using the native speaker as the "gold standard" (especially for languages that have many native varieties), using a native speaker baseline gives **greater validity** to claims that learner performance is affected by some variable. A native-speaker baseline may, for example, help distinguish which performance features are due to language processing difficulties and which are due to the particular design of the task (see also *infra*).

1.3.2 Fluency in a foreign language

Like native speech, learner speech is interspersed with all kinds of disfluencies. Learners are generally assumed to produce fewer disfluencies and speak faster as their perceived fluency (and their proficiency) increases: the underlying assumption seems to be that, **the higher the automatisation of procedural linguistic knowledge, the higher the fluency** (Pawley & Syder 1983; Sajavaara 1987).

More specifically, longitudinal and cross-sectional studies on L2 fluency provided evidence that increases in fluency level (and fluency perception) are clearly related with **improvements in temporal fluency measures** such as speech rate and mean length of runs (Lennon 1990; Towell 1987; 2002). Evidence suggests that a higher fluency level is, however, not correlated with gains for other measures (Baker-Smemoe *et al.* 2014; Iwashita *et al.* 2008; Tonkyn 2012). In a longitudinal study, Derwing *et al*. (2009) found no indication of any difference between the **mean length of unfilled pauses** in L2 English across three collection times spread over 2 years. These results contradict a previous study by Lennon (1990), but the learners in the two studies have different mother tongue backgrounds, and the time-span considered is also different, which may account for the diverging results in the two studies.

As is the case for many other language-related characteristics, some aspects of L2 fluency are in fact attributable to each person's **individual and idiosyncratic speech characteristics**. Idiosyncrasies have been claimed to "pervade [a] person's oral performance in any language and are not specific to the learner's performance in a particular language" (Guz 2015:235). Besides, **learner variables** and personality traits (Dewaele 1996; Dewaele & Furnham 2000) as well as **L1** and **L1 speech habits** (Derwing *et al.* 2009; Hincks 2010) also affect L2 fluency.

Research has further shown that L2 fluency is also affected by external factors, such as the **learning context** (Freed, Segalowitz & Dewey 2004; Götz 2017). For example, Freed *et al*. (2004) analysed the fluency development of learners in three settings over one semester. The intensive summer immersion programme group improved the most, followed by the study-abroad group, while learners in an at-home setting (i.e. formal classroom) did not show any significant gains. Such results raise numerous questions about the nationally advocated

language policies, and about the worrying lack of improvement in formal classroom settings, although further data are obviously needed to assess the evolution of each group over a longer period of time.

A last important factor affecting L2 fluency is **speaking task** (e.g. De Jong *et al.* 2012b; Derwing *et al.* 2004; Dumont 2017b; Ejzenberg 1996; Ellis 2009). Ellis (1985:241) convincingly argues that "[d]ifferent tasks call for different types of knowledge, each of which calls for knowledge types that vary in terms of analycity and automaticity". Likewise, Trosborg (1995:102) stressed that "different tasks require (and develop) different types of knowledge, and the knowledge which is acquired in one type of setting is not necessarily applicable to, or available for, a different kind of task". In other words, he posits that the knowledge – or **fluency – acquired for one speaking task might not be directly transferable to other tasks**, which has considerable implications for language learning and assessment.

Previous research purporting to task effects has borne out that there is a clear **beneficial effect of pre-planning** and of the amount of pre-planning time on fluency (Crookes 1989; Ellis 2009; Foster & Skehan 1996; 1999; 2009; Goldman, Auchlin & Simon 2013; Mehnert 1998; Ortega 1999; Skehan & Foster 1997). The **level of interaction** in speaking tasks is another important factor affecting fluency: as compared to monologues, dialogic speech has been shown to favour fluency in terms of speed and repair measures, and length of unfilled pauses (Ejzenberg 2000; Michel 2011; Riggenbach 1989; Witton-Davies 2014). Clear **task structure** also supports fluency (Cucchiarini, Strik & Boves 2002; 2010; Ejzenberg 1996; Foster & Skehan 1996; 1999), but increase in **task complexity** leads to a cutback in fluency (Skehan 1998; Skehan & Foster 1999).

From a cognitive psychological point of view, the findings in the above mentioned studies are broadly consistent with the so-called **Trade-off Hypothesis** (Skehan 1999), which states that, given that attentional capacity and working memory are limited, committing attentional resources to one aspect of performance may have a negative impact on others. In particular, there seems to be a tension between three important aspects of performance, namely complexity, accuracy and fluency. The Trade-off Hypothesis predicts that committing attention to one of those three areas might cause a lower performance in the others. More specifically, speaking task characteristics (e.g. structure or complexity) and task conditions (e.g. planning conditions and level of interactivity) induce learners to direct their attention to different dimensions of language performance.

1.3.3 L1 vs. L2 fluency

One of the central questions in analyses of learner language is the question of how L2 speech compares and differs from speech production in a first language. Kahng (2014:809) claims

that "[o]ne of the most noticeable differences between speech in first language (L1) and second language (L2) is found in fluency".

Two main approaches to the relationship between native and non-native speech fluency can be distinguished (Guz 2015:235–236). On the one hand, learner fluency can be investigated **from the point of view of the target language**. For example, French-speaking learners of English could be compared with native English speakers. In this approach, L2 fluency is analysed in terms of deviations from L1 standards. Alternatively, learners at different proficiency levels (or, more rarely in (dis)fluency research, of different mother tongue backgrounds) can be compared and contrasted. This approach helps "identify the possible source of certain non-standard features" (Gilquin & Granger 2015:424). On the other hand, L2 fluency can also be investigated **from the point of view of the learner's mother tongue**. For example, the performance of French-speaking learners of English could be compared with their performance in French, i.e. their mother tongue. In this approach, a fluency gap can be observed between the learner's performance in the L2 and his/her assumed more fluent performance in the native language.

A. L2 fluency from the perspective of the target language

A wealth of studies support the commonly held assumption that learner speech is more disfluent (i.e. **contains more disfluencies**) than native speech. Evidence supports that non-native speech is produced at a considerably lower speed, and there is a higher incidence of pauses and other hesitation phenomena (Cucchiarini, Strik & Boves 2000; Deschamps 1980). Learners are claimed to become more fluent as their proficiency in the target language increases (Freed 2000; Towell, Hawkins & Bazergui 1996), but even very advanced L2 learners have been found to be less fluent in the target language than their native counterparts (Kahng 2014).

The gap between L1 and L2 fluency has generally been explained in terms of **differences in language knowledge, processing, and degree of automaticity**[14] **in the target language** (Guz 2015; Kormos 2006). More specifically, Kormos (2006:154) explains that the difference in fluency between L1 and L2 production "might be caused by a number of factors such as the deficient knowledge of L2 lexis, syntax, morphology, and phonology, attentional resources needed for suppressing L1 production procedures, and greater demands on self-monitoring"[15]. More specific similarities and differences will be reviewed in Section 1.4.

[14] Automaticity refers to "the absence of attentional control in the execution of a cognitive activity, with attentional control understood to imply the involvement, among other things, of intention, possibly awareness, and the consumption of cognitive resources, all in the service of dealing with limited processing capacity" (Kahneman 1973; in Segalowitz & Hulstijn 2005:371).

[15] Note that this argument seems equally valid for the fluency gap between the learners' L2 and the learners' mother tongue.

In the domain of **testing and assessment**, the fluency level of learners is generally also assessed according to a native standard. Not only do these tests seem to assume that native-like performance is the final goal for L2 learners, but they also disregard variability in native speaker performance: **assessment on grounds of a unitary native norm should, in fact, be reappraised** (*cf.* also Bosker 2014:9).

B. L2 fluency from the perspective of the mother tongue

In the second perspective, learner fluency is analysed from the point of view of the learners' mother tongue. The general assumption is that **learners are considerably more fluent in their native language**. To date, however, little empirical evidence has provided support to this assumption. In her comparison of Swedish learners of English talking in Swedish and in English, Hincks (2010) showed that the learners talked with a significantly lower speech rate in their L2 (a drop by 23%) and the mean length of their speech runs was also significantly shorter (about a quarter shorter) than in their native language.

Recent studies probing more closely into the relationship between L1 and L2 fluency suggest that the **fluency characteristics in a speaker's L2 are strongly related to those in the L1** (De Jong *et al.* 2015; Segalowitz 2010). Early studies showed that pause patterns, length, and distributions partly result from the speaker's personal speaking style and are likely to permeate both the learner's mother tongue and his/her L2 (Deschamps 1980; Raupach 1980). More recently, De Jong *et al*. (2009; 2012a; 2015) provided strong evidence that a large part of fluency-related phenomena are characteristic of the way individuals speak in general, and not just typical of their speech production in the L2. Speech rate, length of pauses, and the number of filled pauses are, for example, all significantly related in a speaker's L1 and L2 (De Jong *et al*. *ibid*.). In other words, speakers who speak slowly, produce long unfilled pauses and many filled pauses in their L1 are likely to speak slowly, produce long unfilled pauses and many filled pauses when speaking in their L2, too. By contrast, the number of L2 unfilled pauses (De Jong, Schoonen & Hulstijn 2009) and repetitions (Rose 2013) appears not to be correlated with the L1. Furthermore, Derwing *et al*. (2009) also found a close relationship between L1 and L2 temporal measures. Cox & Baker-Smemoe (2013), however, indicated that the strength of the relationship was stronger for lower levels of fluency than for higher levels, which suggests that, as learners' fluency progresses, their hesitation patterns may become more similar to those of the target language (and hence, more unmarked).

In conclusion, evidence suggests that **learners' fluency in the target language and in their L1 should be considered in conjunction** and that **L2 fluency measures should ideally be corrected for L1 characteristics**. Guz (2015:237) further claims that "L2 fluency is not L2 specific but results from an [sic] individual, idiosyncratic speech differences".

1.3.4 Summing up

Before moving on to the next section, which looks at the previous literature on the fourteen fluency measures under investigation in this thesis, it is useful to briefly situate our approach and to clarify the meaning of some relevant terminological terms in the frame of this thesis.

The perspective on fluency adopted in this thesis conforms to Segalowitz's (2010) three-way distinction between cognitive, utterance and perceived fluency. However, the focus lies on **utterance and perceived fluency**, and no direct claims about cognitive fluency will be made. Seen in this light, fluency encompasses **a set of observable features** in the learners' and native speakers' utterances which can be objectively measured, quantified, and qualified, and analysed both from the speakers' and the listeners' perspective. These features include not only temporal variables, but also non-time-related features (e.g. repetitions and repairs).

Bearing in mind that previous literature has provided evidence that disfluencies may either be used as a symptom of high cognitive effort, or as a signal that has a positive effect on the listener's comprehension, the neutral term ***(dis)fluency feature*** will henceforth be adopted to refer to the ambivalent role such features can take. Likewise, ***(dis)fluency*** will be used to stress the fact that fluency and disfluency are not two polar opposites, but the two ends of a continuum. The terms *fluency* and *disfluency* will be used only to refer to those end-points.

Furthermore, while I acknowledge that some **non-verbal aspects** of communication also contribute to fluency (*cf.* Götz 2013a) and that this area is definitely worth probing into, I will not investigate this aspect as the data do not lend itself to this type of analysis.

Lastly, due to the nature of the data used in this study, the **learners' fluency in their mother tongue** will not be considered, but detailed attention will be paid to individual variation.

These considerations round off the first two main parts of this chapter. The next section turns to a brief survey of the fourteen (dis)fluency features under investigation.

1.4 AN ARRAY OF (DIS)FLUENCY FEATURES

Just like defining (dis)fluency is not straightforward, the establishment of its tangible components is not without difficulty either. There is, in fact, no generally agreed upon operationalisation of (dis)fluency (Foster & Skehan 2009:281). Different approaches to delineating (dis)fluency features may, however, be pointed out, like for example (*cf.* also Kormos 2006:162):

- the exploration of the **temporal aspect** of speech production (Dechert & Raupach 1980; Grosjean 1980a; Raupach 1980);
- the examination of **interactive mechanisms** (Ochs, Schegloff & Thompson 1996; Riggenbach 1991);
- the analysis of **phonological aspects** of fluency (Gut & Fuchs 2017; Hieke 1984; Isaacs & Trofimovich 2011; Wennerstrom 2000);
- the study of **formulaic speech** (Götz 2013a; Towell, Hawkins & Bazergui 1996; Wood 2010);
- the investigation of **repair phenomena** (Blackmer & Mitton 1991; Corley 2010; Hedström 1984; Kormos 1999; Levelt 1983; Pillai 2006; Postma, Kolk & Povel 1990; Witton-Davies 2010);
- the relationship between **speed, breakdown and repair** fluency (Bosker *et al.* 2013; Lahmann, Steinkrauss & Schmid 2015; White 1997);
- the analysis of fluency in the frame of the **CAF** (complexity-accuracy-fluency) framework (Derwing & Rossiter 2003; Ellis & Yuan 2004; Housen, Kuiken & Vedder 2012; Larsen-Freeman 2006; 2009; Skehan 2009; Skehan & Foster 2007; Tonkyn 2012; Yuan & Ellis 2003).

Empirical studies in (dis)fluency research have explored (dis)fluency features from three main angles. (Dis)fluency has been investigated from a **longitudinal** perspective (Freed 1995; Towell, Hawkins & Bazergui 1996). Researchers have also **compared** fluent and non-fluent (or less fluent) learners, e.g. learners and native speakers, or learners from different proficiency/(dis)fluency levels (Götz 2013a; Guz 2015; Gráf & Huang 2017; Riggenbach 1991). Lastly, **perceived (dis)fluency scores** or levels have been correlated with empirical measurements of utterance fluency (Bosker *et al.* 2014; Götz 2013b; Kormos & Dénes 2004; Préfontaine 2013a; Rossiter 2009).

The two central (and interrelated) problems that arise from previous studies are **the way features are defined** and the **way they are quantified**. A focal example is speech rate: it may be defined as the number of syllables or words per minute. Depending on the definition, the

quantification will differ. Likewise, there is little consistency as to the definition of an unfilled pause: while some adopt a threshold, others do not, and when a threshold is adopted, it differs greatly from study to study (see below for more details).

The next sections set out to provide an overview of the literature on the fourteen (dis)fluency features under investigation in this thesis. I will first explore temporal and breakdown variables, before moving on to repair fluency variables.

1.4.1 Speech rate

The two main measures of the rate of speech delivery are speech rate (SR) – which is the most widely used measure – and articulation rate (AR).

Speech rate measures the **ratio between the amount of speech produced and the time needed to produce this speech**. Speech rate is often seen as an inclusive measure of (dis)fluency in the sense that, "as it includes pause time, it can be considered to cover both the encoding of ideas and of the speech forms used to communicate them, inclusive of the time needed to retrieve the forms from memory stores" (Towell 2012:62). On the other hand, speech rate has been criticised for being a hybrid measure that actually confounds rate of speech and pausing time (De Jong *et al.* 2012b). It is difficult to know, for example, whether a higher speech rate is due to a higher articulation rate or to a smaller amount of pausing time. The other measure of rate of speech, **articulation rate**, "focuses on the amount of time required for a speaker to physically produce speech" (Ginther, Dimova & Yang 2010:382). In other words, it gives an indication of the **amount of speech as a function of the articulation time** (i.e. the speaking time minus the time devoted to pausing). As such, AR has been claimed to be the "purest" measure of the rate of speech (Witton-Davies 2014:71). A disadvantage of this measure, however, is that it is greatly affected by the threshold adopted for measuring unfilled pauses: the lower the minimum threshold, the more pausing time will be excluded, and vice versa.

Speech (and articulation) rate has been explored in a first and second/foreign language perspective. The rate of speech has been shown to **vary between** as well as **within individuals** (Jacewicz, Fox & Wei 2010; Quené 2008; 2013; Tsao, Weismer & Iqbal 2006) and some evidence has also been provided supporting **cross-linguistic variation** (Pellegrino, Coupé & Marsico 2011). A stream of research has focused more particularly on **task-related differences**. An important finding is that speech rate is consistently higher in dialogic tasks than in monologues (Ejzenberg 1997; 2000; Kowal, Wiese & O'Connell 1983; Riggenbach 1989; Tavakoli 2016). The relationship between rate of speech and information density or **comprehension** has also been investigated (e.g. Anderson-Hsieh & Koehler 1988; Griffiths & Beretta 1991; Lane *et al.* 1973). For example, Munro and Derwing (2001) highlighted a

curvilinear relationship between speech rate and comprehensibility by demonstrating that both slow and high speech rates are related to lower comprehensibility.

With respect to the practical measurement of the rate of speech, although the number of **syllables per second** has routinely been employed (e.g. Baker-Smemoe *et al.* 2014; Cucchiarini, Strik & Boves 2000; Derwing *et al.* 2009; Griffiths 1991; Rossiter 2009), some studies report speech (articulation) rate as the number of **syllables per minute** (e.g. Ginther, Dimova & Yang 2010; Kormos & Dénes 2004; Mehnert 1998; Towell 1987; Towell, Hawkins & Bazergui 1996). Moreover, while some studies base their measures on **unpruned** syllables (e.g. Derwing *et al.* 2004; Rossiter 2009), other use the **pruned** counts, i.e. excluding hesitations such as repairs and/or false starts and/or repetitions and/or filled pauses etc. (e.g. Kormos & Dénes 2004; Préfontaine 2013b). In a sense, using pruned speech (or articulation) rate is also a hybrid measure as it combines (or confounds) rate of speech and rate of repairs. Counting syllables is, however, "a tedious job and [it] is often cast aside due to time constraints" (De Jong & Wempe 2007:52). It is probably one of the reasons why a number of studies have expressed speech rate as the amount of **words per second or minute**, such as Lennon (1990), Freed (1995) and Freed *et al*. (2004) or Riggenbach (1991).

Figures associated with **speech rate in native language** display **considerable range**. Early studies reported speech rates ranging from c. **120 to 180 words per minute** (about two to three words per second) (Levelt 1989; Maclay & Osgood 1959). Brand and Götz (2011) measured an even higher mean rate in a corpus of native British English: 218 words per minute on average. When it comes to the measurement in terms of syllables, studies reported mean rates of about 210 syllables per minute (Ginther, Dimova & Yang 2010; Rohr 2017) or 4 syllables per second (i.e. c. 240 syllables per minute) (Guz 2015; Hincks 2010), also with considerable range – from **140 to 260 syllables per minute** in Goldman-Eisler (1968).

Learners' speech rates are on average **lower** than native speakers': about 160 words per minute (ranging from 117 to 190 wpm) in German learners of English (Brand & Götz 2011), 2.7 syllables per second in Polish learners of English (Guz 2015), 3.12 syllables per second in Swedish learners of English (Hincks 2010). Interestingly though, Bosker and Reinisch (2015:online publication) noted that, although learners' speech rate is indeed significantly lower than native speakers', "nonnative speech is implicitly *perceived as faster* than temporally-matched native speech" (my emphasis), which leads them to suggest that the additional cognitive load of listening to a non-native accent speeds up rate perception. A large part of learners' variability can, however, be attributed to differences in language mastery as speech rate has been proved to be strongly and positively **correlated with proficiency level** (Baker-Smemoe *et al.* 2014; Ginther, Dimova & Yang 2010; Iwashita *et al.* 2008; Kormos & Dénes 2004; Osborne 2010). Besides, **speech rate in a first and a foreign language are also correlated**: speakers who speak relatively slowly (or quickly) in their mother tongue also tend to do so in the L2 (e.g. Cox & Baker-Smemoe 2013; Derwing *et al.*

2009; Guz 2015; Towell, Hawkins & Bazergui 1996). Yet, inconsistent evidence has been found for Japanese learners of English (Rose 2013).

1.4.2 Mean length of runs

The mean length of speech runs (MLR) gives an indication of the **amount of speech between unfilled pauses** (as a matter of fact, it is thus also an indicator of the frequency of unfilled pauses). Mean length of runs is seen as "a measure of the ability of the speaker to encode units of speech" (Towell 2012:62): calculating the mean length of runs helps to determine the "level of routinization of knowledge representation" (*ibid*.:121) as well as the "level of access to all the syntax and lexis the speaker controls" (*ibidem*).

Although speech runs are delimited by pauses, they do not necessarily represent a semantic or syntactic unit in speech. They may "reflect a word, a phrase, a sentence or a series of sentences depending on the task and the rate of output" (Grosjean 1980b:40), but "[l]onger runs suggest that more elements of speech are being combined in a shorter space of time" (Towell 2012:62).

MLR is typically expressed in **syllables per run**, but some studies also report this measure in **words per run** of speech (e.g. Gut & Fuchs 2017).

As rightly stressed by Götz (2013a), defining the boundaries of speech runs has raised a lot of debate. Most researchers consider only **unfilled pauses as run boundaries**, but **filled pauses** are sometimes also taken into account as run delimitations (e.g. Derwing *et al.* 2004). Besides, defining a run as a stretch of speech between unfilled pauses leads to the key issue revolving around the measurement of UPs, i.e. their cut-off point[16]. Whilst most studies follow the standard set by Goldman-Eisler (1968) by adopting a 0.25 second cut-off point (e.g. Grosjean 1972; 1980c; Kormos & Dénes 2004; Préfontaine 2013a; Tavakoli 2016), other thresholds have been chosen as well, such as 0.4 second (Derwing *et al.* 2004) or 1 second (Iwashita *et al.* 2008; Mehnert 1998).

Native speech runs have been found to be **significantly longer** than learners', with reported L1 means around 12 to 14 syllables and L2 means of about 8 syllables per run (Ginther, Dimova & Yang 2010; Guz 2015; Hincks 2010; Rohr 2017). According to Guz (2015), MLR is not correlated in a speaker's L1 and L2. With respect to **task-induced differences**, results so far are mixed. Gut (2009) noted that for both native and non-native speakers, the mean length of runs is higher when reading a scripted passage compared to when retelling a story (i.e. a non-scripted task). Unlike her, Cucchiarini *et al*. (2010), who focused on learners of Dutch, found no evidence supporting that there might be a difference between read and

[16] This aspect will be discussed at length in Section 1.4.4.

spontaneous speech in terms of mean length of runs. Besides, a study of B2 learners of English revealed that runs in dialogic speech are significantly longer than runs in monologic tasks (Tavakoli 2016). Several studies have further reported that mean length of run strongly **correlates with both (self-) perceptions of (dis)fluency** (e.g. Cucchiarini, Strik & Boves 2000; Kormos & Dénes 2004; Préfontaine 2013a; Préfontaine, Kormos & Johnson 2015) **and L2 proficiency** (Baker-Smemoe *et al.* 2014; Ginther, Dimova & Yang 2010; Kahng 2014; Kormos & Dénes 2004).

1.4.3 Phonation time ratio

Phonation time ratio (PTR) summarises the amount of time filled with speech as a percentage of the total time and is calculated by dividing the amount of time spent articulating by the total time of the speech sample. Obviously, such a measure also heavily relies on the accurate identification and measurement of all unfilled pauses.

While phonation time ratio is the time spent speaking, another measure, "pause-time ratio" is the time spent pausing. The two measures are the opposite sides of the same coin and come in complementary distribution. PTR should ideally be considered together with a measure of the frequency of unfilled pauses and their mean length to identify whether between-speaker differences are due to a higher/lower frequency of unfilled pauses, or whether they are due to longer/shorter pauses.

Kahng (2014) has claimed that native English speakers devote about 17% of their time on unfilled pauses, which amounts to a **phonation time ratio of c. 83%**. By comparison, he calculated that Korean **learners** of English use about 32% of their time on pauses, that is, a **PTR of 68%.** In their study of Hungarian learners of English, Kormos and Dénes (2004) estimated the PTR of low-intermediate learners at c. 52% and that of advanced learners at c. 69%.

Analyses have shown strong **correlations between PTR and (dis)fluency perception** (Cucchiarini, Strik & Boves 2000; Kormos & Dénes 2004): the higher the phonation time ratio, the higher the perceived fluency. Phonation time ratio has also been shown to increase with increasing proficiency level (Iwashita *et al.* 2008).

Another important factor that influences phonation time ratio is the type of speaking task. PTR is for example **higher in dialogues** than in monologues (Moniz 2013; Tavakoli 2016), **higher in pre-planned tasks** than in non-prepared tasks (Foster & Skehan 1996), and **higher in read speech** than in spontaneous speech (Cucchiarini, van Doremalen & Strik 2010).

1.4.4 Unfilled pauses (and mean length of unfilled pauses)

Research into the use of unfilled pauses (UPs) is probably the most abundant among the (dis)fluency features analysed in this thesis. UPs have first been analysed in monolingual speakers in terms of frequency, distribution and length, and their investigation was then expanded to cross-linguistic and SLA contexts.

Research findings indicate that unfilled pauses appear at a **high frequency in speech**: Biber *et al*. (1999) found a frequency of more than **19,000 unfilled pauses per million words** (1.9 phw), Götz (2013a) obtained higher values at **4 UPs per hundred words (phw) on average for native speakers** and as many as **15 UPs phw for advanced German learners** of English. However, in their article comparing read and spontaneous learner speech, Cucchiarini *et al*. (2002) highlight the great **impact of task type** on the frequency of pauses. Whereas in their data, advanced learners uttered about 10 UPs per minute in read speech, they produced about three times as many in spontaneous speech (and their mean duration was more than doubled (0.34 ms vs c. 1 sec)). Research has generally shown that the frequency of UPs is a **reliable indicator of perceived (dis)fluency** (e.g. Préfontaine, Kormos & Johnson 2015; Riggenbach 1991), though others have found otherwise (e.g. Kormos & Dénes 2004).

The **length of unfilled pauses** also seems to raise a lot of interest and controversy. A number of studies investigating the relationship between mean length of pauses and (dis)fluency perception confirm the existence of a **relationship between mean length of pauses and (dis)fluency perception**. Whilst Bosker *et al*. (2013) found a significant negative relationship between (number and) mean length of UPs and perceived fluency for learners of Dutch, Préfontaine *et al*. (2015) highlighted a positive relationship with learners of French producing longer average pause times judged to be more fluent (i.e. longer UPs are judged to be more fluent). This latter finding was related to Grosjean and Deschamps' (1975) observation that French L1 speakers pause longer than English native speakers: longer L2 French unfilled pauses are thus arguably more native-like. Yet, a number of studies also found that the **duration of pauses might be less important than their frequency**: Cucchiarini *et al*. (2002:2870) observed that, whereas fluency ratings are strongly related to the number of pauses per minute both in read and spontaneous speech, it is less so for mean length of pauses in read speech, and there is "almost no relation at all with perceived fluency" in spontaneous speech. This confirms their earlier finding (Cucchiarini, Strik & Boves 2000) that "less fluent speakers, in general, do not make longer pauses than more fluent speakers, but they do pause more often" and that "mean length of silent pauses seem[s] to have almost no relation at all with perceived fluency", which is also supported by Kormos and Dénes (2004).

The frequency and length of unfilled pauses has been shown to be **correlated in a speaker's L1 and L2**: speakers who tend to pause a lot in their native language also tend to pause frequently in their L2, and speakers who tend to produce long pauses in their L1 also tend to do so in a foreign language (e.g. Cox & Baker-Smemoe 2013; Derwing *et al.* 2009; Rose 2013).

The divergence between the results obtained so far for UP frequency and length may be partly explained by **methodological discrepancies**: in addition to the plethora of different task types or L1s and/or L2s used, researchers do not seem to agree on the **minimum threshold** for what should count as an unfilled pause. A quick review of the literature shows that many different benchmarks are used – Table 1-1 provides a succinct overview, with selected references, of the range of thresholds that are commonly used by L1 and L2 researchers. As can be noted, the lower thresholds range from 0.10 sec to as high as 1 or even 3 seconds. Importantly, the justification for adopting one or the other is often practical or arbitrary (see e.g. Kirsner, Dunn & Hird 2003). Towell *et al*. (1996:91–92), for example, set their threshold at 0.28 seconds because "the speed at which the mingograph ran and the squared paper on which the printouts were produced meant that each small box represented .04 of a second. The experimenters could easily see when a line was crossed but could only measure with extreme difficulty the amount needed for 0.25 seconds, i.e. 6 boxes plus one fifth of a box. For purely practical reasons, it was decided to use 0.28 seconds". This explanation rightly stresses the great **technical challenges** facing the UP researcher, and highlights the importance and **benefits of new technologies**, for example, for the automatic detection and measurement of unfilled pauses.

Lower UP threshold	Selected references
0.10 sec	Griffiths (1991); Trofimovich and Baker (2006)
0.20 sec	Cucchiarini, Strik and Boves (2002); Kormos and Dénes (2004); Segalowitz (2010)
0.25 sec	De Jong *et al*. (2012a); Préfontaine, Kormos and Johnson (2015); Tavakoli (2016); Towell (2002)
0.28 sec	Towell *et al*. (1996)
0.30 sec	Baker-Smemoe *et al*. (2014); Tonkyn (2012)
0.35 sec	De Jong *et al*. (2012b)
0.40 sec	Derwing *et al*. (2009); Freed (2000); Tavakoli (2011)
1.00 sec	Iwashita *et al*. (2008)
3.00 sec	Fulcher (1996)

Table 1-1: Lower UP thresholds in the literature

The issue of the under-evaluated consequences of the choice of a lower threshold has been brought to the foreground in several studies. Kowal *et al*. (1983:385), for example, noted that between 50 and 71.5 % of unfilled pauses occur in the duration interval between 0.25 and 1 second. The argue that "[a] considerably higher cut-off point excludes part (or all) of these pauses, lowers the percentage of pause time/total time, lengthens mean pause duration and increases mean phrase length, and finally yields a slower articulation rate – all quite independently of any real change in the data". In their article, De Jong and Bosker (2013)

aimed to find the optimal cut-off point for L2 research: they found that, although the correlations between fluency ratings and mean duration of unfilled pauses were always high, irrespective of the threshold (they however advise researchers to adopt a 250 ms threshold)[17], for the number of pauses, **the higher the cut-off point, the higher the correlation with fluency ratings** – but, as they rightly explained, by setting a high threshold, counting the number of long pauses is confounded with measuring the duration of UPs. Furthermore, Campione and Véronis (2002), in their large-scale corpus study of unfilled pause duration, also showed the dangerous effects of these different thresholds when comparing findings. They highlighted the fact that the **distribution of pauses appears as trimodal**, and distinguished brief (below 200 ms), medium (between 200 and 1000 ms) and long (above 1000 ms) pauses, the latter occurring only in spontaneous speech. The methodology they adopted as well as the subsequent results contrast with less empirically-based classifications such as those suggested by Goldman-Eisler (1961a) or Riggenbach (1991), which are still commonly used in L2 research.

Finally, the **distribution and the function of pauses** have also been explored. Most researchers agree to distinguish pauses that appear at **syntactic boundaries** and those that occur **within syntactic units**. Drommel (1980, in Carter & McCarthy 2006) also separates intentional pauses (the so-called "T-Pauses" that fulfil communicative functions and are claimed to be perceived as natural by the listener) from unintentional unfilled pauses (the "D-Pauses", which are motivated by planning demands and which generally occur at major grammatical transition points). Chafe (1980), for his part, highlights the **multi-functionality of pauses**, and argues that, whilst speakers pause *between* phrases and clauses when they make a decision about what to say next, they stop *within* phrases and clauses when they experience difficulty in deciding how to verbalise something they already have in mind. Götz (2013a:184) observes that UPs within clauses and within constituents are overused by German non-native speakers (see also e.g. Brand & Götz 2011; Davies 2003; Lounsbury 1969). She also explains that unfilled pauses within clauses are "very frequently caused by the learners' search for the appropriate lexical item or correct grammatical form they want to use". In addition, Chambers (1997:540) stresses that "[b]ecoming fluent therefore is not about speaking faster (articulation rate), but about pausing less often and *pausing at the appropriate junctures in an utterance* (my emphasis)".

1.4.5 Filled pauses

Biber *et al*. (1999:1053) define filled pauses (FPs) by contrast with unfilled pauses (UPs): whereas the latter are occupied "by silence", a filled pause is defined as "a vowel sound, with

[17] Their study also concluded that there was no optimal threshold either between the mean log duration of UPs and vocabulary knowledge as a measure of L2 proficiency.

or without accompanying nasalization". The fact that these two phenomena are referred to under the umbrella term "pauses" makes the implicit assumption that they have the same underlying function, which might be misleading (*cf.* Campione & Véronis 2005).

According to Clark and Fox Tree (2002:92), filled pauses tend to be built around central vowels in the language, are generally brief (though they may be extendable), may contain a nasal consonant (*m* or *n*) and are subject to speaker-preferences. In terms of transcription, curiously, and although no difference in pronunciation is implied, FPs are usually transcribed as *eh* and *erm* in British English and *uh* and *um* in American English. Departing from the general view that considers them real pauses (see e.g. Corley & Stewart 2008 for a discussion), Clark and Fox Tree (2002) argue that filled pauses may be regarded as real words (i.e. "linguistic units that have conventional phonological shapes and meanings and are governed by the rules of syntax and prosody" (*ibid.*:75)) because "*uh* and *um* must be planned for, formulated, and produced as parts of utterances just as any other word is".

Filled pauses are often argued to be **one of the preferred strategies** for hesitating for both native speakers and learners, before drawls, repeats or false starts for example (Grosjean & Deschamps 1975; Grosjean 1980c). Depending on the nature and the properties of the corpora[18], various frequencies have been reported in the literature. In **native language**, Biber *et al*. (1999:1054) estimate the frequency of filled pauses at 13,000 per million words (i.e. 1.3 phw). Götz (Brand & Götz 2011; Götz 2013a), who investigated speech management strategies in interviews, found that British English native speakers use 2.27 FPs phw, i.e. nearly twice as many filled pauses on average as the frequency reported by Biber *et al*. (1999:1054). With regard to L2 speech, **learners** are said to produce significantly more pauses than native speakers: German learners of English, for example, utter 5.12 FP phw on average (range: 1 to 14 filled pauses phw) (Brand & Götz 2011; Götz 2013a). A high frequency of filled pauses has also been observed by De Jong *et al*. (2012a) in the language of intermediate to advanced learners of Dutch with varied L1s, who produced 11.8 FPs phw on average. Although diverging results were obtained for the correlation between the frequency of FPs in a speaker's L1 and L2, studies investigating the **relationship between L1 and L2 fluency** (e.g. Guz 2015) confirm the hypothesis that pausing occurs more frequently in a speaker's L2 than in their mother tongue.

Several researchers have investigated the link between pausing behaviour and **L2 proficiency**. No such correlation could be observed in Iwashita *et al*. (2008) or Baker-Smemoe *et al*. (2014), hence indicating that highly proficient learners do not pause less often than lower-proficient L2 speakers. But despite the absence of a link with L2 proficiency, the

[18] Among others, the type of speaking task has a great influence on the frequency of filled pauses: there are for example more FPs in dialogues than in monologues (Tavakoli 2016) and in complex tasks than in cognitively more simple tasks (De Jong et al. 2012b).

frequency of FPs has been shown to be a good **predictor of fluency rating** (e.g. Bosker *et al.* 2013; Foster & Skehan 1999; Rossiter 2009).

The literature is rife with possible reasons, roles and impacts of filled pauses. Clark and Fox Tree (2002) provide quite a comprehensive account of researchers' interpretations of filled pauses. On the part of the speaker, their presence is generally associated with **preparedness problems**. As such, they tend to be viewed negatively, especially in more formal settings where admitting to lack of preparedness may undermine a speaker's authority. Clark and Fox Tree (2002:98) explain that "[c]ourses on public speaking train people to speak without *uh* and *um*, and the best public speakers are successful. In all of the recorded inaugural speeches by US presidents between 1940 and 1996, for example, there is not a single *uh* or *um*". In a less extreme view, filled pauses are associated with **planning problems** due to higher cognitive load, and reflect the fact that speakers are e.g. searching for a word, are in doubt, are asking for help, or want to keep the floor. As a consequence, filled pauses are more likely to happen at places where cognitive load is heightened, such as the beginning of utterances or phrases (e.g. Barr 2001; Maclay & Osgood 1959), with more complex or unfamiliar topics (e.g. Bortfeld *et al.* 2001; De Jong *et al.* 2012b) and before low frequency and unpredictable words (e.g. Beattie & Butterworth 1979; Levelt 1983). Fox Tree (2001:320) further suggests that *um* and *uh* do not have exactly the same function: whereas *uh* is indicative of a short delay and facilitates lexical identification, *um* is a signal of a long upcoming suspension and does not have this facilitating effect. This different function, he argues, "might be what underlies a number of disparate proposals about the functions of *um*s and *uh*s" (but see e.g. Fraundorf & Watson 2011).

The analysis of the role of filled pauses has also focused on the way they positively **impact the listener**. It has been claimed that, because they **add time for cognitive processes to unfold**, the presence of filled pauses (and other interruptions) in the discourse is beneficial to the listener (e.g. Brennan & Schober 2001; Watanabe *et al.* 2008). Prior work (e.g. Fox Tree 2001; Corley, MacGregor & Donaldson 2007; Fraundorf & Watson 2011) has found that, not only do filled pauses **heighten immediate attention** to upcoming speech, but they also have longer-term effects: when FPs precede a word for example, the word is processed differently and is better recognised in subsequent memory-tests.

1-1: Cliticised filled pauses (from Clark & Fox Tree 2002:101)

> **but-uh** (0.2) **we-um** (1.1) uh **have-uh** (0.1) eight to twelve airplanes **that-uh** enter the airspace **right-uh** in front of the crowd

1-2: FP after a truncation (from Clark & Fox Tree 2002:102)

> no but fr= **uh** but from that point of view it would be odd

It is noteworthy that filled pauses do not always appear in isolation, but can also occur in "**chunks of disfluencies**": Levelt (1989) claims that FPs are used jointly with 30% of repairs, and Riggenbach (1991) observes that the presence of chunks of unfilled and filled pauses

seems to be indicative of non-fluent speakers (fluent speakers produce very few of them). Moreover, Clark and Fox Tree (2002:101) stress that the filled pauses *uh* and *um* "are often cliticised onto prior words and never onto following words", especially with introductory conjunctions as in "an.duh" (*and uh*), "bu.tuh" (*but uh*), "so.wuh" (*so uh*), and "i.fuh" (*if uh*). An example of cliticised filled pauses can be seen in 1-1. The authors note, however, that filled pauses are never cliticised with a word fragment preceding them (Example 1-2).

1.4.6 Restarts (aka repairs)

Restarts (RSs) have been referred to in the literature by varied terms: restarts (Riggenbach 1989), (self-)repairs (Levelt 1989; Pillai 2006), reformulations (Foster & Skehan 1996; Tavakoli 2016), retrace-and-restarts (Biber *et al.* 1999), (self-)corrections (De Jong *et al.* 2012a; Huensch & Tracy-Ventura 2016) etc. Despite this diversity, what those terms have in common is that they refer to the treatment after some kind of trouble in speech production.

Schegloff, Jefferson and Sacks (1977), who were among the first to focus on restarts in conversation, suggest a categorisation of restarts on the basis of who initiates and who carries out the repair. Restarts may be initiated by the speaker him-/herself or by someone else, and they may be repaired either by the speaker or not, which results in four classes of repairs. The authors (*ibid.*, *cf.* also Levinson 1983) observe that self-initiated self-repairs are the most frequently occurring type of restart, followed by self-initiated other-repair, and other-initiated self-repair, and, finally, other-initiated other-repair.

From a psycholinguistic perspective, Levelt (1983; 1989) distinguishes two major classes of repairs, which may be made to correct – phonological, morphological, syntactical or lexical (Fathman 1980) – errors or to "say the same thing in a more felicitous way". The former category is termed "error repairs" and the latter "appropriateness repairs". No matter what the category is, the sequence of words that contains a repair may be subdivided into **three sub-parts**: the *reparandum*, the *edit interval*, and the *repair*. Corley (2010:708) describes this three-tier structure as follows:

> The reparandum consists of the material which will be corrected, and often shows prosodic signs of the upcoming repair. The edit interval follows a suspension of speech, and may include a filler such as *uh*, typically with a long vowel duration (Shriberg, 2001). The repair comprises the information which replaces the reparandum; in more complex cases, it may be preceded by a repetition of all or part of the pre-repair utterance [...].

In addition to those three parts, the "suspension of speech", or "moment of interruption", is often identified. It corresponds to "the point at which the flow of speech is interrupted for editing" (Levelt 1983:44). It can take place either during the utterance of the troublesome item (which would result in a truncation), or shortly after it, but research has shown that it is more likely to occur at word boundaries than within a word (Du Bois 1974; Levelt 1989; Nooteboom 1980).

In Example 1-3 below, the reparandum (i.e. the item to be repaired) is *left*. The moment of interruption occurs three syllables after the reparandum (after *again to*), and the edit interval immediately follows this suspension: it corresponds to , *uh … ,*. Lastly, the repair, which is the "correct version of what was wrong before" (Levelt 1983:44), is *pink*.

1-3: The three parts of a repair (from Levelt 1983:44)

```
Go from left again to, uh …, from pink again to blue
```

According to Levelt (1983; 1989), the *repair* in the restart itself can be realised in different ways. It can involve the **simple replacement** of the troublesome word(s), the repetition of some word(s) prior to the troublesome word which is then replaced (i.e. replacement with **anticipatory retracing**), or a fresh start where the speaker starts anew with new material that was not part of the original utterance (i.e. a false start)[19]. These three realisations of repairs are illustrated in 1-4 to 1-6, respectively. The author further argues that the way of repairing depends on the type of repair: correcting errors can be done while preserving the rest of the syntax of the original utterance (i.e. simple replacements); in appropriateness repairs, speakers generally insert new materials.

1-4: Repair with simple replacement

```
Go from left pink to blue
```

1-5: Repair with anticipatory retracing

```
Go from left from pink to blue
```

1-6: Repair with fresh start

```
Go from left it is pink
```

Going a step further, Hedström (1984) as well as Erman (1987) make a clearer conceptual distinction between structures where speakers attempt to "resume the initial syntactic structure even when they add to, specify, reframe or otherwise modify the informational content of their utterance" (Hedström 1984:79; in Denke 2009:115) and complete abandonments of the original syntactic plan (i.e. Examples 1-3 to 1-5 vs. Example 1-6). This distinction is also adopted in the present study: false starts are considered a separate category (see Section 1.4.7). Furthermore, in an attempt to avoid terminological confusion, the term *restart* is used in this study to refer to sequences of words where speakers attempt to resume their original syntactic structure after some trouble, and the term *repair* is used to refer to the third part of the sequence of words that contains a restart, after the *reparandum* and the *edit interval*.

[19] Incidentally, those three types of realisation of *repairs* illustrate the fuzzy boundary between restarts, repetitions and false starts.

Restarts are actually **not very common** in speech. Bortfeld *et al*. (2001) report 1.94 restarts per 100 words across a series of dialogue tasks in native language. In learner speech, a mean of 1.6 RS per hundred words has been reported in the speech of intermediate to advanced learners of Dutch (De Jong *et al.* 2012a). Evidence further indicates that restarts occur **less frequently in dialogues** than in monologic speech (together with lower rates of pausing and higher speech rates) (Witton-Davies 2014). Besides, Temple's (1992) analysis of learners and native speakers of French reveals that half of the native repairs pertain to the search for, or repair of, the noun whereas learner repairs rarely involve nouns (they revolve more frequently around the use of articles and verbs).

The literature seems to suggest that there is **no relationship between restarts and perceived (dis)fluency**. For instance, Cucchiarini, Strik and Boves (2002) did not find any relationship between fluency ratings and "number of disfluencies" (which covers, among others, corrections). Likewise, Kahng (2014) and Bosker *et al*. (2013) did not find any strong association between restarts and perceived (dis)fluency or L2 proficiency.

Restarts are also claimed to be **often accompanied by other (dis)fluency features**, such as filled and unfilled pauses, which can be explained by an increase in cognitive effort. Whereas small-scale modifications of the linguistic form or of the content do not considerably increase the cognitive load, larger-scale changes in the informational content require "significantly greater processing effort", which can be reflected in the use of (dis)fluency features such as pauses between the interruption point and the onset of the correction (Kormos 2000:157). Levelt (1983), for example, reports that 28% of appropriateness repairs and 62% of error repairs are used conjointly with another (dis)fluency feature, the filled pause *er* being the most frequently used. However, Götz (2013a:69; also Riggenbach 1991) argues that restarts, even when they are accompanied by other (dis)fluency features, do not necessarily render the speech less fluent as restarts are generally perceived as a natural phenomenon of speech and are sometimes barely noticed by the listeners.

1.4.7 False starts

False starts are included in Skehan's (2003) typology of repair fluency variables. They refer to speakers **attempting to produce an utterance, but giving up mid-way and starting anew with fresh material that was not part of the original and interrupted utterance** (e.g. Levelt 1989; Nacey & Graedler 2013) – an illustration is provided in 1-7. Contrarily to restarts, false starts are thus characterised by semantic and grammatical incompletion. Fox Tree (1995:710) further writes that:

> False starts occur when speakers start to say something, but then decide to abort their utterances and begin again. For reference purposes, the aborted information will be referred to as the *false start* and the new information replacing it as the *fresh start.* In *for a really champion one you can – it's gonna be twenty cents*, *you can* is the false start and the fresh start

> begins with *it's*. […] In this example, as in the experimental disfluencies, the fresh start completely replaces the information supplied in the false start.

This citation also sheds light on the **terminological fuzziness** and confusion that often exists between the terms **"false start", "fresh start", and "restart"**. A number of researchers consider false starts as extreme cases of repairs, or as the *reparandum* in a restart: they thus merge the two categories together (e.g. Du Bois *et al.* 1993; Freed 2000; Iwashita *et al.* 2008). Alternatively, other researchers assume that false starts and restarts involve different cognitive, strategic and/or psycholinguistic processes (Kormos 2000) and thus keep the ontological distinction between the two categories (e.g. Grosjean 1980c; Levelt 1989; Nacey & Graedler 2013; Pallaud, Rauzy & Blache 2013). Besides, in quite a number of studies, false starts are disregarded altogether. One reason might be that, due to their **strikingly sparse frequency** in spoken discourse (1 in every 60 syllables in Grosjean & Deschamps 1975)[20], they may have been overshadowed by other, more frequent (dis)fluency features such as pauses on the grounds that more frequent features are likely to have a greater impact on fluency. Moreover, false starts are very **difficult to detect automatically**, and their identification requires a great deal of manual work. These factors probably largely explain the low number of empirical, large-scale studies on false starts.

1-7: False start (EN030-S)

> it's in the shape of a big glass pyramid (0.290) and inside **you have** (0.640) the roof comes down in the day to let the sun in

Despite their very low frequency, false starts are typically seen as **detrimental to fluency** (e.g. Dister 2007; Pallaud, Rauzy & Blache 2013). As was the case for restarts, research has shown that they **tend to occur conjointly with one or more other (dis)fluency features, such as unfilled or filled pauses** (Pallaud, Rauzy & Blache 2013; see also Kormos 2000), which further accentuates the interruption caused by the false start.

Studies investigating the **link between false starts and L2 proficiency** found no significant correlation between the frequency of FSs and proficiency level (e.g. Baker-Smemoe *et al.* 2014; Iwashita *et al.* 2008). These results highlight a stark contrast with the intuitive and widespread belief that beginners produce more false starts than more advanced learners, who may have a wider range of speaking strategies and a more solid lexico-grammatical knowledge to avoid using such abrupt interruptions (see e.g. the CEFR descriptors; Council of Europe 2001).

[20] They are "a hardly observable feature in speech" (Wisniewski 2015), and they are even less frequent in very controlled speaking tasks such as reading aloud tasks than in spontaneous speech.

1.4.8 Repetitions

With **reported rates** of c. 1.5 per hundred words in native speech (e.g. MacGregor, Corley & Donaldson 2009) and 2 to over 4 per hundred words in learner speech (e.g. De Jong *et al.* 2012a for learners of Dutch; Gilquin 2008; and Götz 2013a for learners of English), repetitions (also often called repeats) are yet another typical measure of (dis)fluency (e.g. Hasselgren 1998; Lennon 1990; Möhle 1984; Riggenbach 1991).

Foster *et al*. (2000:368) define repetitions as follows: "A repetition is where the speaker repeats previously produced speech. […] However, it is necessary to distinguish between those repetitions which indicate [disfluency], and those which are used for rhetorical effect". Most researchers indeed agree that repetitions fall into two clearly distinct categories (e.g. Huddleston & Pullum 2002; Candea 2000): Candea (*ibid*.) calls them FR. ***faits de langue*** (i.e. language-related repetitions) and FR. ***faits de parole*** (i.e. speech-related repetitions).

The **language-related repetitions** category refers to repetitions that have an intended oratory effect (such as intensity, gradation or emphasis). In this case, **repetitions fulfil a semantic function**, they do not contradict the principle of language linearity and are characteristic of both written and spoken language. Huddleston and Pullum (2002:561) illustrate the use of such an intensificatory repetition by the following example (1-8), where the repetition of *long* is actually intended as an intensification of the adjective, and is equivalent to the meaning of "very long".

1-8: Intensificatory repetition (from Huddleston & Pullum 2002:561)

```
it was a long long way
```

As Huddleston and Pullum (*ibidem*) argue, "[t]he construction [oratory repetitions] should be distinguished from that where a repetition arises in hesitant speech". By contrast with language-related repetitions, **speech-related repetitions** are involuntary and do not have a semantic purpose: they bear witness to a breakdown in language linearity and are a feature specific to spoken language – when they do appear in written language, it is only when the writer tries to imitate spoken language. An illustration of a speech-related repetition is shown in 1-9.

1-9: Two speech-related repetitions (FR042-F)

```
time to (0.240) to really do (0.230) fieldwork for example this is this is rather
impossible
```

The **use and origin of repetitions** has long been debated about in the literature, but they are generally said to be used to hold the floor or to allow more planning time (e.g. Beliao & Lacheret 2013; Foster, Tonkyn & Wigglesworth 2000). In their in-depth article on repetitions, Clark and Wasow (1998) rightly point out that repeating words or sequences of words takes extra time and effort, and it is redundant in the discourse.

Repetitions are characterised by a specific – and complex – **internal structure**. Candea (2000) distinguishes the **repeatable** from the **repeated**. In a famous quote, Candea (*ibid*.:315) says that:

> Any repetition forms a block in speech that includes at least two elements: a first element that we will call the "*repeatable*" and a second element, identical to the first, that we will call the "*repeated*". It goes without saying that, in theory, any unit produced in speech is in principle a *repeatable,* and it is only the presence of a *repeated* immediately afterwards that makes this repeatable actually enter into the composition of a block that we call *a posteriori* a "repetition" [italics and underlining original][21]

The repeatable corresponds to the (sequence of) word(s) that is (are) originally uttered by the speaker, and the repeated refers to the second (and third etc.) utterance of the same word(s). When the repeated consists in two or more repetitions, she calls it "multiple". Henry and Pallaud (2004) use the terms FR. *répétition simple* when there is only one repeated, FR. *repetition double, repetition triple* etc. when there are two, three etc. repeated.

Besides, Candea (2000) stresses that the **repeatable and the repeated can either be adjacent** – Henry (2002) calls these "direct repetitions" (FR. *répétition directe*) – or be **separated by other (dis)fluency features**, such as pauses, but that these do not change the meaning of the repeatable – "associated repetitions" (FR. *répétitions associées*) in Henry's terms.

In their extension of the model of repairs from Levelt, the "commit-and-restore model of repeated words", Clark and Wasow (1998) went a step further and identified the following four stages in the structure of repetitions:

- The initial commitment to the constituent, which corresponds to Candea's (2000) repeatable (the first *I* in example 1-10);
- The suspension of speech (just after *I*);
- The hiatus in speaking – the material between the suspension and the resumption, which can be empty, filled by a filled pause (as in the illustration) or, most frequently, by an unfilled pause (Fathman 1980);
- The restart of the constituent that was suspended, which equals Candea's repeated (in the example, the repetition of the pronoun *I* after the filled pause).

[21] Original quote: [T]oute répétition forme un bloc dans la parole qui comporte au minimum deux éléments: un premier élément que nous appellerons le « *répétable* » et un deuxième élément, identique au premier, que nous appellerons le « *répété* ». Il va de soi qu'en théorie toute unité produite par la parole est en principe un *répétable* et ce n'est que la présence d'un *répété* immédiatement après qui fait que ce répétable va entrer effectivement dans la composition d'un bloc que nous appelons a posteriori une « répétition » (Candea 2000:315; italics and underlining original).

1-10: The four stages of repetitions (adapted from Clark & Wasow (1998: 220))

```
I uh I wouldn't be surprised that
```

In the repeatable and the repeated, Biber *et al*. (1999:1055; see also e.g. Fathman 1980; Gilquin 2008) argue that **it is more common for single words to be repeated** than for sequences of words or whole utterances. They also mention that repetition of parts of words is frequent. When they analysed the **nature of the word(s) in the repeatable**, they found that personal and possessive pronouns, as well as conjunctions show a strong tendency to occur in repeats. They conclude that "repeats and filled pauses, as hesitation phenomena, show parallel tendencies to co-occur with certain word classes". In another study (Gilquin 2008:139), and contrarily to Biber *et al*.'s (1999) results, prepositions have been shown to figure prominently in native speakers' and especially French learners' repetitions – the author also includes a comparative list of repetitions in the two speaker groups and shows that many are shared by NSs and NNSs. Overall, it appears that function words are repeated far more often than content words (e.g. Clark & Wasow 1998; Fox & Jasperson 1995; Maclay & Osgood 1959), arguably because "they tend to come first in major constituents" and because "they tend to be more accessible and easier to pronounce" (Clark & Wasow 1998:210).

As far as the **number of repetitions** is concerned, Biber *et al*. (1999:1055) write that "the likelihood of the repetition decreases sharply with the **number of words repeated**, so that the overwhelming majority of examples are of a single repeat (e.g. *the the*)" and that "[t]here are extremely few instances of three or more repeats (e.g. *the the the…*)". This claim is heavily supported by other researchers' findings. In his investigation of the Switchboard corpus, Kapatsinski (2004), for example, found that 79% of repetitions are one-word repetitions, 18% are two-word repetitions, and a small 3% are three-word repetitions. Gilquin (2008) also confirmed that the higher the number of repeated words, the less likely it was to occur, and she showed that this also held for learners of English.

In learner language, despite the intuitive feeling that they are an important part of fluency, repetitions have been found to have **no correlation with (dis)fluency judgements** (Kormos & Dénes 2004). They seem **not to be indicative of proficiency level** either (Baker-Smemoe *et al.* 2014; Iwashita *et al.* 2008; Rose 2013).

1.4.9 Truncated words

Truncated words (also referred to as *interrupted words*, *interruptions,* or *amorces de mots* in French) can be defined as "an interruption of morphemes in the course of enunciation"[22] (Pallaud 2002:79). They are words that are not uttered in their entirety because the speaker

[22] Original: "une interruption de morphèmes en cours d'énonciation" (Pallaud 2002:79).

has stopped at some point, but they need to be distinguished from apocopes, where the abbreviation is made voluntarily or for rhetorical purposes. A typical example of a truncation is shown in 1-11.

1-11: A truncation (FR002-S)

```
so er her fa= his father decided to take me by car
```

Truncated words have been reported to occur at a **frequency of 3 to 6 per thousand words**: they are thus far less frequent than other (dis)fluency features such as pauses (Pallaud 2002; Henry & Pallaud 2004). In learner language, this (dis)fluency phenomenon has been reported to be as frequent as 11 truncated words per thousand words (Gilquin 2008). The frequency of truncations has, however, been found to greatly vary depending on the speaker and researchers consequently concluded that they are likely to reflect speakers' individual speaking behaviour.

Henry and Pallaud (2004:204–205) distinguish so-called *list phenomena* ("phénomènes de listes") from *syntactic ruptures* ("rupture syntaxique") depending on whether what follows the truncation can be situated at the same syntactic location (list phenomena) or whether it does not belong to the same syntactic unit (ruptures). Their analysis of truncations resulted in a **three-tier typology of interrupted words**, which includes the following categories:

- List phenomena:
    - **Completed interruptions** ("*amorces complétées*"): the word that was interrupted is completed and has "the same syntactic place"[23]. Three patterns can be observed: the interrupted word may be completed (1) without repeating the *amorce*, (2) with the repetition of the *amorce* (in which case, the word is uttered in full after having been interrupted), or (3) with the repetition of the *amorce* and of other lexical elements that were uttered prior to the interruption. Completed interruptions are illustrated in Example 1-12.
    - **Modified interruptions** ("*amorces modifiées*"[24]): the word that was interrupted is abandoned but is replaced by another word at the same syntactic place, as in 1-13.

[23] This corresponds to an "entassement paradigmatique", or what Blanche-Benveniste (1997) calls "le piétinement syntaxique".

[24] *Interruptions corrigées* ("corrected interruptions") in Pallaud (2002).

- Rupture phenomena:
    - **Unfinished interruptions** ("*amorces laissées inachevées*"): the truncated word is abandoned, and what follows has a different syntactic function. An example of unfinished interruption is provided in 1-14.

1-12: Completed interruption with repetition of the truncation (Henry & Pallaud 2004:204)

c'est vrai que c'est pas **b- beau** d'associer les deux choses

1-13: Modified interruption (Henry & Pallaud 2004:205)

on va + attaquer l'autre **b- morceau** l'autre moitié du dos

1-14: Unfinished interruption (Henry & Pallaud 2004:205)

alors je vais euh faire un petite **diver-** on va diverger là pour expliquer ça euh au début

Henry and Pallaud's (Pallaud 2002; Pallaud & Henry 1995; Henry & Pallaud 2004) results showed that **completed truncations are the most frequent** (about two thirds of all truncations), before unfinished and modified interruptions. A similar categorisation has been made by Gilquin (2008): in her study, where she contrasts native and EFL learner's use of hesitation markers, she distinguishes **stutters** ("when the complete word comes immediately after the truncation"), **delays** ("when the complete word comes later in the utterance") and **repairs** ("when the complete word does not occur in the utterance at all"). While the first two seem to correspond to the category of completed interruptions in Henry and Pallaud's terms, the latter equals the category of modified and unfinished interruptions. Gilquin's results confirm that completed truncations account for the majority of truncations both in native and in learner speech, but they also show that, not only do **EFL learners overuse truncations**, but they also differ from native speakers in terms of type of truncation. While the proportion of stutters and delays is not statistically different in NS and NNS speech, **L2 learners use more repair truncations**. Overall, research findings do not support Levelt's (1989:481) proposal that "by interrupting a word, a speaker signals to the addressee that that word is an error" (by contrast with complete words that should be interpreted as correctly delivered).

A number of researchers (e.g. André & Tyne 2010) have pointed out that truncations may also favour **discursive collaboration**: the interlocutor, upon hearing the truncated word (which is often preceded by other (dis)fluency phenomena), may produce the completion him-/herself, thereby ensuring the syntactic completion of the utterance.

1.4.10 Foreign words

The term "foreign words" is closely related to the notion of code-switching. Code-switching can be briefly defined as the **alternating use of (at least) two languages** in the same

conversational event (Eldridge 1996; Greggio & Gil 2007): "foreign words" thus refer to the specific items uttered in another language than the main language of the discourse.

Although it used to be regarded as an avoidance strategy – that is, an indicator of disfluency –, the use of foreign words by language learners has come to be seen as a **highly purposeful** phenomenon in spoken interactions. Research suggests that foreign words have an important role in facilitating interaction, as well as in facilitating foreign language learning, and have a multifarious number of more specific functions. They may, for example, be used to **elicit an equivalent meaning** in the target language, to **clarify a message** that has already been transmitted in one code but not understood, to provide a "**stopgap**" while the word(s) in the target language is (are) being retrieved; as a **"resource expansion" strategy** etc. (Eldridge 1996; Liebscher & Dailey–O'Cain 2005; Poulisse 1987; Sert 2005).

Nacey and Graedler (2013) explored the use of foreign words in the Norwegian component of LINDSEI as a **compensation strategy**. They argued that foreign words are a highly effective device that positively contributes to a smooth flow of conversation. They, however, suggested that the fact that the interviewers understood the learners' mother tongue was one of the probable reasons why code-switching was so effective in the Norwegian component of LINDSEI (see also De Cock 2015a; 2017a). The authors further described three ways in which foreign words may be used: they may be inserted without any modification (i.e. "code switching", *cf*. Example 1-15), modified to follow the rules of the target language (i.e. "foreignising", as in 1-16), or literally translated from the L1 (i.e. used as calque, as shown in 1-17).

1-15: Code-switching (from Nacey & Graedler 2013:348)

> my father helped out with the **stabbur** (Norwegian storage house on pillars)

1-16: Foreignising (from Nacey & Graedler 2013:348)

> they were just swimming around really fast like the **stims** of fish (Norwegian word for *shoal*)

1-17: Calque (from Nacey & Graedler 2013:348)

> if you become a teacher in Norway you normally have end up in (em) **. children's schools** (calque from Norwegian *barneskoler*)

More recently, De Cock (2015a; 2015b; 2017a; 2017b) discussed the use of foreign words in other components of LINDSEI[25]. The studies reveal that the frequency of foreign words varies quite markedly across the various components, with the French- and German-speaking learners using over twice as many foreign words as the Spanish-speaking learners. The author points out that these results are probably related to the fact that the interviewer in the French

[25] In 2015 and 2017a: the Dutch, French, German, Italian and Spanish components. In 2017b: the Bulgarian, Chinese, Greek, Japanese, Polish, and Swedish components.

and German components of LINDSEI shares the L1 of the learners, which was not the case for the other components analysed in her studies. De Cock highlighted that the foreign words come overwhelmingly, but not exclusively, from the learners' L1, and that they feature in both the interviewers' and the learner interviewees' contributions. She identified the following four functional categories of foreign words in learner speech:

- **lexical bridges,** which help learners bridge vocabulary or lexical gaps that are either unknown or temporarily inaccessible. In 1-18, for example, the learner uses the Chinese *fan shu* because the English equivalent (*sweet potato*) is temporarily inaccessible;
- **cultural and institutional bridges,** which denote aspects of the education system, folklore etc. typical of the learners' country (Example 1-19);
- **pragmatic and discourse bridges,** which fulfil basic pragmatic or discourse functions in the learners' mother tongue and are largely spontaneous. In Example 1-20, the learner for example uses the word *enfin* ("well") before correcting himself;
- in **direct speech reporting,** as illustrated in Example 1-21, or in **metalinguistic discussions** (Example 1-22).

1-18: Foreign words as lexical bridge (from De Cock 2017; LINDSEI_CH)

> em with their own things what what we call er **<foreign> fan shu </foreign>** in Chinese <laughs> I don't know what

1-19: Foreign word as cultural/institutional bridge (from De Cock 2017; LINDSEI_GR)

> I liked erm the . **<foreign> la Tour Eiffel </foreign>**

1-20: Foreign word as discourse/pragmatic bridge (from De Cock 2015; LINDSEI_FR)

> because we are (er) two: **<foreign> enfin </foreign>** we we are three: children in my family and (er) two of us . are studying here so (er) they

1-21: Foreign word in direct speech reporting (from De Cock 2017; LINDSEI_GR)

> and when I returned back to: my school . er the first of eh primary school . er I was eh . ta= . I was saying some Spanish words like . **<foreign> salud </foreign>**

1-22: Foreign words in metalinguistic discussion (from De Cock 2015; LINDSEI_SP)

> and: er and also because I I like how Mexicans li= th= the way of of speaking of Mexican people erm and: one of the most interesting things that I found er in seeing the film was the accent and and also the vocabulary for instance words such as beer **<foreign> cerveza </foreign>** in Spanish they they call it **<foreign> chela </foreign>** I think </B>

De Cock (2017b) showed that there are **differences in terms of preferred functional category across components**: whilst cultural/institutional bridges are the preferred functional category in the French, German and Spanish components of LINDSEI, pragmatic/discourse bridges are predominant in the Dutch component and lexical bridges in

the Italian component. From the point of view of (dis)fluency, she also highlighted that learners tend to use filled and unfilled pauses, as well as explicit acknowledgements of lack of knowledge (e.g. *I don't know how you say it in English*), approximations and discourse markers (e.g. *sort of*, *kind of*) with lexical bridges.

There seems not to be a strong relationship between proficiency and the use of foreign words. Eldridge's (1996:304) study on code-switching in Turkish learners of English, for example, has shown that it **may not be correct to assume that the greater the proficiency, the fewer the foreign words**. It would, however, be interesting to examine the use of the four functional categories of foreign words identified by De Cock (2015a; 2015b; 2017a; 2017b) across proficiency levels. More specifically, because lexical bridges are more closely related to automatisation processes and ease of retrieval, they may reveal a different pattern than, e.g., cultural/institutional bridges. Moreover, while learners arguably have the exclusivity of lexical and pragmatic bridges, it does not seem impossible that the other two categories may come up in native speech too.

1.4.11 Lengthenings

Lengthenings can be basically defined as the **stretching out of a sound** or a syllable for longer than typical. A technical definition is enunciated by Campione and Véronis (2004:120), who write that:

> hesitation lengthenings are characterized by a continuous vowel of longer than normal duration, of constant vocal quality and [...] which is associated with a fundamental frequency (Fo) that is flat or very slightly descending[26]

Although this phenomenon may be referred to by several terms, including **drawls**, **elongations**, **prolongations** of sounds, or **sound stretches** (e.g. Chambers 1997; Eklund 2001; Götz 2013a; Rohr 2017), this thesis adopts the term *lengthening*.

Lengthenings may actually occur purely as a **result of phonological processes** – stressed syllables are for example longer than unstressed syllables (Fokes & Bond 1989; Oller 1973) – or they may carry a **pragmatic effect** such as marking emphasis or turn-taking (Du Bois *et al.* 1993; Kohler 2006; Ladd 1996). As opposed to these two functional uses, lengthenings are also claimed to be a **hesitation phenomenon** regularly used by native and non-native speakers to buy additional planning time. Those lengthenings are characterised by their "unnaturalness" (Grosjean 1980c:42) and have been argued to signal "a delay already in process" (Clark & Fox Tree 2002:86) rather than the initiation of delays. Incidentally, Eklund

[26] Original quote: "[les] allongements d'hésitation [...] se caractérisent par une voyelle continue de durée très supérieure à la normale, de qualité vocalique constante et [...] associée à une fréquence fondamentale constante (Fo) plate ou très légèrement descendante" (Campione & Véronis 2004:120).

(2001:5) has underlined that lengthenings and filled pauses have in common that "they both signal hesitation by means of vocalization and duration". A regular pitfall of studies into lengthenings is the lack of measurement of the lengthened segments (generally due to the unavailability of time alignment): in fact, it is not rare that lengthenings are detected perceptively.

To date, lengthenings remain the poor relation of (dis)fluency features in L2 and contrastive L1-L2 studies. As underlined by Campione and Véronis (2005:43; see also Duez 1998), this is probably due to the fact that most (phonetic) studies used to be based on "laboratory speech" at the expense of spontaneous oral speech. This is also in part due to the (technical) difficulties associated with the analysis of lengthenings, which should ideally aim to take into account "not only variability between speakers (and their mean length of syllable) but also the phonological weight of the syllable, the stress on the syllable, and its position within the intonational contour" (Rohr 2017:332).

Previous investigations of lengthenings have shown that they are more **common** than most other (dis)fluency features in native speech and are outnumbered only by filled and unfilled pauses (Eklund 2000; Eklund & Shriberg 1998; Grosjean 1980c). They have been reported to occur with a frequency of about 0.3 per hundred words in native English conversation (Gilquin 2008), and a frequency of c. 1.3 per hundred words in native Swedish dialogues (Eklund 2001). In **learner language**, they are **even more endemic** with rates approaching 2 lengthenings per hundred words (Gilquin 2008; Rohr 2017).

In her survey of lengthenings (which she terms "prolongations") in learner and native speech, Rohr (2017) showed that, although lengthenings may theoretically apply to words of any length, the majority occur on **monosyllabic words** (both in NS and in NNS speech), with only a very small proportion on disyllabic words. Furthermore, although lengthenings may occur in syllable-initial position (as in *a:nd* or *i:f*), in medial position (e.g. *bu:t*) or in syllable-final position (*she:*, *so:*), lengthenings are not evenly distributed within words. A ratio of 30-20-50 for initial, medial and final position has for example been found by Eklund (Eklund 2000; Eklund 2001; Eklund & Shriberg 1998). Rohr's (2017) results do support the **prevalence of syllable-final vowel lengthening** in both native and learner speech, but her data revealed that native speakers rarely prolong medially (only in 6% of the cases), whereas intermediate learners of English prolong in initial and medial positions more regularly and relatively equally (19% and 18%, respectively). These differences, the author argues, are likely to contribute to listeners' intuitive perception of differences between native and non-native use of lengthenings.

It has also been reported that native speakers show a marked **preference for lengthening function words over content words** (c. 98% vs. 2%, respectively – Rohr (2017)). While this is also the case in learner speech, the proportion of lengthened content words is much higher (c. 84% vs. 16%). Among the frequently lengthened function words, the **articles *a* and *the*** have crystallised a lot of attention. Those two articles have the particularity of having a

"weak" and a "strong" lengthened form (*the:* and *the[i:]; a:* and *a[ei]*). Clark and Wasow (1998) found that, in the London-Lund corpus, *the[i:]* – which, they claim, disrupts continuity – occurs with a much higher frequency than *the:*. Gilquin (2008) further found that native speakers and learners differ in their use of the weak forms *the:* and *a:*, which the learners overuse. She advances that (*ibid*.:136) "[i]t is probably not a coincidence that these two forms correspond to the normal mode of lengthening for the French articles, namely the lengthening of the final vowel ([ə] for the definite article *le* and [œ̃] for the indefinite article *un*)"[27] and concludes that "[l]earners seem to shy away from the special drawls (those that use a different sound) and show a predilection for the mode of lengthening they are used to in their mother tongue".

1.4.12 Discourse markers

Since the 1970s, analyses of discourse markers (DMs) have surged in the literature. Discourse markers are linguistic elements such as *well* or *you know* "which do not contribute to the propositional content of the utterance which they modify [and which] are frequent in conversation, where they express the speaker's attitudes to the addressee, negotiate background assumptions, express emotions and contribute to coherence" (Aijmer & Simon-Vandenbergen 2003:1123). A typical example of a discourse marker is shown in 1-23.

1-23: A discourse marker (from Gilquin 2016:224)

```
we play it like every day . four or five hours
```

DMs have been analysed from a variety of perspectives, such as marking **discourse coherence** (Halliday & Hasan 1976), or signalling **sequential relationship** (Fraser 1990; 1999). They have also been approached from a Relevance Theory perspective (Schiffrin 1987). As underlined by Fung and Carter (2007; also Jucker & Ziv 1998), the various viewpoints on discourse markers, the range of analytical categories, and the difficulties in accounting for them adequately in theoretical terms are reflected in the **multiplicity of terminology**. The variety of labels includes: "[sentence] connectives" (Degand 2000; Halliday & Hasan 1976; Ozono & Ito 2003), "smallwords" (Hasselgren 2005), "modal particles" (Aijmer 1997), "discourse particles" (Aijmer & Simon-Vandenbergen 2003; Schourup 1985), "pragmatic markers" (Watts 1988), "pragmatic expressions" (Erman 1987), "interactive discourse markers" (Povolná 2009). "Discourse markers" (Fuller 2003; Fung & Carter 2007; Romero-Trillo 2002) is, according to Schourup (1999:228), merely "the most popular" among the large

[27] It also corresponds to the typical French filled pause *euh.*

panel of labels[28]. Incidentally, Neary-Sundquist (2013) notes that some of the terminological differences stem from whether the focus of the study is on speech (in which case, "discourse marker" is generally preferred) or on writing (where "connective" is more widespread). In this thesis, I have adopted the term "discourse markers". Note that a subcategory of discourse markers, namely conjunctions, are discussed separately in Section 1.4.13.

Another consequence of the profusion of approaches to discourse markers is that it is "unfeasible to draw up one **unified definition of DMs** that could apply to all studies in the field" (Buysse 2007:80). Attempts at definitions can be found in (*inter alia*) Schiffrin (1987) or Fraser (1999), but here I will merely list the **features shared by most discourse markers**, as discussed by Schourup (1999:230–234). The first three are "all frequently taken together to be necessary attributes of DMs", the remaining features are "less consistently regarded as criterial for DM status" (*ibid*.:232).

- **Connectivity**: discourse markers relate units of discourse. Schourup stresses that this connectivity can be understood in different ways (e.g. relating textual units, or relating an utterance to the wider context of utterance) (*ibid*.:231);
- **Optionality**: DMs can be removed from their host utterance without altering its grammaticality;
- **Non-truth-conditionality**: DMs do not contribute to the truth-conditions of the proposition expressed by an utterance;
- **Weak clause association**: discourse markers usually fall outside the syntactic structure, or are loosely attached to it;
- **Initiality**: DMs prototypically occur at the beginning of utterances;
- **Orality**: DMs are typical of spoken discourse;
- **Multi-categoriality**: DMs are heterogeneous with respect to syntactic class.

In her account of native and learner discourse markers, Müller (2005:4–8) also lists seven core characteristics of discourse markers, four of which overlap with Schourup's list (namely optionality, initiality, orality and multi-categoriality), but she also includes:

- **Phonological features**: discourse markers have a range of prosodic contours;

[28] Note also that discourse markers and filled pauses are sometimes referred to in the literature with the umbrella term "fillers". Filled pauses are said to be "non-lexical", and discourse markers are "lexical" and have been termed "lexical fillers", "lexicalised filled pauses" or "verbal fillers" (Rohr 2017; Rose 1998; Stenström 1994).

- **Lack of semantic content**: DMs are not totally void of meaning, but do not add to the propositional content of the utterance;

- **Multifunctionality**: they generally fulfil more than one function.

Although there have been studies on larger or smaller sets of discourse markers (e.g. Crible 2017a; Denke 2009; Müller 2005), previous research has mainly focused on **individual markers:** *I think* (Aijmer 1997)*, so* (Buysse 2007), *like* (Fox Tree 2006), and *well* (Schourup 2001) have probably received the most scholarly attention so far (Buysse 2015:59). Discourse markers have been analysed in different languages, **predominantly in English** (e.g. Müller 2005), but also in Spanish (Campillos Llanos & González Gómez 2014), Hungarian (Dér & Markó 2010), or Bulgarian (Fielder 2008). There is also some evidence from the contrastive literature that the use of discourse markers might be **language-specific**. The French language, for example, appears to prompt a higher rate of DMs (Crible 2017a; Vinay & Darbelnet 1995).

From the point of view of (dis)fluency, discourse markers have sometimes been (and, to some extent, still are) **stigmatised** as being "markers of unclear thinking, lack of confidence, inadequate social skills, and a range of others [sic] undesirable characteristics" (Crystal 1988:47), in other words, disfluencies. An increasing body of researchers now agrees that discourse markers have **discursive and pragmatic functions** that facilitate the listener's understanding: DMs have, among others, been shown to function as structuring devices, fillers, repair markers, and interpersonal markers (Aijmer 2011; Denke 2009; Erman 1987; 2001; Jucker & Ziv 1998:1; Müller 2005). Buysse (2015), for example, showed that, both in learner and native speaker speech, one of the most frequent uses of the discourse marker *well* is to enable the speaker to change an already-produced content or structure, as in Example 1-24, where "the interviewee interrupts her utterance and reformulates it instantaneously: the painter does not continue a painting but starts a new one" (Buysse 2015:73)

1-24: Speech management function of 'well' (from Buysse 2015:73; Dutch component of LINDSEI; DU023)

> <B> then she sits down again and he continues **well** he makes a new portrait I think </B>

While a considerable body of research has been devoted to the study of discourse markers in native languages, the amount of research in learner languages is more limited. Moreover, whereas a number of studies focus on single discourse markers (e.g. Buysse 2009; Fox Tree & Schrock 2002; Schourup 2001), others adopt a much wider perspective (Crible 2017a; Gilquin & Granger 2015). Previous research into **native and learner use** of discourse markers has yielded an overall underuse of discourse markers in learner language (Buysse 2010; Gilquin 2008; Gilquin & Granger 2015; Müller 2005). What also emerges from these previous studies is that learners use a smaller range of discourse markers. As explained by Aijmer (2004:182; my emphasis), "[t]he major difference between learners and native speakers has to do with

the **frequency of *individual* markers**": some specific discourse markers may, in fact, be overused by learners. Müller's (2005) study, for example, showed that, while DMs are underused by German-speaking learners of English, this is not the case for the DM *well* (see also e.g. Buysse 2015; Gilquin 2008). Moreover, Gilquin and Granger (2015; see also Aijmer 2011) convincingly demonstrated that the mother tongue of the learners as well as idiolectal preferences of L1 and L2 speakers are also highly relevant in analyses of discourse markers.

Differences between learners and native speakers are, however, not limited to the quantitative use of individual (or of a set of) markers: the **qualitative use** can also be particularly indicative. For example, in an investigation of the DM *well*, Aijmer (2011:231) found that, above all, learners use *well* "as a fluency device to cope with speech management problems but underuse it for attitudinal purposes". This, she hypothesises, indicates that, even at an advanced level, learners remain unfamiliar with interpersonal communication strategies in the target language. In a similar vein, Gilquin and Granger (2015; see also Gilquin 2016) examined the DM *you know* in the speech of French-speaking and Polish learners of English, as compared to native speakers of English. They focussed more particularly on the placement of the DM by making a distinction between interruptive and non-interruptive uses of *you know* (Examples 1-25 and 1-26, respectively). The authors found that, although French-speaking learners underuse this DM, the breakdown of interruptive vs. non-interruptive uses of *you know* is similar to that found for the NSs. By contrast, Polish learners, who overuse the DM, tend to use it in interruptive structures. The qualitative use of *you know* by Polish learners thus marks them as "particularly disfluent and non-native-like, contrary to what the quantitative analysis would have suggested" (*ibid*.:433).

1-25: Interruptive use of the DM 'you know' (from Gilquin & Granger 2015:433)

```
if we you know make some= something legal
```

1-26: Non-interruptive use of the DM 'you know' (from Gilquin & Granger 2015:433)

```
it's just like a curtain you know so you've gotta change it
```

In learner speech, the frequency of discourse markers is positively **correlated with proficiency level or perceived fluency level** (Hasselgren 2002; Neary-Sundquist 2013). The **range** of DMs remains **relatively limited** even at the most advanced stage(s) of proficiency (Buysse 2010; Gilquin 2008; Neary-Sundquist 2014) and learners "can be presumed to stick to those pragmatic markers they are most familiar with" (Buysse 2015:84). In this respect, however, it is noteworthy that Rose (2000; Rose & Ng 2001) offered some evidence about the benefit of an instructional approach in interlanguage pragmatics.

Hasselgren (2002) reported that there is a **correlation between the use of "smallwords"** (i.e. discourse markers) **and fluency**: she provides evidence that the more smallwords are used in a native-like way in learner speech, the greater the fluency. Likewise, other studies have shown that, owing to their **discourse and pragmatic functions** (Müller 2005; Neary-Sundquist 2013) in spoken discourse, discourse markers **enhance learner fluency** (Hasselgren

2005; Towell, Hawkins & Bazergui 1996) and the **coherence** (Halliday & Hasan 1976; Schiffrin 1987) of an interaction.

1.4.13 Conjunctions

The class of discourse markers is generally defined functionally, but the linguistic elements considered as discourse markers may belong to several categories. One of the grammatical categories of DMs is that of conjunctions[29]. Like other categories of discourse markers, conjunctions mark logical relationships between words, phrases, clauses and sentences (Carter & McCarthy 2006). Neary-Sundquist (2013) argues that looking at different types of discourse markers together (i.e. "core" discourse markers like *you know* and *I mean* on the one hand, and conjunctions on the other) might hide important differences in their use, and that **discourse markers and conjunctions should be investigated separately**. In her study, Neary-Sundquist provided some evidence that discourse markers and conjunctions do not follow exactly the same acquisition pattern: while both the frequency of discourse markers and that of conjunctions rise with proficiency level, this rise is much sharper for discourse markers than for conjunctions. He therefore suggests that discourse markers might be more useful than conjunctions in discriminating between proficiency levels. Similarly, Schiffrin (1987; see also Shriberg 1994) also draws a distinction between "core" discourse markers such as *well,* and items such as *and, but,* and *so.* Bearing this in mind, in this thesis, conjunctions are considered a separate category.

Three conjunctions in particular seem to have crystallised researchers' attention, namely ***and***, ***so***[30], **and** ***but***. These three conjunctions have been shown to be particularly pervasive in both native English and native French (Bestgen 1998; Buysse 2007; 2009; Fraser 2005; Fung & Carter 2007). They are also very polysemous, with a core meaning which can be nuanced depending on the context (Crible 2017b; Fung & Carter 2007). From a functional point of view, Fraser (2005) argues that *and* is an elaborative marker, *so* an inferential marker, and *but* a contrastive marker. For Crible (*ibidem*), *and* is a generic coordination, *but* a contrastive conjunction, and *so* indicates epistemic consequence or logical effect. All three items are, however, characterised by a high degree of multifunctionality.

Despite Pawley and Syder's definition of fluency as "the native speaker's ability to produce fluent stretches of spontaneous *connected* discourse" (Pawley & Syder 1983:191 my

[29] There is considerable debate about the use of the terms "conjunction", "connector", "coordinator", connective", "conjunct" etc. See Biber *et al*. (1999:79–80); Cosme (2007:238–250); Quirk *et al*. (1995:442; 918–935); or Neary-Sundquist (2013:112–113) for a discussion. In this thesis, the term "conjunction" is adopted.

[30] Although *so* differs in some respect from *and* and *but,* it may be regarded as a "marginal member" of the category of conjunctions (Huddleston & Pullum 2002:1319; see also Quirk et al. 1995). For reasons of clarity, *so* will thus be referred to here with the term "conjunction".

emphasis; see also Fillmore 1979), to date, only a limited number of studies has analysed conjunctions from the point of view of (dis)fluency. Yet, *and* has been shown to be a "marker of continuity" (Bestgen 1998:758; also Schiffrin 1987) whose function is to signal to the hearer that two or more utterances are highly related. *And* can thus create **chunks of closely related events**. Additionally, the conjunction may also be used "to connect two sentences that lack coherence" (Bestgen 1998: 757). In such cases, *and* appears to be the **trace of some production difficulty** (see Altenberg 1987; Jisa 1984; Peterson & McCabe 1987; Rose 1998; Spooren 1997). The conjunction *but* can link utterances by expressing contrast, denial, or dissonance between utterances (Biber *et al.* 1999; Fraser 1988; Schiffrin 1987). Schiffrin (*ibid*.:164) also notes that *but* is regularly used as a "sequential conjunction" with repetitions or restarts when a speaker wishes to return to a prior concern. With respect to *so*, most prior research has focused on *so* for marking result (Schiffrin 1987) or inference (Blakemore 1988; 2002; Fraser 1999) between utterances. *So*, especially when it is used in initial position, may also function as a "**marker of cohesion**" (Bolden 2009; Howe 1991), as a "topic developer" or "topic sequencer" (Johnson 2002), as a "marker of elaboration" (Buysse 2009), or as a marker of "emergence from incipiency" (Bolden 2006). In learner speech, *so* has been shown to have as many as ten functions (Buysse 2012). An interesting property of *so* is that, unlike *but*, it can combine with the conjunction *and* (Huddleston & Pullum 2002), as illustrated in 1-27.

1-27: Combination of 'and' and 'so' (from Huddleston & Pullum 2002: 1319)

this may make the task seem easier **and so** increase self-confidence

Moreover, it is noteworthy that conjunctions are **generally preceded by an unfilled pause**, or **followed by a filled pause** (Hansson 1999; Rose 1998), with which they may be cliciticised (*cf*. Section 1.4.5 and Example 1-1).

Lastly, there is some evidence that *and, so,* and *but* are **positively related to proficiency and perceived fluency level** (Hasselgren 2002; Neary-Sundquist 2013). However, some researchers have pointed out that less fluent speakers seem to use these three conjunctions more frequently than more fluent learners. This, Buysse (2014) argues, might be due to the interplay between three factors. First, in a desire to come across as proficient speakers, learners use an abundance of conjunctions. Second, less fluent speakers have "a more limited array of markers with which they feel on safe ground because these are highly frequent in the target language" (*ibid.*: 32): they thus tend to stick to these highly frequent conjunctions (see also Hasselgren 1994). Lastly, learners may use markers that functionally resemble similar items in their mother tongue.

All in all, what emerges from previous research is that *and, so,* and *but*, which are particularly endemic in speech, contribute to the cohesion of the discourse by connecting utterances. Yet, the patterning of conjunctions with filled and unfilled pauses also seems to indicate that, to some extent, they are also the trace of ongoing cognitive processes.

1.5 CONCLUSION

In this chapter, fluency and disfluency have been situated in the wider context of L1 and L2 research. The chapter has provided an overview of how these concepts came to life and evolved through time. It also contrasted learner and native speaker (dis)fluency, before exploring the main findings with respect to 14 (dis)fluency features. The overall impression that transpires from this literature review is that, although a great deal has already been written about (dis)fluency and (some) (dis)fluency features, there are still a number of weaknesses and drawbacks in the field which deserve further attention.

First, it is never stressed enough that so-called disfluencies are the **normal accompaniment of both native and non-native speech**. In many cases, they fly away unnoticed, and only a detailed transcription can reveal them: they "are simply traces of the act of producing speech in real time and therefore an integral part of spoken language" (Temple 1992:29). (Dis)fluency features are in fact the **two sides of a coin**: whereas they may indicate planning difficulties (i.e. cognitive disfluency), they may also be used strategically by the speaker and exploited by the listener (i.e. they may increase perceived fluency) (Clark & Wasow 1998). Fluency can thus not simply be equated with the absence of disfluencies. Specifically, more research needs to be carried out to **explain the way in which native speakers violate the idealised rules of spontaneous speech while the discourse itself is still (most generally) perceived as fluent**.

Several factors that affect native and non-native (dis)fluency have been reviewed. Pre-**planning time**, high level of **interaction**, and clear **task structure** have been found to have a beneficial effect on L1 and L2 (dis)fluency (Cucchiarini, Strik & Boves 2002; Foster & Skehan 2009; Mehnert 1998; Riggenbach 1989). **Cross-linguistic differences** have also been pointed out, with, e.g., French L1 speakers pausing less frequently but longer on average than English L1 speakers (Grosjean & Deschamps 1975). Whilst such results definitely contribute to the richness of the field, they also indicate that straightforward comparisons between studies should be approached with care.

Furthermore, previous research indicates that some aspects of (dis)fluency are in fact attributable to each person's **individual and idiosyncratic speech characteristics**. On the one hand, native speakers and learners have been shown to have highly variable hesitation patterns, and several (dis)fluency profiles have been identified among a supposedly homogenous group of German-speaking learners of English and British English native speakers (Götz 2013a; de Leeuw 2007). On the other hand, it has been proved that several (dis)fluency characteristics of a learner in his/her L2 are strongly related to those in his/her L1, especially the temporal variables (De Jong *et al.* 2015; Segalowitz 2010). All of these results suggest that (1) if possible, the learners' L1 and L2 (dis)fluency should be considered in conjunction, and that (2) it is very important to take individual differences into account as (dis)fluency is actually multi-faceted. In this thesis, due to the nature of the data, it will

unfortunately not be possible to take the learner's L1 speech habits into account, but individual differences and (dis)fluency patterns will be taken into account.

Lastly, although the field of L1 and L2 (dis)fluency research has given rise to many insightful – contrastive, cross-linguistic, longitudinal etc. – studies, it has been plagued by a lack of agreement with respect to the elements contributing to (dis)fluency, their terminology, definition and measurement. In particular, considerable research remains to be done to fully gauge the extent to which **different measurements of the same phenomenon** affect subsequent research findings. A case in point is the lower threshold of unfilled pauses, or the measurement of frequencies per minute or per hundred words. These aspects obviously extend beyond the scope of this thesis (but see e.g. Campione & Véronis 2002; De Jong & Bosker 2013; Dumont 2017a; Kowal, Wiese & O'Connell 1983).

More generally, to date, little research has tried to determine the reason why a speaker uttered a particular (dis)fluency feature. This is a very hard, if not impossible enterprise especially because investigations are usually carried out *a posteriori* from a corpus of recorded speech (Corley & Stewart 2008). I believe that much insight could be gained by adopting a **multidisciplinary approach**, for example, by combining findings from LCR and SLA with those from psycholinguistics, speech pathology, or even neurology.

Against this backdrop, the next chapter explores the contribution of spoken corpora to (dis)fluency research.

Chapter 2 THE CONTRIBUTION OF SPOKEN (LEARNER) CORPORA TO (DIS)FLUENCY ANALYSIS

A spoken language corpus is a corpus consisting of recordings of speech which are accessible in computer readable form, and which are transcribed orthographically, or into a recognised phonetic or phonemic notation.

Sinclair (1996:28)

Chapter 2 is a corollary to Chapter 1 and focusses on the data and methods used in (dis)fluency research. The first three sections provide an introduction to spoken corpora, i.e. collections of texts that can be used to measure fluency "on the part of the speaker" (Götz 2013a:4). The fourth section then considers fluency assessment – "fluency on the part of the listener" (*ibidem*) – and gives an overview of fluency assessment grids and criteria.

In the present chapter, I discuss the following questions: What is a spoken corpus (Section 2.1)? What are the specificities of spoken learner corpora and how do these characteristics relate to (dis)fluency research (Section 2.2)? What are the pitfalls and untapped potentials of spoken corpora in the frame of (dis)fluency research (Section 2.3)? And, lastly, how are spoken corpora used in the domain of language testing (Section 2.4)?

2.1 SPOKEN CORPORA

2.1.1 Well er… what is a learner/native spoken corpus?

Linguistic research makes ample use of data produced by the users of a given language. Over the years, the nature and utilisation of the data has slowly shifted from made-up examples that were mostly used as a way of illustration to the exploitation of authentic and longer excerpts as a basis for large-scale analyses. One type of resource linguists use is corpora. McEnery and colleagues define a corpus as "a collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety" (McEnery, Xiao & Tono 2006:5). While the specific language variety a **native corpus** seeks to be representative of is the language produced by native (L1) speakers, that is, speakers expressing themselves in their mother tongue, a **learner corpus**

aims to be representative of the interlanguage of second or foreign language (L2) speakers. Representativity is ensured by the systematic collection of texts (in the sense of linguistic productions) that is governed by a number of criteria such as the medium of communication (written or spoken), the status of the writers/speakers, the target language etc.

The main criterion that is used to distinguish corpora is the **medium of communication**. While most corpora to date are based on written data such as essays or newspaper articles, some corpora rather focus on, or at least include, spoken data from telephone conversations or radio broadcast for example. The **number of spoken corpora** is slowly increasing, but is still **very low** compared to the wealth of written corpora. As underlined by Ballier and Martin (2015:107), "[o]ne reason for this scarcity is that spoken corpora are more costly (in terms of money, time and technology) to collect and annotate". One of the most famous corpora of English, the *British National Corpus* (BNC), for example, includes only 10 per cent of spoken language.

The data in spoken corpora typically comes in the form of **written transcriptions of spoken discourse** that was previously recorded (see also Section 2.3.1 for more details on written transcriptions as well as on the relationship between audio recordings and transcriptions). A distinction is usually made between "**mute**" and "speech" spoken corpora (Gilquin 2015). The former only contain transcriptions and can be illustrated by the spoken component of the *British National Corpus* (BNC-Spoken[31]) or the *Louvain International Database of Spoken English* (LINDSEI). In "**speech**" spoken corpora, besides the written transcriptions, the corresponding sound files are also made available (e.g. in the form of .wav files). The *Michigan Corpus of Academic Spoken English* (MICASE), the *Santa Barbara Corpus of Spoken American English* or the *LeaP* corpus are examples of such corpora. In some speech corpora, the sound files have been aligned with the written transcription through time alignment, so that it is possible to listen to small portions of the actual speech by playing the sound file (Ballier & Martin 2015:110). Besides, researchers have steadily been trying to collect **multimodal corpora** (e.g. the *Insight Interaction corpus*[32] and the *IFA Dialog Video corpus*[33]) that bring together resources made up of sound files, transcripts and video recordings (that are ideally also time aligned with sound and transcript).

While written transcriptions of oral data may be treated like written texts and queried with tools such as *WordSmith Tools*, time aligned corpus data require more **specific tools**, such as *Praat* or *EXMARaLDA*, which can be used to both generate a score (or "partiture") with the transcription and represent the acoustic data through a spectrogram, with plotted

[31] Note, however, that (the majority of) the audio recordings from the BNC-Spoken have recently been released (*cf*. BNCweb; http://corpora.lancs.ac.uk/BNCweb/; last accessed 10/03/2018).

[32] https://www.arts.kuleuven.be/midi/corpora-tools/insight-interaction-corpus (last accessed 24/04/2017).

[33] http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFADVcorpus/ (last accessed 24/04/2017).

fundamental frequencies or intensity curves for example, thereby allowing the researcher to investigate both discourse and acoustic features.

2.1.2 The advent, and uses, of spoken corpora

The first notable corpus of spoken English that was made available for research purposes, the **London-Lund Corpus of Spoken English** (Svartvik 1990), is a testimony to the birth of spoken corpus linguistics. The corpus consists of 100 texts of 5,000 words of spoken British English (i.e. 500,000 words in total) collected between 1959 and 1990. Although the fruit of great efforts (especially in the level of detail of the transcripts), the corpus also shows some major drawbacks, such as the unavailability of the original sound recordings (see Campoy & Luzón 2007).

Following the development of such (still relatively small) spoken corpora, the 80s and 90s witnessed the birth of so-called "**mega-corpora**" such as the COBUILD (the *Collins Birmingham University International Language Database*), the BNC, the Switchboard or the CANCODE (the *Cambridge and Nottingham Corpus of Discourse in English*). These mega-corpora (which may contain both spoken and written discourse) are much larger in size than the first generation corpora – 20 million words for the spoken component of COBUILD, nine million words for the spoken BNC, five million words for the CANCODE etc. – and they contain data from a variety of settings, sociolinguistic contexts and spoken genres. Although many languages are represented in spoken corpora, corpora related to English by far outnumber those in other languages.

Spoken corpora that have been collected more recently (or corpora containing a spoken section) tend to be somewhat smaller in size (as compared to the millions of words of mega-corpora), but also more specific. They may focus on (Campoy & Luzón 2007):

- **national varieties:** the *International Corpus of English*, the *Santa Barbara Corpus* etc.;
- **dialectal varieties:** the *Limerick Corpus of Irish English* or the *Freiburg English Dialect Corpus*;
- **specific time-spans:** e.g. the *Diachronic Corpus of Present-Day Spoken English;*
- **age-defined categories of speakers**: the *Child Language Data Exchange System* (CHILDES) for children; the *Bergen Corpus Of London Teenage Language* (COLT) for teenagers; the *Multimedia Adult ESL Learner Corpus* (MAELC) for adults etc.;
- **specific text types and genres or domains:** the *Michigan Corpus of Academic Spoken English* (MICASE) and the *British Academic Spoken English Corpus* (BASE) for academic and professional discourse;

- native speaker and **non-native speaker** discourse: the *Louvain International Database of Spoken English Interlanguage* (LINDSEI), the *Louvain Corpus of English Conversation* (LOCNEC), the Japanese learner corpus NICT JLE; the *Vienna-Oxford International Corpus of English* (VOICE – English as a Lingua Franca).

Moreover, famous publishers also provided incentive for the collection of new spoken corpora, like the *Longman Spoken Corpus* or the *Cambridge Spoken Learner Corpus*.

Spoken corpora have been **used for a variety of purposes**, depending to a great extent on the criteria used for their compilation as well as the availability (or not) of audio files and of some type of annotation of the data. One of the primary (and earliest) purposes of spoken corpora is obviously to gain **better insights into the nature of orality** by examining its specificities: spoken grammar or hesitations have, for example, been extendedly analysed (see e.g. Leech 2000). Research has also turned to the pragmatics of spoken language (e.g. discourse markers or conversational acts), as well as to acoustic properties such as prosody, intonation or phonetics (e.g. Gut 2009). **Socio-linguistic** investigations into how spoken language varies depending on the genre, the topic, or the origin of the speakers have also been carried out. **Speech technology** also makes ample use of spoken corpora both for the creation and the improvement of speech generation systems (text-to-speech) or speech recognition systems (speech-to-text), which are now increasingly commonly used. Lastly, evidence from spoken (learner) data has been instrumental in the domain of **language learning**. For example, a growing number of learner dictionaries (Collins, LEAD[34], Longman etc.) now also offer authentic spoken examples, and/or contrast spoken and written usage. Corpus-based textbooks used in language education are also slowly gaining momentum.

[34] https://uclouvain.be/en/research-institutes/ilc/cecl/lead.html (last accessed 16/03/2017).

2.2 SPOKEN LEARNER CORPORA AND (DIS)FLUENCY RESEARCH

The previous section provided a general introduction to spoken corpora. The present section zooms in on spoken *learner* corpora and discusses some of their specificities that are particularly relevant in the framework of this dissertation.

As described previously, spoken learner corpora consist in systematic collections of (near-) natural linguistic material produced by L2 speakers and stored in electronic format. Spoken learner corpora differ from spoken native corpora because the speakers are **second or foreign language learners** (SL and FL). As highlighted by Granger (2008), however, although some learner corpora do contain data from SL learners, the majority focuses on FL learners who learn a language in a country where the target language is not predominant or does not have an official status (like English in Belgium).

Spoken learner corpora also differ from datasets from many earlier SLA studies because they aim to be **representative** of learner language. Representativity is ensured by the systematic selection of the texts – in this case, transcripts of spoken data – to be included, which is based on "**explicit design criteria**" (Granger 2008:1427; see also Gilquin 2015; Nesselhauf 2004) that pertain both to the learner (mother tongue, gender, proficiency level etc.) and to the situation or task (topic, genre, time-pressure, exam setting etc.). These variables considerably affect learner language, and it is thus a crucial requirement that they are controlled for.

Another aspect that characterises (spoken) learner corpora (and ensures representativity) is the **(near-) naturalness** of the material they contain (Granger 2008)(Granger 2008). Unlike other types of data like experimental data, learner corpora generally favour **less constrained types of productions**, that is, "data that reflects as closely as possible 'natural' language use (i.e. language that is situationally and interactionally authentic) while recognising that the limitations facing the collection of such data often obligate researchers to resort to clinically elicited data (for example, by using pedagogic tasks)" (Ellis & Barkhuizen 2005:7). So, in practice, spoken corpora (and perhaps learner corpora even more so than native spoken corpora) span a *continuum* of degrees of naturalness, from the more natural (such as spontaneous conversations) to the more constrained (e.g. picture descriptions or reading-aloud tasks). Nesselhauf (2004:128) advises for data collected with more control to be considered "peripheral learner corpora" and Gilquin (2015:10) further underlines that "[w]hen so much control is exerted that the learner is no longer free to choose his/her own wording, for instance in the case of a reading-aloud task, the term 'learner corpus' will normally be

avoided" (see also Granger 2012)[35]. Note also that, when the data has been gathered from both natural and less natural contexts, the term "database" may be used (Gilquin 2015:10). This is the case, for example, for LINDSEI, which is made up of, in decreasing order of naturalness, spontaneous dialogues, monologues on a set topic and picture descriptions[36].

Spoken learner corpora, like written learner corpora (and corpora in general), differ in their **degree of accessibility**. The *Learner Corpora around the World* list[37] (henceforth LCW – see also Figure 2-1), which is maintained by the Centre for English Corpus Linguistics, is a good starting point for a journey into learner corpora as it offers an up-to-date record of available learner corpora "around the world".

Corpus	Target language	First language	Medium	Text type/ task type	Proficiency level	Size in words	Project director	Availability
The Arabic Learner Corpus (ALC)	Arabic	66 languages	written and spoken	Narrative and discussion	Intermediate and advanced	written: c. 283,000 audio: c. 3h30	Abdullah Alfaifi & Eric Atwell	Available
The Pilot Arabic Learner Corpus	Arabic	English	written	Narrative	Intermediate and advanced	c. 9,000	Ghazi Abuhakema Reem Faraj Anna Feldman Eileen Fitzpatrick Montclair State University, USA	
The Jinan Chinese Learner Corpus (JCLC)	Chinese	50 languages	written	Exams and assignments	Beginners, intermediate and advanced	c. 6 m. Chinese characters c. 9,000 texts	Maolin Wang Shervin Malmasi Minggxuan Huang	
The AKCES/CZESL corpus (Acquisition corpora of Czech/Czech as a second language)	Czech	Various	written and spoken	Student essays and interviews	Various	2 m.	Karel Sebesta Charles University in Prague Technical University in Liberec, Czech Republic	Available

Figure 2-1: Screenshot of the LCW list

In what follows, I discuss some of the major features pertaining to the collection of spoken learner corpora (but see also e.g. Gilquin 2015; Granger 2008; Granger, Gilquin & Meunier 2015a), and refer both to the LCW list[38] and to the data used in previous studies from the field of L2 (dis)fluency if the data have not been made available.

[35] Yet, according to some researchers (e.g. Atwell, Howarth & Souter 2003; Ballier & Martin 2015; Gut 2014), even highly controlled data such as decontextualized sentences or read-aloud text passages do qualify as peripheral types of learner corpora.

[36] It must be pointed out, however, that the term "corpus" seems to be used as a generic term too. Gut (2004; 2012), for example, refers to LeaP as a corpus, although it includes both readings of short stories and free speech.

[37] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (last accessed 10/03/2018).

[38] The list of learner corpora available in Spring 2017.

2.2.1 Spoken vs. written learner corpora

It is striking to see that, among all the learner corpora that are listed in the LCW list, only **a quarter** (37; 24%) **pertain to speech,** while the majority are corpora of written language (105; 67% – see Figure 2-2). Some corpora (15; 9%) also contain both spoken and written material, such as the *Longitudinal Database of Learner English* (LONGDALE[39]). Although the proportion of written learner corpora still by far outweighs that of spoken corpora, their number is **steadily going up,** which seems to indicate a renewed and growing interest in oral communication in learner corpus research.

Like spoken native corpora, many spoken learner corpora are "mute" corpora and consist in **transcriptions only** (which often contain some mark-up or tagging). Yet, some **also make audio-recordings available,** and those are sometimes also time aligned with the transcriptions. The LINDSEI[40] database and the *NICT Japanese Learner English* corpus (NICT JLE[41]) (Izumi, Uchimoto & Isahara 2004; 2012), for example, belong to the former category, and the *Spanish Learner Language Oral Corpora* (SPLLOC[42]) and the corpus PAROLE (*PARallèle Oral en Langue Etrangère*[43]) are examples of time aligned corpora.

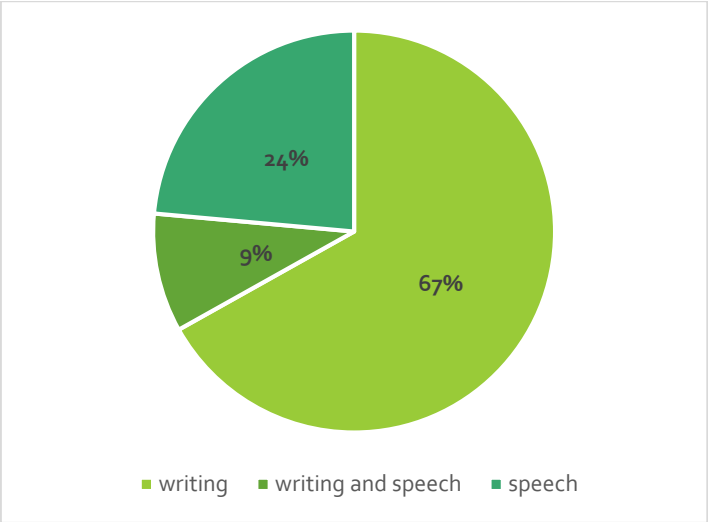


Figure 2-2: The proportion of spoken learner corpora in the LCW list

In the frame of a research into (dis)fluency, access to the audio data is absolutely essential because, as will be made clear in Section 2.3, sole reliance on the transcriptions might lead to

[39] https://uclouvain.be/en/research-institutes/ilc/cecl/longdale.html (last accessed 10/03/2018).

[40] In the CD-ROM edition, only transcripts are made available. However, the national teams that collected the components do have access to the original audio recordings.

[41] https://alaginrc.nict.go.jp/nict_jle/index_E.html (last accessed 23/02/2017).

[42] http://www.splloc.soton.ac.uk/index.html (last accessed 23/02/2017).

[43] https://talkbank.org/access/SLABank/French/PAROLE.html (last accessed 23/02/2017).

dangerous pitfalls. This holds true especially for unfilled pauses, which are often claimed to be at the core of (dis)fluency, and which are used to calculate other temporal measures. Failure to resort to empirical measures of the frequency and length of unfilled pauses (i.e. using audio data, or, even better, time aligned data), might lead researchers to draw erroneous conclusions.

2.2.2 L2s and L1s

Two fundamental properties of learner corpora and their use in (dis)fluency research are the **target language** (L2) and the **mother tongue** (L1) of the speakers. While most spoken learner corpora[44] are **monolingual** (i.e. they contain language from one target language only), some contain linguistic productions in two or more target languages: four spoken corpora in the LCW list (8% – see Figure 2-3) are **multilingual**, such as the COREIL corpus (Delais-Roussarie & Yoo 2010a; 2010b) or the PAROLE corpus. Among the range of L2s, **English** figures predominantly (56% of the monolingual corpora contain learner English[45]), but spoken corpora embrace a larger variety of L2s, including French, Czech, Arabic, Russian and many other languages, as illustrated in Figure 2-4.

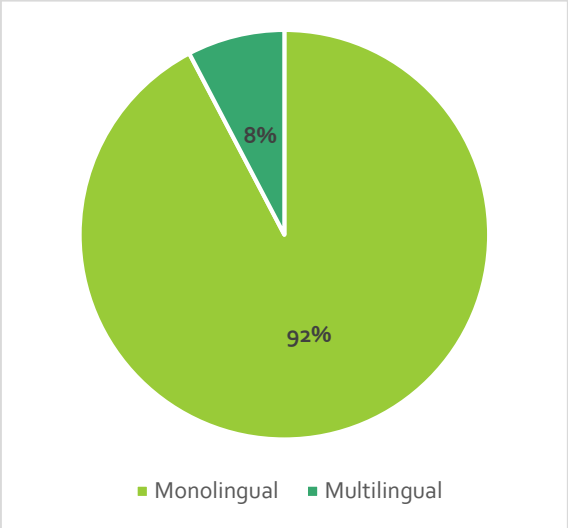


Figure 2-3: Proportion of mono- and multilingual spoken learner corpora (based on the LCW list)

[44] I use here the term "spoken learner corpora" to refer to both learner corpora that contain speech only, and speech and writing (i.e. the 24% and the 9% in Figure 2-2, respectively).

[45] Note also that all multilingual corpora in the LCW list include an L2 English component.

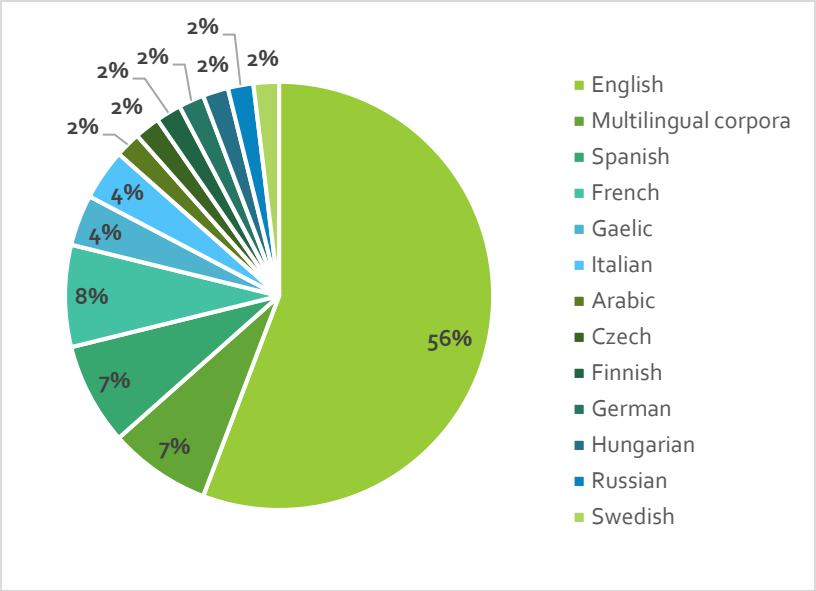


Figure 2-4: L2s in spoken learner corpora (based on the LCW list)

With respect to the **mother tongue** of the learners, again, while some learner corpora are restricted to learners from the same mother tongue background (e.g., NICT JLE corpus contains data from Japanese-speaking learners only), others include more than one L1. A typical example of a multi-L1 spoken learner corpus is LINDSEI: in the CD-ROM edition, this large database contains data from learners of 11 different mother tongue backgrounds, namely Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish and Swedish - nine more components will be included in the second version of LINDSEI, namely the Arabic, Basque, Brazilian Portuguese, Czech, Finnish, Lithuanian, Norwegian, Taiwanese, and Turkish components.

From the point of view of language, it is thus possible to find three types of spoken learner corpora:

- Corpora that include productions of learners from the same mother tongue background in one target language (i.e. **1 L1 and 1 L2**), such as the *NICT JLE* corpus, the *EVA Corpus of Norwegian School English*[46] (Norwegian learners of English) or the *InterFra* corpus[47] (Swedish learners of French);

- Corpora that include productions of learners from different mother tongue backgrounds in one target language (**2 or more L1s and 1 L2**), such as the *Corpus*

[46] http://clu.uni.no/icame/ij21/eva-corp.pdf (last accessed 23/02/2017).

[47] http://www.su.se/romklass/interfra (last accessed 23/02/2017).

Parlato di Italiano L2[48], the *DIAZ* corpus[49] or the LINDSEI database. Such corpora allow, for example, contrastive interlanguage analyses (Granger 2015) where two interlanguages are compared against each other in entirely comparable text types and settings, and can be used to produce generic pedagogical tools such as learners' dictionaries;

- Corpora that include productions of learners from the different mother tongue backgrounds in different target languages (**2 or more L1s and 2 or more L2s**), such as the *European Science Foundation Second Language Database*[50] (ESF), the *PAROLE* corpus or the *University of Toronto Romance Phonetics Database*[51] (RPD).

In the field of **L2 (dis)fluency research**, studies tend to focus more on data from learners of the same language background speaking in one target language (i.e. 1 L1 and 1 L2). A closer look at the L1s and L2s reveals a nice **kaleidoscope of language combinations**: Derwing *et al*. (2004) for example analysed the English of Mandarin Chinese learners; Freed *et al*. (2004) as well as Préfontaine and Kormos (2015; 2016) worked on English-speaking learners of French; Guz (2015) studied the fluency of 45 Polish learners of English and Kahng (2014) that of 31 Korean learners of English. A number of researchers have however adopted a wider approach and looked at the L2 of learners from several L1s together (though without necessarily contrasting subgroups), including Cucchiarini *et al*. (2000; 2002; 2010) (L2 Dutch), De Jong and Bosker (2013) (L2 Dutch of Turkish and English-speaking learners), De Jong *et al* (2012a) (43 different L1s) or Mehnert (1998) (31 learners of German from 9 different L1s).

2.2.3 Proficiency level

Learner corpora and L2 (dis)fluency studies display a marked difference with respect to the **level of proficiency** of the learners, as well as **the way it was assessed**. These factors, however, are of prime importance for the interpretation and comparison of research findings.

There appear to be two main ways of specifying the proficiency level of the learners in spoken corpora: **CEFR levels** and (more frequently) the **beginner/intermediate/advanced** triad. The labels "beginner", "intermediate" or "advanced" are often attributed based on external criteria computed in terms of the number of years the learner has been studying the L2, the

[48] http://elearning.unistrapg.it/osservatorio/Corpora.html (last accessed 23/02/2017).

[49] http://www.language-archives.org/item/oai:talkbank.org:SLABank-Spanish-DiazRodriguez (last accessed 23/02/2017).

[50] http://www.mpi.nl/tg/lapp/esf/esf.html (last accessed 23/02/2017).

[51] http://rpd.chass.utoronto.ca/ (last accessed 23/02/2017).

year at university[52], the amount of time spent abroad etc. Research, however, indicates that such global measures can be operationalised in different ways by different researchers (see Ortega & Byrnes 2008 on four ways of operationalising advancedness) and are not always reliable (see e.g. Callies & Götz 2015a; Gass & Selinker 2008; Granger & Thewissen 2005). In this respect, CEFR levels are arguably more robust because each level has a corresponding descriptor but, in fact, few are the corpora where the stated proficiency level has actually been assessed based on those descriptors.

Although it appears from the LCW bibliography that most spoken learner corpora cover two or more proficiency "levels", researchers into learner (dis)fluency may well choose to focus on **one proficiency level** exclusively: Derwing *et al*. (2004; 2009), for example, concentrates on beginners; Rossiter (2009) on learners of an intermediate level; Tavakoli (2011; 2016) specifies that her learners have a B2 level; and Riazantseva (2001) analysed the fluency of very advanced learners. Alternatively, a broader perspective may be adopted, where either **two main proficiency bands** are analysed contrastively (e.g. Bosker *et al*. (2013); De Jong (2016) and Ginther *et al*. (2010), who analysed intermediate to advanced learners), or the **whole proficiency continuum** is explored (e.g. Cucchiarini *et al*. (2000; 2002); Préfontaine and Kormos (2016) analysed learners at a beginner, intermediate and advanced levels).

With regard to the **method used to assign proficiency levels** in spoken corpora, a useful distinction is generally made between learner-centered methods and text-centered methods (Carlsen 2009; 2012), which broadly correspond to Atkins et al.'s (1992) distinction between external and internal criteria, respectively. While in the former, characteristics of learners (i.e. external to the linguistic production) such as age or institutional status are used to assign a proficiency level to the corpus or transcriptions, internal criteria are essentially linguistic. Most corpora in the LCW list have used external criteria such as institutional status to assign proficiency level. As an organisational convenience, and because of the importance of (the evaluation of) proficiency/fluency level in the present dissertation, more details on the practical assessment of learner proficiency/fluency level are provided in Section 2.4.

2.2.4 Age and number of learners

The **age and number of learners** are also prone to a great level of fluctuation in learner corpora and datasets used in L2 (dis)fluency research, ranging from children to adults, and from very small to larger numbers of speakers (some L2 (dis)fluency studies are recorded in Table 2-1). Although it is very difficult – if not impossible – to say how many learners a corpus

[52] Two spoken corpora in the LCW list use purely external criteria to indicate the learners' proficiency level. The learners in the *English Speech Corpus of Chinese Learners* (ESCCL) have, for example, a "middle school and college" proficiency level.

should ideally contain, the limited number of learners included in some analyses does raise **concerns in terms of representativity**.

Reference	No. of learners	Age of the learners	Proficiency of the learners
Cucchiarini *et al*. (2002)	60 + 57	Adults	Beginner to advanced
De Jong *et al*. (2012b)	189	Undergraduates	Intermediate to advanced
Derwing *et al*. (2009)	32	Adult immigrants	Beginner
Fathman (1980)	75	Children	Beginner
Ginther *et al*. (2010)	125	Teaching assistants	High intermediate and advanced
Götz (2013a)	50	University students	Advanced
Lennon (1990; 1995)	4	University students	Advanced
Tavakoli (2016)	35	EAP students	B2 level
Towell (2002)	12	University students	Intermediate
Towell (1987)	2	Undergraduates	Advanced
Trofimovich and Baker (2007)	20	Children and adults	Intermediate and native-like
Trofimovich and Baker (2006)	30	Adults	Varied

Table 2-1: Age, number of learners and proficiency level in L2 studies

2.2.5 Time of collection

Time is an important characteristic in a corpus. In the same way as written corpora, spoken learner corpora may aim to capture language in a **synchronic** or in a **longitudinal** (i.e. diachronic) perspective: the former are collected at a single point in time and provide a "snapshot of learners' knowledge of the target language at a particular moment", and the latter provide "a representation of the evolution of their knowledge through time" (Gilquin 2015:14) because they are collected at successive points.

To date, most spoken corpora have focused on learners in a **synchronic perspective** (90% in the LCW, as illustrated in Figure 2-5). The added difficulty in collecting longitudinal spoken data (e.g. due to longer collection time or to drop outs) is reflected in the very **low number of longitudinal spoken learner corpora** available: in the LCW list, there are but five spoken learner corpora that contain longitudinal data, among which the *Spoken and Written English Corpus of Chinese Learners* (SWECCL) and the *Corpus of Young Learner Interlanguage* (CYLIL). One other project is the *Longitudinal Database of Learner English* (LONGDALE), which aims

at collecting written and spoken data from the same learners at different time points over a period of (at least) three years. Despite their rarity, these corpora are very precious resources to track down changes in second/foreign language acquisition.

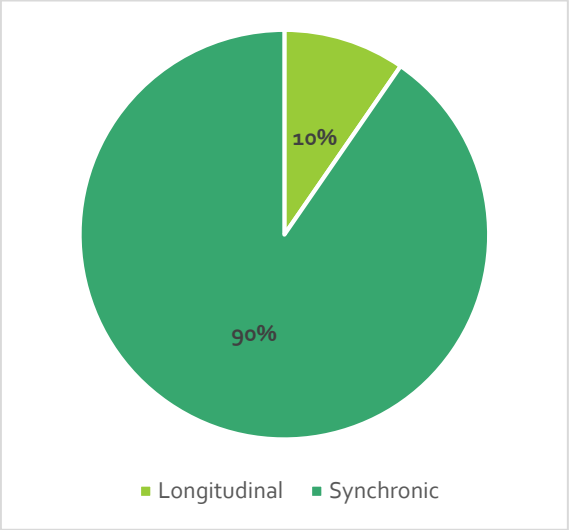


Figure 2-5: Proportion of synchronic and longitudinal spoken learner corpora in the LCW

Likewise, most **L2 (dis)fluency studies** so far have analysed the fluency characteristics of learners synchronically. Investigations into the development of fluency over time are scarce (Table 2-2 below offers a selection of such studies) and vary in terms of time-spans – from 2 months to c. 4 years.

Reference	Time span	Learner characteristics
Derwing *et al*. (2009)	3 time-points over 2 years	16 Slavic and 16 Mandarin-speaking learners of English
Freed *et al*. (2004)	2 time-points over 1 semester	28 English-speaking learners of French in 3 learning contexts
Lennon (1990)	2 time-points over 6 months	4 German-speaking learners of English
Lennon (1995)	2 time-points over 2 months	4 German-speaking learners of English
Tonkyn (2012)	2 time-points over 10 weeks	24 postgraduate learners of English (varied L1s)
Towell (1987)	3 time-points over 4 years	2 English-speaking learners of French
Towell (2002)	4 time-points over 4 years	12 English-speaking learners of French

Table 2-2: Some longitudinal L2 fluency studies

2.2.6 Speaking tasks

The **centrality of speaking tasks** in spoken learner corpus research[53] is now a widely-accepted position (see e.g. Swales 2009). Speaking tasks such as read speech, picture description, or an interview impose different demands on the speaker (Butterworth 1980), and research has shown that variations in a speaker's output can, indeed, be attributed to the speaking task and its properties. Task variables include, *inter alia*:

- **the degree of naturalness**: to what extent is the speech constrained (e.g. read speech vs. free discussion)?
- **the preparation (or "planning") time**: did the learner prepare what he/she was going to say? How long did he/she have to plan his/her speech?
- **the degree of interaction**: did the learner engage in an interactive task (viz. a dialogue or a multilogue) or not (i.e. a monologue)?
- **the degree of complexity**: to what extent is the task cognitively complex?
- **the topic familiarity**: did the learner talk about a familiar topic (e.g. holidays vs. nuclear energy)?

In the LCW list, most spoken learner data collections sample **more than one speaking task** at a time (i.e. they qualify as "databases"). To give but three examples, the ANGLISH[54] database (Tortel 2008) contains both readings of texts as well as spontaneous dialogues; LeaP also contains four types of speech styles (including the retelling of the story and free speech in an interview situation), and the *Louvain International Database of Spoken English Interlanguage* (LINDSEI) consists of three speaking tasks (a set topic, a free discussion and a picture description). This variety of speaking tasks within the same database is, I believe, a great advantage in learner corpus research. It first increases the **reusability** of the data for different research purposes. Also, it opens the way to investigations into **how some speech property** (e.g. speech rate) **may vary (or not) across different communicative situations** for a same speaker[55] (e.g. Dumont 2017b; Skehan & Foster 2012; Tavakoli 2016). Besides, if comparable native speaker data are available for the same speaking tasks, it becomes possible to investigate whether variations in performance are due to a learner processing an L2 or whether these variations are task-induced (if learners and native speakers are affected similarly across tasks, then L2 processing is not the cause of these variations). This might for

[53] Although I focus here on learner corpus research, many of the task variables that affect the learners' interlanguage also affect native speakers' speech (e.g. Foster & Tavakoli 2009; see also Tavakoli 2009).

[54] The corpus is available at: http://sldr.org/voir_depot.php?id=731&lang=fr&sip=1 (last accessed 19/03/2017).

[55] Of course, data from different speaking tasks may also be pooled to increase the amount of data per speaker.

example contribute to the delineation of some sort of hierarchy of speaking tasks depending on their level of difficulty, with implications for language teaching and assessment.

The buzzing activity in the field of **learner speech and L2 (dis)fluency** is revealed by the many different speaking tasks that have been used over the years, as can be observed in Table 2-3. Although it appears from this table that task denomination is still in need of some homogenisation, the variety of the tasks positively highlights the richness of analyses of learner speech.

Speaking task	Selected references
Monologue	De Jong *et al*. (2012b; 2012a); Derwing *et al*. (2004); Tavakoli (2011; 2016)
Cartoon description	Derwing *et al*. (2004); Deschamps (1980); Grosjean (1980a); Kormos and Dénes (2004); Préfontaine (2013a); Rossiter (2009); Tavakoli (2011)
Argumentative task	Bosker *et al*. (2013)
Narrative	Ortega (1999); Préfontaine and Kormos (2015; 2016); Tavakoli (2009; 2011)
Decision-making	Levkina and Gilabert (2012)
Giving an opinion	Ginther *et al*. (2010); Iwashita *et al*. (2008)
Spontaneous speech	Cucchiarini *et al*. (2002; 2010); Goldman *et al*. (2010); O'Brien *et al*. (2007); Riggenbach (1991); Segalowitz and Freed (2004)
Interview	Baker-Smemoe *et al*. (2014); Fathman (1980); Freed (1995); Freed *et al*. (2004); Fuller (2003); Götz (2013a); Temple (2000)
Telephone conversation	Kapatsinski (2010)
Multilogue	Butterworth (1980)

Table 2-3: Speaking tasks in L2 (dis)fluency research

Owing to the **growing awareness that task influences a speaker's output**, a body of research has turned to investigate and contrast speaking performances across task variables (see e.g. Butterworth 1980; De Jong *et al.* 2012b; Foster & Skehan 1996; Goldman-Eisler 1961a; Levkina & Gilabert 2012). In what follows, I briefly summarise the findings on two such task variables that have attracted a lot of attention, namely the number of speakers actively engaged in the discourse and the extent of planning of speech.

2.2.6.1 Monologues vs. dialogues

Monologues involve the production of sequences by one speaker, whereas dialogues[56] are "prototypically a joint enterprise involving more than one person" (Cameron 2001:87) and where the speakers take turns to talk in a collaborative manner (Wilson & Zimmerman 1986). But these two types of speech do not only differ because of the **number of speakers** involved, but also with respect to a number of **speech characteristics**. Multi-partite discourse, unlike monologues, is for example characterised by (unclaimed) inter-turn silent pauses, interruptions by another speaker, overlapping speech and turn-taking (see e.g. Edwards 2001; Tavakoli 2016).

Although linguists typically associate oral language with conversation (i.e. spontaneous dialogue), in the domain of learner speech and (dis)fluency research, the **principal source** of empirical material for research used to come – and, though to a lesser extent, still comes (*cf*. Table 2-3 above) – nearly exclusively from **monologic tasks**. Research results were then more or less implicitly generalised to interactive tasks and dialogues (Horowitz & Samuels 2005; O'Connell & Kowal 2008; Tavakoli 2016).

Tavakoli (2016:136) claims that the frequent use of monologic tasks in SLA and learner corpus research can be attributed to 3 main factors, namely the **degree of control**, the **predictability of the outcome** and the **clarity and ease of procedures and measurements**. Dialogues, contrarily to monologues, involve complex pragmatics, which leads to less controlled and less predictable performances. Moreover, the interactive nature of dialogues renders difficult the handling of, e.g., simultaneous speech, inter-turn silent pauses, or cross-turn phenomena (e.g. a repeat after an interruption by the interlocutor). Another, but related, issue concerns the fact that the speech canal may either be occupied by one speaker, more than one speaker, or no-one: this greatly complicates the measurement of temporal variables (Goldman, Auchlin & Simon 2013). Because all these issues are inexistent (or can easily be circumvented) in monologues, it is in fact not surprising that research has started the investigations of speech with a – in a sense – less complex material.

The danger for (dis)fluency research lies in the temptation to transpose the monologic conceptualisation of the construct onto dialogic speech. Studies that examined the differences between monologic and dialogic (dis)fluency have found that **dialogues are characterised by a higher fluency** – a higher speech rate, less pausing time and fewer repairs (Bell 2003; Michel 2011; Witton-Davies 2014). In a between-participant study, Tavakoli (2016) used 1-minute monologues (retellings of an experience) and 3-minute dialogues (argumentative discussions) produced by 35 English for Academic Purposes students (B2 level). While her findings on speech rate, pausing time and rate of repair phenomena are

[56] I use here the word "dialogue" as a cover-term for interactions between two or more than two people (i.e. "multilogues").

consistent with previous results (i.e. dialogues being characterised by a higher fluency), she found little difference in terms of number and location of pauses between monologues and dialogues, which seems to indicate that the mode has little influence on how often and where L2 speakers pause. On a methodological level, her analysis also explored whether – and demonstrated that – **the choice and operationalisation of fluency measures in dialogues has an impact on the results** (a case in point is the handling of inter-turn unfilled pauses).

Two reasons have been put forward to explain the fact that dialogues are more fluent than monologues (e.g. Tavakoli 2016; Webber 2001). First, it has been argued that speakers engaged in dialogues can **use the time** when the other speaker speaks to plan their next utterance, which favours the fluent and smooth delivery of the turns. Second, it might also be that **having an interlocutor** genuinely encourages speakers to engage more actively in the discourse, and to take the interlocutor's needs into account.

2.2.6.2 Planning and planning time

Another task-related variable that has been proved to influence L2 speech production is related to planning. Planning is one of the two phases of language production, together with the execution phase (e.g. Clark & Clark 1977; Levelt 1989), and it can be operationalised in two ways: pre-planning – **planning before a task** – and online planning – **planning during a speaking task**. To date, most studies have analysed the effect of pre-planning (e.g. Foster & Skehan 1996; Mehrang & Rahimpour 2010; Ortega 1999; Wigglesworth & Elder 2010), and limited attention has been devoted to online planning (e.g. Ellis 2009; Nakakubo 2011; Yuan & Ellis 2003).

Previous research has indicated that the activation of linguistic procedures (such as lexical retrieval, or the retrieval of phonological forms) requires a high level of cognitive control for learners, especially at lower levels of proficiency. It has thus been posited that pre-planning may be effective in reducing cognitive load, thereby allowing learners' attentional resources to attend to other aspects of language such as linguistic complexity, grammatical accuracy or fluency (e.g. Crookes 1989; Foster & Skehan 1996). Several studies have consistently reported that, when learners are given some time to plan their speech in advance, they produce significantly **more fluent and more complex language** (e.g. Ahmadian & Tavakoli 2011; Foster & Skehan 2009; Levkina & Gilabert 2012; Ortega 1999). Working within the context of task-based instruction, Foster and Skehan (1996) also provided evidence of an interaction between pre-planning and task type: in their study, the effects of planning on fluency (and complexity) were greater for more cognitively demanding tasks.

Although the studies mentioned above are fairly consistent in their results, they employed different planning durations, from 1 to c. 10 minutes (and different speaking tasks), so it is difficult to compare the findings more precisely. A stream of research has however looked at whether the **duration of pre-planning** also has an effect on language production (e.g. Foster & Skehan 1996; Levkina & Gilabert 2012). Mehnert (1998) for example reports on a study that

investigated four groups of learners of German. The control group had no planning time available and the other 3 groups, 1, 5, or 10 minutes of planning time, respectively. The results of this study show that **fluency does increase with greater planning time**. Mehnert also investigated how other aspects of speech performance relate to planning time. Lexical density, like fluency, also proved to increase with longer planning time. Accuracy improved with only 1 minute of planning but did not increase further with longer planning times. Lastly, complexity was higher with the longest planning time condition.

In conclusion, it seems that it is not only pre-planning that has a positive effect on learners' fluency[57], but also the extent of this planning.

2.2.7 Select overview of spoken learner corpora

The previous sub-sections set out to introduce spoken learner corpora and the factors that are of major importance for their collection, with particular emphasis on those that have been shown to affect L2 (dis)fluency. As we have seen, spoken learner corpora capture a variety of L2s and L1s. The age, proficiency level and the number of learners, too, may be very different depending on the corpus. Lastly, speaking tasks have been shown to be important cornerstones in spoken corpus design because their impact on research findings (in this case, on (dis)fluency measures) is immediate and clearly significant.

Although this diversity in terms of spoken learner corpus properties is definitely proof of the buzzing life in the field, it is also a likely reason for the **mixed results** in the (dis)fluency literature. It is also a reminder of the **caution** one must exercise in referring to earlier work because it may at times become tempting, yet tricky, to straightforwardly compare results. Most (dis)fluency researchers are aware of this issue, and take utmost care to only refer to those studies that are best comparable with their own, and I will try to do so too.

As a conclusion of this section devoted to spoken learner corpora, an overview of some corpora is provided in Table 2-4, with indications of – among others – the mother tongue and target language of the learners, their level of proficiency, and the type of speaking task they performed.

[57] Note also that Foster and Skehan (2009:211) showed that "planning affects both NS and NNS in similar ways although effects on NNSs are slightly weaker."

Corpus / database	Type	L2	L1	No. of participants	Proficiency level(s)	Speaking task(s)	Duration	Transcription and annotation
LINDSEI Gilquin *et al*. **(2010)**	Mute	English	Varied	50 per component	Intermediate to advanced	Informal interviews made up of: (1) a set topic; (2) a free discussion and (3) a picture description	c. 15 min. per interview (i.e. c. 12 hours per component)	Orthographic
EVA	Written and spoken data (mute)	English	Norwegian	62	Intermediate	3 picture-based tasks	Unknown	Orthographic
InterFra	Mute (audio files are available) Longitudinal + written data	French	Swedish	8 + 18	Intermediate to advanced	Interviews, retellings of video clips and picture story	Unknown	Orthographic
DIAZ	Time aligned Longitudinal	Spanish	German Swedish Icelandic Korean Chinese	8	Unknown	Semi-spontaneous structured interviews (and experimental data from questionnaires)	Unknown	Orthographic
NICT JLE Izumi *et al*. **(2004; 2012)**	Mute	English	Japanese	c. 1300	9 proficiency levels	Interview tests	c. 15 min. per interview (c. 300 hours in total) c. 2 million words	Orthographic Error tagging and speech features
SPLLOC Mitchell *et al*. **(2008)**	Time aligned	Spanish	English	120	Beginners, intermediate and advanced	Narratives, picture-based interviews, pair discussions (and clitic production)	c. 40 hours c. 270,000 words	Orthographic

PAROLE Hilton *et al*. **(2008)**	Time aligned	Italian French English	Varied	68	Beginners, intermediate and advanced	Summary, commentaries	c. 20 min. per speaker c. 20,000 words	Orthographic
ANGLISH Tortel **(2008)**	Time aligned	English	French	40	Intermediate and advanced	Read speech, repetition task, unprepared monologues	c. 6 hours (including NS component)	Orthographic
LeaP Gut **(2004; 2009; 2012)**	Time aligned	German English	Various	101	Intermediate to advanced	Read speech, prepared speech, free speech, story retelling, nonsense word lists	c. 12 hours	Phonetic and phonological transcriptions
LONGDALE Meunier *et al*. **(2010)**	Written and spoken (mute) data; longitudinal	English	Various	117 in the French component	Intermediate to advanced	Informal interviews	Unknown	Orthographic

Table 2-4: Some representative spoken learner corpora and databases

2.3 THE REPRESENTATION AND EXPLOITATION OF SPOKEN CORPUS DATA

Decisions relating to the properties of corpus design (such as those outlined in the previous section) are of fundamental importance because they guarantee that the data can be used to meet the specific research objectives of the researcher. But before spoken corpora can be subjected to analysis, the oral data have to be captured adequately and in a way that allows researchers to investigate them. The final section of this chapter addresses different considerations behind the transcription and representation of (learner and native) spoken discourse.

Besides the practical aspects of the recording of speakers (which, due to space constraints, I will not discuss here, but see, e.g. Podesva & Zsiga (2013)), there are a number of important considerations that need to be taken into account for the construction and exploitation of spoken corpora, including the **transcription**, the **time alignment** of the transcriptions and the audio recordings, and the **linguistic annotation** of the data. These steps are not entirely independent: on the contrary, they interact and influence each other (see e.g. Adolphs & Carter 2013). For example, the research aims may determine the degree of detail (and the spatial arrangement) of the transcription as well as the scale of time alignment, and the scale of time alignment in turn determines the granularity of linguistic annotations (or vice versa). In this section, I will describe the three aforementioned steps in the representation of spoken data, with particular attention to their exploitation in the framework of (dis)fluency research.

2.3.1 Transcribing spoken language

Spoken discourse is ephemeral in essence, and as soon as it has been uttered, it "flies away" into the abyss of oblivion (*verba volant*!). Linguistic analyses of spoken language based only on an audio signal are thus an impossible enterprise, as contended by Blanche-Benveniste (2000:24; my translation)[58]:

> One cannot study speech through speech, relying on the memory one has of it. One cannot, without the help of visual representation, walk through speech and compare its pieces.

[58] Original quote: "On ne peut pas étudier l'oral par l'oral, en se fiant à la mémoire qu'on en garde. On ne peut pas, sans le secours de la représentation visuelle, parcourir l'oral en tous sens et en comparer les morceaux".

Fortunately, investigations of oral data can be carried out through written texts that function as a proxy for the primary data (*scripta manent!*). Somewhat paradoxically then, **spoken language is (primarily) analysed based on written transcripts**[59].

The act of transcribing is by no means easy or neutral and each decision has profound consequences not only on the utility of the transcripts, but also on the interpretation of the findings, as rightly underlined by Kendall (2008:337):

> [...] the act of transcription, especially by beginning transcribers, is often undertaken as a purely methodological activity, as if it were theory neutral. Each decision that is made while transcribing influences and constrains the resulting possible readings and analyses [...]. Decisions as seemingly straightforward as how to lay out the text, to those more nuanced – like how much non-verbal information to include and how to encode minutiae such as pause length and utterance overlap – have far-reaching effects on the utility of a transcript and the directions in which the transcript may lead analysts.

The representation of spoken data is, indeed, one of the biggest challenges in spoken corpus linguistics, suffice it to say that it has been metaphorically equated with a "black hole" (McCarthy 1998:13). Bearing this in mind, Edwards (1992; but see also Edwards 2001), in her chapter on discourse transcription, outlines two general design goals of written transcriptions of oral data:

- **authenticity**, or the fact that "transcripts preserve the information needed by the researcher in a manner which is true to the nature of the interaction itself" (Edwards 1992:4);
- **practicality**, or the fact that "its conventions be practical with respect to the way in which the data are to be managed and analysed, for example, easy to read, apply to new data sets, and expand if needed for other purposes" (*ibid.*:4).

She further mentions different principles subserving these goals, including readability and computational tractability. The latter has to do with the **systematicity and predictability in transcribing**. Failure to meet this requirement may either result in "underselection" (i.e. overlooking relevant instances) or in "overselection" (i.e. retrieval of non-relevant instances along with the relevant ones). The principle of readability aims to ensure that information in transcripts is preserved in a form which enables the researcher to extract the target information quickly. One main aspect of this principle is **time-space iconicity**, according to which temporally prior events are encountered earlier on the page then temporally later events. Another major aspect of readability pertains to the issue of the **spatial arrangement of the speaker(s)' turns**. Three main arrangements are possible, aka vertical (or "linear"

[59] For an overview of the historical attempts to transcribe discourse, see Edwards (2001:338–343). The author interestingly notes that the earliest attempt to capture spoken language in writing dates back to ancient Greece and the golden age of the art of oratory. The "entextualisation" of speech is thus by no means recent; only the means to do so have changed.

(Adolphs & Carter 2013:13)), column and partiture (or "musical score" (*ibid.*:13)). A brief description of each is included in Table 2-5 below.

Spatial arrangement	Definition
Vertical	Speakers' turns are displayed one above the other chronologically. Time is preserved in the vertical dimension.
Column	Speakers' turns are arranged in columns (one column per speaker). Time is preserved in the vertical dimension.
Partiture	Events are displayed horizontally; the talk of each speaker is arranged on a different line on the score. Events on the same vertical axis represent simultaneous acoustic events produced by different speakers. Time is preserved in the horizontal dimension.

Table 2-5: Three types of spatial arrangements in written transcriptions

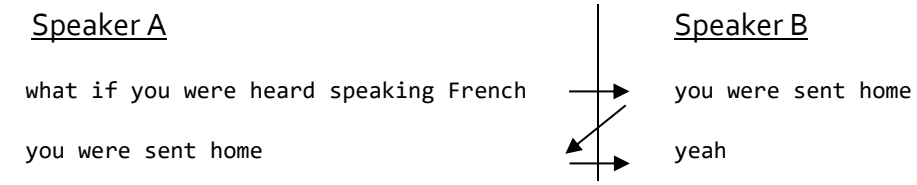
An example of each display is shown in Example 2-1 through 2-3 (the arrows indicate the reading direction).

The **choice of a spatial display** is not as neutral as it might seem: it also **affects the perception of the data**. Vertical arrangement, the most widely used spatial arrangement, tends to bias the reader to perceive speakers as equally engaged in the interaction (it also makes it difficult to show overlaps). By contrast, a display in columns biases the reader to perceive asymmetries between the speakers (the left-most speaker being the most dominant). It is also less readable when many speakers are involved in an interaction. Finally, the partiture display, in addition to giving an impression of equal communicative status between the speakers (like the vertical arrangement), is very efficient in capturing interaction: it emphasises turn taking, and clearly shows (even many) simultaneous utterances as well as the timing and sequencing of turns.

2-1: Vertical arrangement (LINDSEI - FR006-F)

```
A     what if you were heard speaking French
B     you were sent home
A     you were sent home
B     yeah
```

2-2: Column arrangement (LINDSEI - FR006-F)



2-3: Partiture display (LINDSEI - FR006-F)

A	what if you were heard speaking French		you were sent home	
B		you were sent home		yeah

Irrespective of the spatial choice adopted, careful consideration ought to be given to several other aspects related to transcribing and transcriptions. The most important ones are discussed below.

2.3.1.1 How much to transcribe

Transcripts can roughly be defined as "line-by-line account[s] of what was actually said" (Jenks 2011:2). But, in addition to providing a detailed account of the **words that have been uttered**, they may also aim at capturing ***how* those words or utterances have been uttered** by the original speakers (for example by encoding some paralinguistic or prosodic information).

To date, there is no universal way of transcribing spoken data: depending on their theoretical or methodological underpinnings, researchers may choose among **transcription systems of varying degrees of detail** to capture the verbal, prosodic, and paraverbal aspects of spoken language – and each transcription system carries important implications for the types of research that are made possible. Jenks (2011), in his book on the transcription of talk and interaction, distinguishes **five types of transcription detail** (Table 2-6), ranging from (the narrative and) the orthographic[60] transcription of the words that have been produced, to transcriptions that also include interactional and/or paralinguistic and/or multimodal features. For example, while an orthographic transcription might be appropriate for an analysis of morphological patterns or of the lexical bundles used by learners of English, an analysis of (dis)fluency rather requires (in addition to the transcription of the words) a detailed interactional and pragmatic account of pauses, reformulations and possibly also other prosodic features.

The types presented in Table 2-6, Jenks advises, should however "only be used as a starting point to determine what types of detail can be transcribed": the **level of granularity to be**

[60] Jenk's (2011) orthographic transcription should not be confused with one of the two main *ways* of transcribing (i.e. orthographic vs. phonetic).

adopted is always dependent on the researcher and his/her research aims (see also Adolphs & Carter 2013:11–12) and, also, on practical constraints. Time and funding allowing, it is worth considering producing richer transcriptions (i.e. with interactional, paralinguistic and/or multimodal mark-up) because, in addition to better reflecting the original audio data, interactional or paralinguistic features might help in interpreting the results more accurately, and such transcriptions might be more easily re-usable in other research projects.

1	narrative	i.e. a narrative account of the communicative event
2	orthographic	i.e. words only
3	interactional	e.g. pauses and overlapping speech
4	paralinguistic	e.g. elongation, voice amplitude, stress, intonation
5	multimodal	e.g. written notes and video stills of gestures

Table 2-6: Five types of transcription details (adapted from Jenks 2011:43)

When the researcher has selected a spatial display as well as the appropriate level of granularity to be adopted in the transcriptions, he/she has to decide how to actually transcribe it. I will first set out different ways in which linguistic information can be encoded in written transcriptions, and then give some more consideration to the transcription of interactional and paralinguistic features.

2.3.1.2 Transcribing linguistic information

A. Vernacularisation and standardisation

There exist two main strategies to address the challenge of representing the colourful ways of oral communication, namely vernacularisation and standardisation (Jenks 2011; see also Ballier & Martin 2015; Nagy & Sharma 2013).

Transcribers may **transcribe speech as it is being spoken**. The so-called vernacularisation strategy "seeks to capture unique ways and styles in which words and utterances are spoken" (Jenks 2011:19): the transcription is a written representation of all kinds of "pronunciation particulars" (Jefferson 1983). Two transcribing options are possible: one is to use non-standard spelling (or "folk orthographic representations", also referred to as "eye dialect" (Jenks 2011; Nagy & Sharma 2013)), the other is to use the International Phonetic Alphabet (IPA). In **eye dialect transcriptions**, graphic deformations[61] are used to reflect the actual pronunciation (typically contracted forms or vowel/consonant elisions). For example, the form *did you* could be transcribed *did you, didja, didya, did ya* etc. Although the correspondence between symbols and pronunciation is more homogeneous with the **International Phonetic Alphabet**, depending on the actual pronunciation, several

[61] Blanche-Benveniste et al. (1987) call them "trucages orthographiques" (En. spelling tricks).

transcriptions of the same word are also possible (*that* could be transcribed [dæt], [ϴæt], [ðæt] etc.). The two vernacularisation options accurately depict actual pronunciation particulars of – especially non-standard – speakers (such as learners), but they have also spawned legitimate criticism. Firstly, vernacularised transcriptions are **not easy to decipher** and the need might sometimes arise to "oralise"[62] them to understand what has been said (Dister & Simon 2008). Secondly, as stressed by Edwards (2001:324), "(f)or purposes of **computer manipulation** (e.g. search, data exchange, or flexible formatting), the single most important design principle is that *similar instances be encoded in predictably similar ways*" (italics original). Without a consistent way of transcribing, computerised analysis easily becomes inaccurate and misleading (Andersen 2016:324–325):

> For one thing, it causes problems for end users of corpora, who may have to search and analyse more than one variant of the same word for full accountability, without necessarily knowing the full set of variable representations of the same feature. Moreover, it leads to inaccuracy in statistic calculation and in the annotation made by computational grammars that use lexicons as bases for tagging and parsing techniques.

Lastly, vernacularised transcriptions have been claimed to be **socially problematic** (e.g. Dister & Simon 2008; Jenks 2011; Nagy & Sharma 2013): because their focus is on the representation of non-standard forms, such transcriptions may lead to negative social evaluations by reinforcing the stereotypes associated with some social/regional/… groups. In other words, they might provide evidence of such groups using "defective" speech (rather than evidence of pronunciation variants).

Alternatively, verbal discourse can be encoded using standard orthographic spelling – i.e. the standardisation strategy. **Orthographic transcriptions** are the primary mode of representation of speech in a non-oral format (Kendall 2008), and tend to be used for transcribing large amounts of data (Delais-Roussarie & Post 2014:53). Although it does strip away pronunciation idiosyncrasies and might be more difficult to apply to less standardised language varieties (such as the Picard (see Nagy & Sharma 2013), but potentially also (learner and) non-native varieties), orthographic transcription has the double advantage of being less cognitively demanding both for the transcriber and the reader, and of being far more homogeneous and predictable than non-standard spellings[63] (which in turn eases computer manipulations considerably).

---

[62] From French "oraliser", i.e. utter a text out loud.

[63] Note, however, that even highly standardised languages such as native English may include less standardized forms, like filled pauses (*uh/eh/uhm* etc.) or contracted forms (*going to/gonna; do not/donno*) – see e.g. Andersen (2016) for a comparative study of the transcription of filled pauses, interjections, phonological reductions and discourse markers in spoken corpora. Nagy & Sharma (2013) advise researchers to decide on a "standardised" spelling of these forms, which would ideally be included in a transcription protocol that transcribers will refer to during the transcription process. Other considerations include the informed choice of the norm (e.g. British English vs. American English) and the spelling of numbers, abbreviations, and acronyms.

Vernacularisation and standardisation need not be seen as polar opposites: they are the two ends of a continuum, and researchers are free to **use both strategies together** (e.g. use vernacularisation only when relevant in an orthographic transcription) if deemed appropriate for their research purposes.

B. Punctuation

Another issue pertains to the use (or not) of **punctuation** in the transcripts. Blanche-Benveniste *et al*. (1987) advise against the use of punctuation on the grounds that there is no established correspondence between spoken prosody and written punctuation. For example, a dot at the end of a sentence might correspond to nothing in the original audio signal, not even a silent pause. Moreover, the use of punctuation could also misleadingly (and unknowingly) suggest an analysis to the researcher: "punctuation, if integrated too early, prejudges the syntactic analysis and imposes a division on which it is difficult to return" (Blanche-Benveniste, Jeanjean & Monfrin 1987:142; my translation)[64]. Dister and Simon (2008) even go a step further by saying that the exclusion of punctuation in transcriptions is linked to the calling into question of the notion of sentence in speech.

2.3.1.3 Transcribing interactional and paralinguistic features

The transcription of interactional (e.g. pauses, truncations, repetitions, or overlapping speech) and paralinguistic features (laughs, voice quality etc.) in speech requires great attention (and often some training) on the part of the transcriber. Many researchers (e.g. Gilquin 2008; McCarthy 1998; O'Connell & Kowal 1995) have indeed highlighted the fact that transcribers, however skilful and well-intentioned they might be, may at some point unknowingly correct mistakes the speakers produced, delete redundant repetitions, or simply be "**deaf**" to pauses, hesitations, discourse markers and the like. Several checks of the transcriptions, ideally combined with a thorough annotation of such elements, are thus a prime requirement for the analysis of interactional and paralinguistic features.

With respect to the transcription of interactional and paralinguistic features, Jenks (2011:46) concisely encapsulates the standard convention:

> *In most transcription systems, the standard convention for representing talk and interaction is to use symbols and punctuation markers. For each unique interactional or paralinguistic feature, there is generally a symbol or punctuation marker used to represent it in the transcripts.*

[64] Original quote: "[l]a ponctuation, si on la met trop tôt, préjuge de l'analyse syntaxique et impose un découpage sur lequel il est difficile de revenir".

In the LINDSEI database and the VOICE project[65], for example, transcriptions are interwoven with symbols and punctuation markers: truncated words are marked by a "=" sign; vowel lengthenings are indicated with ":" etc. A case in point, however, is the transcription of **unfilled pauses** (see also e.g. Larsson Aas & Nacey (2017)). It is not uncommon that they are transcribed based on subjective appreciations of their length. In LINDSEI and LOCNEC for instance, one, two, or three dots are used to mark silent pauses depending on their perceived length. In such cases, it is the ***perception* of pause length** that can be studied[66]. The timing of pauses has, however, been made far easier these last few years with advances in computer technologies. Transcribers now also have the possibility of including the precise length of pauses in the transcriptions, thereby allowing the study of the ***actual and measurable* length of the pauses**.

2.3.1.4 The relationship between transcripts and audio recordings

Despite the widely declared assertion (see e.g. Jenks 2011; Kendall 2008) that transcriptions should not be considered substitutes for the original oral data but "additional tools which can be used to help analyse and understand these recordings" (Liddicoat 2007:13), to date, the majority of spoken corpora consist of **transcripts of spoken language only** (i.e. mute corpora). The written representation of speech thus regularly ends up as the primary data used for the analysis. (see also Sections 2.1 and 2.2 *supra*).

The fact that many spoken corpora consist in (and that many analyses of spoken data are based on) written transcripts only might be due to two main factors. First, the act of transcribing is such a painstaking enterprise (much more so than solely recording oral data) that the value of the transcribed data might unconsciously be equalled to the **time and effort** invested in transcribing it. Second, bearing in mind that corpus linguistics has primarily focused on **written genres**, linguists might – at least originally – have approached the spoken mode with the expertise, concepts, and methods they had previously developed for written corpora. For example, transcripts, just like written corpora, are easily searchable by means of corpus tools such as *WordSmith Tools*. Technological improvements enabling more flexible analyses of audio material, the focus of attention has recently turned back to consider the information contained in the audio signal.

While for some types of analyses, the unavailability of audio recordings might not be as much of an issue, "[a] major part of the problem behind the use of transcripts for language research is that the text of a transcript is always an incomplete and interpreted record of the original

[65] The mark-up conventions are available at: https://www.univie.ac.at/voice/documents/VOICE_mark-up_conventions_v2-1.pdf (last accessed 21/02/2017).

[66] As Edwards (2001:332) underlines, "a pause may seem longer if embedded in rapid speech than if embedded in slower speech. […] The perceived length of a given pause is also affected by its location in the discourse. It may seem longer if it is within an utterance than between turns by different speakers."

interaction" (Kendall 2008:337; see also Edwards 2001). Therefore, transcripts "should always be used in conjunction with data recordings and any supplementary data and resources available" because **recordings offer a direct, nearly unadulterated, access to the original linguistic production** (Jenks 2011:4–5).

In the framework of an analysis into fluency and disfluency, the availability of the primary recorded data is obviously absolutely essential because the recordings contain crucial data that cannot be easily encoded (such as speech rate, intonation, or pauses). As I will discuss below (Sections 2.3.2 and 2.3.3), two techniques can be used to improve and strengthen the link between the audio recordings and the written transcriptions, namely time alignment and annotation.

2.3.2 Time alignment

As pointed out in the first section of this chapter, spoken corpora have gained not only in size, but also in diversity. However, the value of a spoken corpus is definitely not restricted to size and diversity: the care and faithfulness in the act of encoding spoken data can be equally valuable (Section 2.3.1). Two other factors, namely time alignment and linguistic annotation, are also crucially important for spoken corpora because they **enrich the amount of information** of spoken corpora and allow for measurements and investigations that would not be possible without them.

2.3.2.1 What is time alignment?

Time alignment consists of the **mapping** of an audio recording and its corresponding written transcription through the creation of **virtual temporal anchors**. Basically, the transcription first needs to be segmented into "units", and the beginning and end of each unit is then attributed its timed equivalent in the recording, thereby weaving a web of links between the two. At the end of the process, it is possible to directly play a specific part of the transcription (see e.g. Adolphs & Carter 2013; Ballier & Martin 2015; Campoy & Luzón 2007; Dister & Simon 2008; Kendall 2008) [67].

An excerpt of the time aligned **Santa Barbara corpus** is shown in 2-4: the first two columns indicate the beginning and end of the text in the last column. Other references of time aligned corpora are included in Table 2-7 below.

[67] For further – technical – details, see e.g. Beaufort and Ruelle (2006); Brognaux *et al*. (2012a; 2012b); and Goldman (2011).

2-4: The Santa Barbara Corpus (SBC003 Conceptual Pesticides)[68]

```
0.00   1.01   MARILYN:            (Hx)   [Okay].
0.30   1.65   ROY:                       [Do you have a par]ticular,
1.65   2.10                      um,
2.10   4.00                      .. [use for the] red peppers,
2.15   3.05   PETE:                 [XXX X]
4.00   6.31   ROY:                as opposed to the yellow or green  pepp[ers].
6.11   6.51   MARILYN:                                                  [No] no,
6.51   7.69                      it was all .. salad peppers.
```

Time alignment of spoken corpora may be achieved **manually** by a human expert researcher, though this technique may quickly become very time-consuming (from 130 to 800 times the recording time (Brognaux *et al.* 2012b)). A number of **tools** to (partly) automatise the process are also available: EasyAlign (Goldman 2011) and SPPAS (Bigi 2015) are examples of user-friendly tools with a graphical interface; HTK (Young *et al.* 2013) and Julius (Lee, Kawahara & Shikano 2001) are examples of Hidden Markov Model (HMM) based recognition toolkits. Train and Align (Brognaux *et al.* 2012b) is a mixed tool that combines a graphical interface with HTK methods. Needless to say, it is also possible (and advisable) to have automatic alignment followed by a phase of manual correction.

An increasing range of **software** has also been developed to access the information contained in time aligned spoken corpora – Praat[69] (Boersma & Weenink 2013), ELAN[70] (Sloetjes & Wittenburg 2008), EXMARaLDA[71] (Schmidt & Wörner 2014) to cite but the most well-known.

2.3.2.2 What are the units of segmentation?

Depending on the research objectives, several units of segmentations are conceivable (see esp. Dister & Simon 2008; also Crookes 1990):

- **segments of identical length**: 5 or 10-second segments for example. The drawback is that speech is interrupted randomly and often in the middle of utterances or words;
- **automatically detected segments**, such as segments between unfilled pauses. The segments may, however, not always correspond to a linguistic unit (utterance, speech act etc.);

[68] The Santa Barbara Corpus is available online at: http://www.linguistics.ucsb.edu/research/santa-barbara-corpus#SBC001 (accessed 2/03/2017); the excerpt comes from the TRN format of the transcription (available at: http://www.linguistics.ucsb.edu/sites/secure.lsit.ucsb.edu.ling.d7/files/sitefiles/research/SBC/SBC003.trn (last accessed 2/03/2017).

[69] http://www.fon.hum.uva.nl/praat/ (last accessed 2/03/2017).

[70] http://tla.mpi.nl/tools/tla-tools/elan/ (last accessed 2/03/2017); Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands.

[71] http://exmaralda.org/en/ (last accessed 2/03/2017).

- **speech turns**, though in some speaking styles, they are far longer than in others (and the very definition of this unit is not always very clear and still sparks off lively debates);
- **prosodically-defined units** such as the "intonation period" (Lacheret & Victorri 2002) or "segments framed by major frontiers" (Mertens 1997);
- other types of linguistically-defined units such as the **T-unit** (as used in Lennon (1990)); the **AS unit** (Foster, Tonkyn & Wigglesworth 2000), which was used e.g. in De Jong (2016); or the **basic discourse unit** (as in Goldman *et al*. (2010));
- **speech acts**;
- **words**;
- …

Typical units of segmentation of large corpora are the utterance-, word- and phoneme-level (see some examples in Table 2-7).

Corpus	**Time alignment**
Santa Barbara Corpus	Utterance-level
Switchboard Corpus[72] (SWB)	Word-level
Wildcat Corpus of Native- and Foreign-Accented English[73]	Word- and phoneme-level
PELCRA[74]	Utterance-level
Machine Readable Spoken English Corpus[75] (MARSEC)	Word-level
Saarbrücken Corpus of Spoken English[76] (SCoSE)	Utterance-level

Table 2-7: Some time aligned spoken corpora

The choice of a unit of segmentation has far-reaching effects on the analysis of the aligned data. In many cases, the segmentation unit is actually used as an **artifact for the annotation of the data**: the unit of segmentation simply functions as unit of annotation. Researchers

[72] https://catalog.ldc.upenn.edu/ldc97s62 (last accessed 2/03/2017).

[73] http://groups.linguistics.northwestern.edu/speech_comm_group/wildcat/transcriptionalignment.html (last accessed 2/03/2017).

[74] http://pelcra.pl/new/time_aligned_pl_27 (last accessed 2/03/2017).

[75] http://www.reading.ac.uk/AcaDepts/ll/speechlab/marsec/ (last accessed 2/03/2017).

[76] http://www.uni-saarland.de/lehrstuhl/engling/scose.html (last accessed 2/03/2017).

considering time aligning their data may also want to take into account the size of the corpus and to reflect on the degree of **re-usability** of the aligned data – note in this respect that some tools, such as EXMARaLDA, allow for **more than one unit of alignment** (e.g. word alignment and segments between pauses).

Having outlined the main principles of time alignment, it is necessary to address its (few) limitations and (many) advantages.

2.3.2.3 What are the advantages and limitations of time alignment?

There are two major limitations to time alignment of spoken corpora. First and foremost, although tools exist to automate the procedure, the manual alignment in the editing phase may be tremendously **laborious and time-consuming**. Of course, the amount of time needed for manual corrections of automatic alignment varies depending on numerous factors, such as the quality of the recordings (the better the recording, the lower the time needed); the unit of segmentation (smaller units typically require more time); the type of speech (monologic speech is more easily and accurately aligned than interactive speech); the type of speakers (children vs. adults; native speakers vs. learners; strong vs. weak accented speakers etc.); the length of the audio recording etc.

Moreover, given the fact that segmentation units are often linguistic in nature, the segmentation step generally **requires advanced linguistic knowledge**. The segmentation into phones is, for example, the prerogative of experienced linguists. Besides, the use of some – fortunately not all – tools also requires **computational or programming skills**, which not all linguists have.

These arguments are, however, quickly counterbalanced by the advantages time alignment can offer.

The main argument in favour of time alignment is nicely summarised by Mello (2014:28): "Today, for a well-informed study of spontaneous speech, transcription is not nearly sufficient – actually, **transcription offered on its own can be a trap and is certain to misguide a researcher off track** in pursuit of language understanding and description, since his/her object of study would be written language (transcription), not oral language" (my emphasis). Transcriptions, as we have seen above, are only a *selection* of the information carried by the speech signal. In other words, part of the data is lost during the conversion from audio signal to written text. In a way, time alignment enables the researcher to "recover" what was lost in transcription. Having a direct access to the primary audio data, indeed, enables more reliable analyses. Consider, for example, excerpts 2-5 and 2-6 (both from the native-speaker corpus LOCNEC). Based on the transcription, they could be interpreted in different ways.

In example 2-5, the presence of a tag between *you know* and *it's really* gives the reader the impression that there is a break between the two (and the line break emphasises this impression here). This illusion of a break could arguably tip the scales in favour of the interpretation of *you know* as a discourse marker. However, no unfilled pause has actually been marked in the transcription: there is thus a clear lack of evidence to interpret this occurrence of *you know*. As regard the two *you*'s, there is no clear clue either in the transcribed data to say if it is a repetition or not.

2-5: LOCNEC - EN022-F

```
I was just on my way down you weren't there when I called you you know <overlap />
it's really it can be quite (erm) you can have problems as well
```

Does the speaker use *you know* in the literal sense as a main clause introducing a sub-clause (*you know it's really…*) or does the speaker use a discourse marker (*you know | it's really*)? And does the speaker repeat the pronoun *you* (*when I called | you you know*) or is the first *you* the direct object of the verb *call* (*when I called you | you know*)?

In example 2-5, the presence of a tag between *you know* and *it's really* gives the reader the impression that there is a break between the two (and the line break emphasises this impression here). This illusion of a break could arguably tip the scales in favour of the interpretation of *you know* as a discourse marker. However, no unfilled pause has actually been marked in the transcription: there is thus a clear lack of evidence to interpret this occurrence of *you know*. As regard the two *you*'s, there is no clear clue either in the transcribed data to say if it is a repetition or not.

2-6: LOCNEC - EN012-F

```
it's harder than I thought I thought it would be easier
```

From a fluency point of view, example 2-6 could be analysed in two different ways. Does the speaker produce two immediately adjacent *I thought* in two different utterances (*it's harder than I thought | I thought it would be easier*)? Or does the speaker stop his utterance mid-way and repeat the beginning of the next utterance (*it's harder than | I thought I thought it would be…*)? There is no element in the transcribed context (e.g. pause in the middle of the two *I thought*, or before the potential repetition) that could help the researcher interpret this excerpt with 100% certainty.

The only way to disambiguate such cases is by going back to the sound file because it gives access to other clues such as pronunciation, prosody or potentially untranscribed (or very small) silent pauses. Without time aligned data, it might quickly become very tempting to classify those cases without listening to the corresponding sound as it is very cumbersome and time-consuming to find the exact three or four milliseconds where these words were uttered in a 15 to 20-minute long file. But, with time aligned transcriptions, all it needs to have access to the sound is a couple of mouse clicks. The researcher can then establish with certainty that example 2-5 is to be interpreted as a discourse marker without repetition of *you*

(*when I called you* | *you know* | *it's*) and that example 2-6 can be analysed as two separate adjacent clauses that coincidentally begin with the two same words (*it's harder than I thought* | *I thought it would be easier*).

Besides the important gain in **reliability of analysis**, time alignment also has consequences in terms of **research possibilities**. Researchers analysing prosody, accent, stress, pausal or temporal phenomena have to toil arduously without time aligned data, but these elements are far more easily accessible (and/or reliably measurable) with aligned corpora: **broader research perspectives** emerge with time aligned data. For example, whereas in non-aligned corpora such as LINDSEI, it is the transcriber's *perception* of unfilled pauses that can be analysed, it is the actual production of unfilled pauses that may be investigated in time aligned corpora.

Lastly, with time aligned data, "it is possible to ameliorate some of the problems inherent in representing speech in text" (Kendall 2008:342). The representation of, e.g. overlaps, pauses and pause length is much clearer, and their transcription often even becomes "unnecessary" as the information "can be reconstructed from the audio itself" (2008:343).

Today, Dister and Simon (2008:16) claim, there is no technical reason anymore not to transcribe oral data in a time aligned fashion.

2.3.3 Linguistic annotation

Linguistic annotation, which is "essentially a development of the transcription stage" (Adolphs & Carter 2013:13), can be defined as "the practice of **adding interpretative, linguistic information** to an electronic corpus of spoken and/or written language data" (Garside, Leech & McEnery 1997:2; my emphasis). More practically, "[t]he task of annotating can be seen as consisting of assigning a label to an element or an interval in the data, where the label marks a specific event in the text or the speech signal" (Delais-Roussarie & Post 2014:47): such "events" can be linguistic in nature – in which case, the label marks a linguistic unit such as a word or phoneme – or it can be paralinguistic – such as silent pauses, or changes in tempo for example. Besides, the term "annotation" can be used to refer to the **end-product** of the practice of annotating: it also pertains to the symbols that are attached to, linked with, or interspersed with the written representation of the linguistic material (Garside, Leech & McEnery 1997).

The linguistic annotation of a corpus is fundamentally distinct from **corpus mark-up**: corpus mark-up "provides relatively verifiable information regarding the components of a corpus and the textual structure of each text" (McEnery, Xiao & Tono 2006:29). By contrast, **corpus annotation**, which is often used as a cover term to refer to **parsing** (i.e. the syntactic analysis of a corpus into its constituents), POS **tagging** (i.e. the allocation of a part-of-speech label to each word), and other forms of annotation (e.g. semantic, prosodic, or error annotation), is

concerned with **interpretative linguistic information**. Interpretation, as Leech (1997:2) stresses, is an intrinsic property of the act of annotating (and of annotations): "[t]here is no purely objective, mechanistic way of deciding what label or labels should be applied to a given linguistic phenomenon". Decisions have to be taken before setting out to annotate a corpus, but also during the annotation process itself (Garside, Leech & McEnery 1997:2–3). Such decisions, as well as recommendations on how to annotate difficult cases, should ideally be documented in an **annotation manual** (see e.g Kübler & Zinsmeister 2014). The manual can not only be used by the annotator during the annotation process (which will increase the consistency of the annotations), but also by the researcher during the analysis (which will improve the quality and reliability of the interpretations).

In spoken corpus linguistics, **the distinction between the representation of the linguistic material and the annotations is sometimes not obvious** and certainly not watertight (Leech 1997:3–4). For example, prosodic labelling of stress or intonation, or of "non-standard" pronunciation, is at one level a representation of the spoken data, and at another level, an interpretation of the same data through the filter of auditory perception. Likewise, indicating silent pauses in a transcription aims at accurately representing the original production, but, in most cases, also depends on the auditory perception, and interpretation, of the audio signal. Note, however, that stand-off annotation systems (which I will be using for the corpus analysis, *cf*. also Sections 2.3.3.3 and 4.2) have the advantage of making clearer this distinction between the linguistic material and the annotations, as compared to inline annotations.

2.3.3.1 The importance and standards of corpus annotation

Although annotating a corpus is known for being extremely **time-consuming** and constrained by the needs of the researcher, the size of the corpus, the tools and manpower available (e.g. Hedeland & Schmidt 2012; Leech 1997), it adds substantial value to a corpus. Leech (1997:2) writes: "[c]orpus annotation is widely accepted as a crucial contribution to the benefit a corpus brings, since it enriches the corpus as a source of linguistic information for future research and development".

There are at least four advantages in annotating a corpus (McEnery 2003; see also e.g. Garside, Leech & McEnery 1997; McEnery, Xiao & Tono 2006), the two main of which I summarise below.

Corpora are only useful when the **information** that they store **is easily and accurately retrievable** (e.g. Delais-Roussarie & Post 2014; Garside, Leech & McEnery 1997; Kübler & Zinsmeister 2014). Consider, for example, a researcher who wants to analyse reformulations or false starts in a corpus of learner speech. The corpus in its raw transcribed version contains no direct information (i.e. a textual indicator of some sort) to extract the concordances of these phenomena. In other words, the raw electronic version of the corpus is insufficient to analyse such linguistic features: to extract this type of information, the researcher first has to

build in information within the corpus by adding annotations. Only then will he/she be able to start quantifying the phenomena under inquiry. As such, annotation also **increases the range of phenomena that can be analysed**.

Besides, an annotated corpus is a more valuable resource than a raw corpus because many **annotations can be re-used** (e.g. Garside, Leech & McEnery 1997; Ide & Suderman 2007; Kübler & Zinsmeister 2014). Grammatical annotations (such as POS tags or syntactic annotations) can easily be handed down to other researchers, who may use them for completely different purposes. The argument of re-usability is very powerful indeed, since "corpus annotation tends to be an expensive and time-consuming business. We do not want to waste resources by 're-inventing the wheel' time and time again [...]" (Garside, Leech & McEnery 1997:5).

For annotations to be accurate, retrievable and re-usable, Leech (1997:6–8) established **six practical guidelines**, or "standards of good practice", that should be borne in mind by corpus annotators (and corpus users:

- **Recoverability**: it should always be possible, and easy, to remove the annotations from an annotated corpus and to revert back to the raw corpus;
- **Extractability**: it should be possible to extract the annotations by themselves from the text;
- **Documentation**: documentation should be available to the corpus user, with detailed information on (1) the annotation scheme, (2) how, where and by whom the annotation was carried out, and (3) an evaluation of the quality (i.e. consistency and accuracy) of the annotations;
- **Caveat emptor**: the corpus user should be made aware that corpus annotation is not infallible, it does not come with a guarantee, but it is offered to the research community as a potentially useful resource;
- **Theory-neutral**: annotation should be based as far as possible on consensual or theory-neutral principles;
- **Standard**: no annotation scheme should be considered as an absolute standard. Annotation schemes are always developed with practical reasons in mind, though convergent annotation principles should also be encouraged.

2.3.3.2 Manual vs. automatic annotation

As summarised in Table 2-8, there are **three basic methods** for annotating a corpus. Corpus annotation can be achieved **fully manually** by a human annotator, usually when no annotation tool is available or when the phenomena to be annotated are very specific (e.g.

reformulations and false starts). As manual annotation is very time-consuming (it has also been claimed to be more accurate (e.g. Hunston 2002; Kübler & Zinsmeister 2014), though accuracy obviously depends on, e.g., what is annotated), it is typically only feasible on corpora of small(er) size.

Manual	**Automatic**	**Semi-automatic**
Human annotator(s) only	Based on methods from computational linguistics	First automatic annotation, then manual post-correction (or interactive interface between the human annotator and the computer program)
Better suited for small corpora	Can be applied to large corpora	Can be applied to large(r) corpora
Very time-consuming, but more accurate for some phenomena	Quick and consistent, but errors may occur	The researcher can edit errors that were produced by the automatic method
All kinds of phenomena can be annotated	Not all phenomena can be annotated automatically	All kinds of phenomena can be annotated

Table 2-8: The three methods for annotating

Alternatively, corpus annotation can be performed **fully automatically** by running predefined probabilistic algorithms on the data (e.g. UCREL's automatic grammatical analysis[77]) and using methods derived from machine learning. Automatic annotation can be easily and rapidly applied to large sets of data and the output is consistent with the rules that have been applied (i.e. the output is reliable). However, the annotations may not always be accurate enough for a particular purpose (errors may occur).

Corpus annotation may also be undertaken **semi-automatically** (computer-assisted method): in this hybrid method, a human annotator goes through the automatically generated annotations and edits them, possibly using an interactive interface with the computer. This method is slower than the automatic annotation, but it is more accurate, and can be applied to large(r) corpora. Some researchers (e.g. Sinclair (1992) in Baker (1997)) however argued that, while using human post-editors may increase accuracy, it also decreases the internal consistency of the annotated data:

> A computer will not deviate from its programming, whereas humans, due to inattention, boredom or overfamiliarity, make slips. Thus a single human post-editor might spot a mistake made by an automatic tagger 99 times out of 100, but would fail to notice every error, thus introducing a level of inconsistency into the data. […] although an automatically tagged corpus

[77] UCREL's POS tagging software is called CLAWS (which stands for Constituent Likelihood Automatic Word-tagging System); see http://ucrel.lancs.ac.uk/claws/ (last accessed 26/04/2017).

might contain a larger proportion of errors, at least those errors would remain consistent throughout the corpus.

(Baker 1997:243–244)

2.3.3.3 Inline and stand-off annotations

Besides these three annotating methods, there are **two fundamentally different ways of adding annotations** to corpora: inline (or embedded) annotation, and stand-off (or standalone) annotation[78] (see e.g. Delais-Roussarie & Post 2014; Leech 1997; McEnery, Xiao & Tono 2006; Palmer & Xue 2010; Rehbein, Schalowski & Wiese 2012; Schmidt 2003).

With **inline annotation**, the textual **material is interleaved with the annotations**. Annotational labels (or "tags") are interspersed next to the eligible element(s) within the primary data itself. This type of annotation can be used for written corpora as well as for spoken data. Notorious examples of spoken corpora with inline annotations are the London-Lund Corpus[79], the Lancaster/IBM Spoken English Corpus (SEC[80]) and the LINDSEI database (Figure 2-6 to Figure 2-8, respectively). As can be seen in the illustrations, inline annotations can be more or less endemic, take various forms, and be used to mark varied properties – part-of-speech, prosody, intonation, pauses etc. – up to a great level of detail. Note also that, to date, the annotation of most learner and native corpora is done using inline annotations.

BRO ⋆ CHÙRE ⋆ for ▮ 139 so I ▮DÍD it ▮ · 140 and ▮then ANÓTHER one ▮ –
141 and
b 142 ⋆[mhm]⋆
>A 141 ▮THÈN they ▷said ▮ 143 well ▮now that you've done THÉSE ▮ 144 and
they've been ▮SÒ SUCCÉSSFUL ▮ 145 we'd ▮like you to do our SÙPER ▮ ·
146 ▮ALPHA△MÀTIC ▮ 147 or ▮SÓMETHING ▮ 148 and ▮this is one of THÉSE ▮
149 that ▮goes SÍDEWAYS ▮ 150 and ▮FRÓNTWARDS ▮ 151 and EM▮BRÓIDERS ▮
152 and ⋆▮DÁRNS ▮ 153 and sews ⋆ ▮BÙTTONS on ▮
b 154 ⋆(– laughs) yes⋆
>A 155 – – and I ▮SÁID ▮ 156 well I ▮don't RÈALLY ▷think ▮ 157 I could ▮WRÌTE ▮ –
– 158 and this was a sort of ▮ninety-six page △BÒOKLET ▮ 159 ▮you KNÓW ▮
160 about ▮that BÌG ▮ ⋆–⋆ 161 [əm] I'd I'd ▮need to GÒ through ▮ 162 ▮each of
the
b 163 ⋆[m]⋆
>A 162 processes at △HÒME ▮ ⋆ · ⋆ 164 I don't think it will be ▮enough just to have

Figure 2-6: London-Lund corpus (paper version)

[78] Note also that, depending on the type of annotation, inline and stand-off annotations can be added manually, automatically or semi-automatically.

[79] http://corp.hum.ou.dk/itwebsite/corpora/corpman/LONDLUND/INDEX.HTM

[80] http://clu.uni.no/icame/manuals/SEC/INDEX.HTM

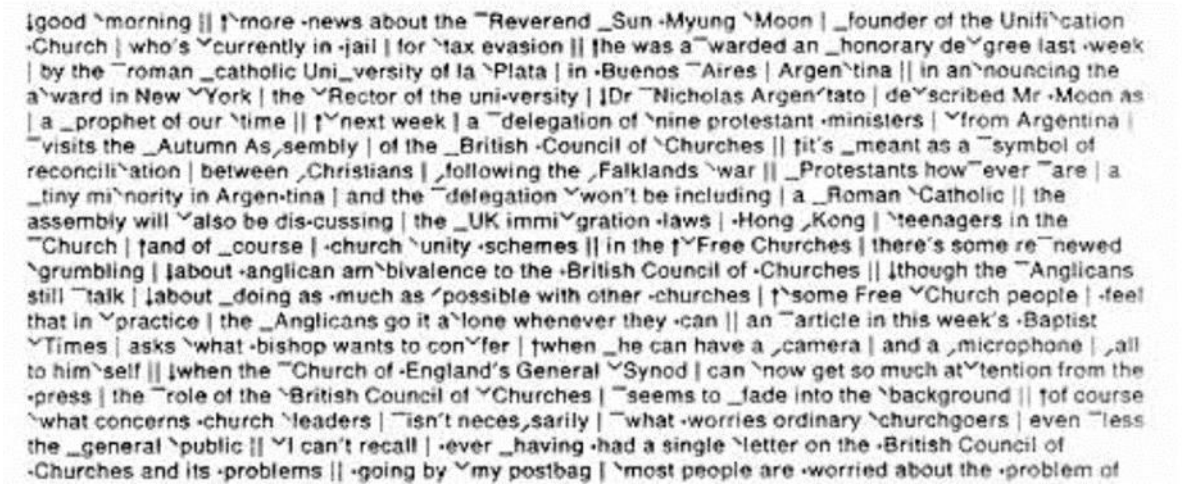


Figure 2-7: SEC corpus (prosodic version)

```
<h nt="FR" nr="FR012">
<S>
<A>  did you manage to find a topic that (er) </A>
<B>   yes <laughs> </B>
<A>  fine okay ... what are you <overlap /> going to tell me about  </A>
<B>      <overlap      />      I'm      going      to      talk      about
 a: a country I visited in </B>
<A>  oh lovely  </A>
<B>  which ins= impressed me .. well it's Canada in fact I went (er) yes I
on= only visited (erm) .. the <X> . the[i:] Eastern part so Quebec </B>
<A>  yeah </A>
```

Figure 2-8: LINDSEI (excerpt from the French component)

While inline annotations certainly have a great advantage in terms of **simplicity of application** (in many cases, a simple text editor is sufficient), they also show some serious **limitations**. First, it is not always easy (or even possible) to **annotate everything** inline. The annotation of phenomena that cover several words (such as repetitions) is, for example, much more difficult than the inline annotation of single units (e.g. word-class membership). Second, whereas inline annotations can achieve a great level of detail, **access to the primary text** and the **readability** of the primary data may become endangered, with potential consequences in terms of **search possibilities**. In this respect, consider the following excerpt from the electronic version of the London-Lund Corpus (Figure 2-9), where automatic searches of the corpus are rendered particularly challenging due to the number of annotations. For example, it seems quite challenging to look for the different realisations of, say, *well*, when many transcriptions are used (^w=ell#, ^well, *^w=ell#, well etc.).

```
1 1  1  10 1 1 B   11 ((of ^Spanish)) . graph\ology#                          /
1 1  1  20 1 1 A   11 ^w=ell# .                                              /
1 1  1  30 1 1 A   11 ((if)) did ^y/ou _set _that# -                         /
1 1  1  40 1 1 B   11 ^well !J\oe and _I#                                    /
1 1  1  50 1 1 B   11 ^set it betw\een _us#                                  /
1 1  1  60 1 1 B   11 ^actually !Joe 'set the :p\aper#                       /
1 1  1  70 1 1 B   20 and *((3 to 4 sylls))*                                 /
1 1  1  80 1 1 A   11 *^w=ell# .                                             /
1 1  1  90 1 1 A   11 "^m/\ay* I _ask#                                       /
1 1  1 100 1 1 A   11 ^what goes !\into that paper n/ow#                     /
1 1  1 110 1 1 A   11 be^cause I !have to adv=ise# .                         /
1 1  1 120 1 1 A   21 ((a)) ^couple of people who are !d\oing [dhi: @]       /
1 1  1 130 1 1 B   11 well ^what you :d\/o#                                  /
1 1  1 140 1 2 B   12 ^is to - - ^this is sort of be:tween the :tw\/o of     /
1 1  1 140 1 1 B   12 _us#                                                   /
1 1  1 150 1 1 B   11 ^what *you* :d\/o#                                     /
1 1  1 160 2 1 B   23 is to ^make sure that your 'own . !c\andidate          /
```

Figure 2-9: London-Lund corpus (electronic version)

The second option (though it is still in its infancy compared to inline annotation) is called **stand-off annotation** (four examples of stand-off annotation are provided in Figure 2-10 to Figure 2-13 below). As the name suggests, stand-off **annotations are stored outside the primary data**, leaving the original text preserved (it thus fully addresses the criticism of **readability** of the raw text). The downside is that searching for information may be more difficult. To avoid this issue, it is thus necessary to develop a way to refer back to the original text and to know which word the tag was applied to. The code "w23" could for example be used with the annotation tag to specify that the tag applies to the twenty third word in the text. In case of **time aligned spoken data**, the temporal anchors created for the alignment can be directly used as links between the text and its annotations.

Although the handling of stand-off annotations in time aligned corpora requires specially-designed tools, such as ELAN[81] (Sloetjes & Wittenburg 2008), Praat (Boersma & Weenink 2013) or EXMARaLDA (Schmidt 2001)[82], stand-off annotations have a number of non-negligible advantages:

- the **format of the tags is free** (tags used for inline annotation are generally more constrained in terms of format);
- the **number of annotation levels in the same text** is virtually unlimited, and it is possible, for example, to have both POS tagging (or alternative POS tagsets), parsing, and other types of annotations in the same file (i.e. one word can be annotated more than once and for different properties or purposes, as illustrated in Figure 2-12 and in Figure 2-13);

[81] http://tla.mpi.nl/tools/tla-tools/elan/ (accessed 28/02/2017).

[82] See e.g. Rohlfing et al. (2006) for a discussion of the strengths and weaknesses of these and other annotation tools.

- stand-off annotation allows **multiple overlapping hierarchies** (i.e. through the creation of mother and daughter tiers);
- stand-off annotation enables **new annotation levels** to be added without causing problems to the others;
- stand-off annotation allows **modifications at one level of annotation** without affecting the other levels;
- etc.

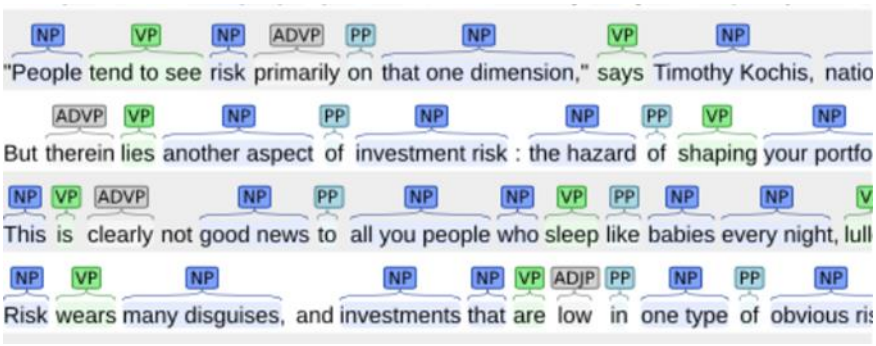


Figure 2-10: Stand-off POS annotation in brat[83]

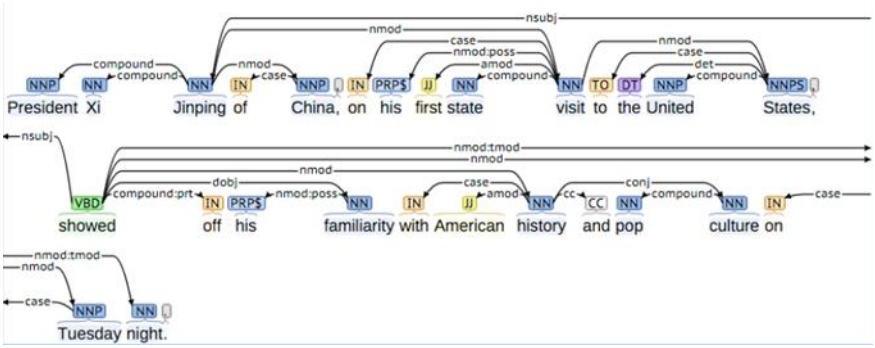


Figure 2-11 : Stand-off annotation of basic dependencies in brat[84]

[83] From: http://brat.nlplab.org/examples.html (accessed 28/02/2017).

[84] http://stanfordnlp.github.io/CoreNLP/ (accessed 20/03/2017).

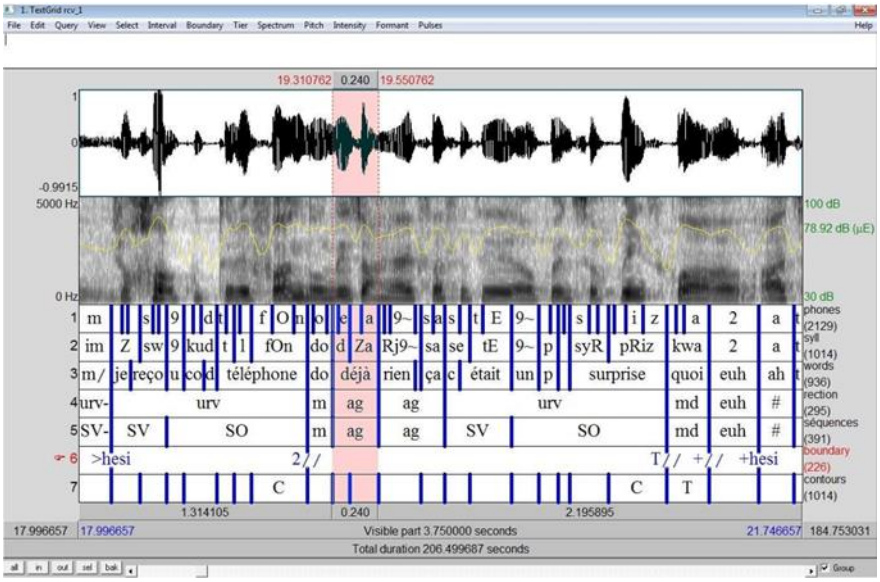


Figure 2-12: Stand-off syntactic annotation in Praat (from Tanguy et al. 2012:2)

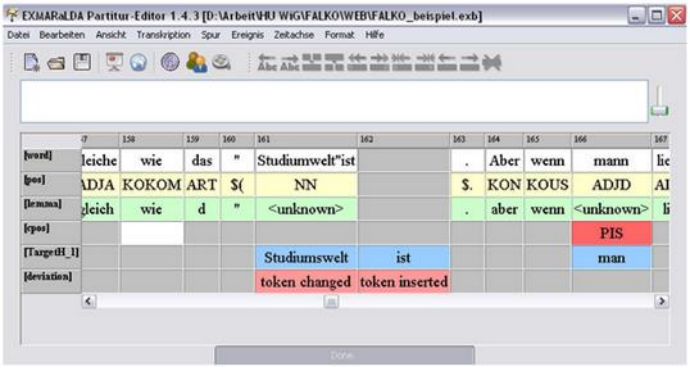


Figure 2-13: Multi-level stand-off annotation in EXMARaLDA[85]

2.3.3.4 Select overview of annotation systems of (dis)fluency phenomena

The annotation of (dis)fluency features has garnered quite a lot of attention from different fields and for different purposes. This section reviews some (dis)fluency annotation systems. It is by no means exhaustive but serves as a starting point for a discussion of their usability in the framework of the present study.

Levelt (1983), whose work primarily focused on repairs, was one of the first who attempted to find regularities in the patterning of disfluencies. He did not develop an annotation system of repairs as such, but tried to dissect their structure, and his work had a great influence on

[85] https://www.linguistik.hu-berlin.de/en/institut-en/professuren-en/korpuslinguistik/research/falko/tools (last accessed 20/03/2017).

the development of later annotation systems of (dis)fluency features. He analysed repairs as consisting in three main parts (as illustrated in Figure 2-14):

- the **original utterance,** which contains the *reparandum* (i.e. the item to be repaired) and the **moment of interruption,** or "the point at which the flow of speech is interrupted for 'editing'" (*ibid.*:44). The interruption can be delayed: in this case, the space between the reparandum and the interruption is called the ***delay** of interruption*;
- the second part is called the **editing phase,** and refers to a period of hesitation which may contain an *editing term*;
- the last part is the **repair**. Repairs can contain retracings of various spans and usually contain an alteration (except in the case of a *covert repair*).

Note that this model of repairs has been then extended to repetitions by Clark and Wasow (1998), who suggested a *commit-and-restore model* of repeated words. They divide repetitions into four stages: the initial commitment; the suspension of speech; the hiatus (Levelt's editing phase) and the restart (see Section 1.2.8 for more details and an illustration).

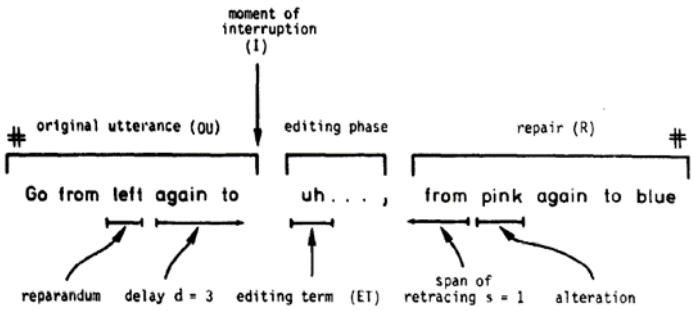


Figure 2-14: The structure of repairs (from Levelt 1983:45)

In the early 90s, **Shriberg** (1994) published an annotation system of disfluencies which is rooted in Levelt's and Clark & Wasow's structure of repairs and repetitions. Her perspective, however, was more explicitly normative as she considered **disfluencies as removable errors**: "[t]he DFs [disfluencies] considered are cases in which a contiguous stretch of linguistic material **must be deleted** to arrive at the sequence the speaker 'intended', likely the one that would be uttered upon a request for repetition" (Shriberg 1994:1). As can be observed in Figure 2-15 she also identified four main "disfluency regions", which she calls **reparandum** (RM), **interruption point** (IP), **interregnum** (IM) and **repair** (RR), respectively. Starting from this basic structure and her assumption about the nature of disfluencies (i.e. removable errors), she developed an innovative and extensive annotation system made up of letters and symbols (Figure 2-16 to Figure 2-18).

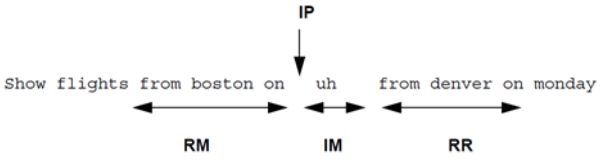


Figure 2-15: The four "disfluency regions" (from Shriberg 1994:8)

Symbol	Explanation	Example
Region-delimiting		
[ ]	onset RM, offset RR	(see all examples below)
.	IP	(see all examples below)
Syntactic-word		
r	repeated word	she she liked it [r . r]
s	word in substituted string	she my wife liked it [s . s s]
i	inserted word	she liked really liked it [r . i r]
d	deleted word	it was very she liked it [d d d .]
Extra-syntactic-word		
f	filled pause	she uh liked it / she uh he liked it [.f] [s . f s]
e	explicit editing term	she sorry he liked it [s . e s]
p	discourse marker	she liked well she liked it [r r . p r r]
Inter-sentence-word		
c	coordinating conjunction	she saw it and and she liked it [c . c]
Diacritics		
-	word fragment	she li- he liked it [s r- . s r]
~	misarticulated word	shle she liked it [r~ . r]
^	contracted word	she'd she'll like it [r^s . r^s]
=	substituted-string fragment	she thought highly she liked it [r s s= . r s]

Figure 2-16: Labelling system – pattern symbols (from Shriberg 1994:57)

Operation	Explanation	Example
RM Deletor	Deletes all words in reparandum. Relies on IP location. Does not rely on symbol type.	~~he likes~~ . sorry she uh likes [~~s — s~~ . e s f r]
f and e Deletor	Deletes filled pauses and explicit editing terms anywhere in the DF. Does not rely on IP location. Relies on symbol type.	he likes . ~~sorry~~ she ~~uh~~ likes [s r . ~~e~~ s ~~f~~ r]

Figure 2-17: Labelling system – correction operations (from Shriberg 1994:58)

Case	Definition and Example	Solution
Serial DFs	2 or more DFs are adjacent (do not overlap) Example: "he he liked uh loved it"	Treat as consecutive, individual basic DFs. he he liked uh loved it [r . r] [s . f s]
Complex DFs	2 or more DFs overlap. Example: "he she she liked it"	Treat as a complex of basic DFs in a hierarchical representation. See section for notational conventions. he she she liked it [R[s . s] . r]
Ambiguous DFs	DF has more than one possible analysis. See section for examples.	Apply 2 rules: 1) delete fewer over more words 2) assume less over more correspondence
Degenerate Cases	Labeling is impossible or inappropriate. See section for examples.	Flag cases but exclude them from analyses.

Figure 2-18: Labelling system – special cases (from Shriberg 1994:58)

Shriberg's seminal work has inspired many researchers, such as Pallaud, Rauzy and Blache (2013), who developed an annotation system of "interruptions" (*cf.* Figure 2-19), and Besser (Besser 2006; Besser & Alexandersson 2007). Likewise, Eklund (2004), and Moniz (2013) also largely based their annotation systems on her labelling method (*cf.* Figure 2-20 and Figure 2-21, respectively).

Reparandum		
Reparandum Type	***R***	Temporary interuption
	I	Definitive interruption
Reparandum_category	***W***	Word reparandum
	P	Phrase reparandum
Lexical_type	***tw***	Tool word
	lw	Lexical word
Break_type B		
	no	No interval
	sp	Silent pause (> 200ms)
	fp	Filled pause
	dc	Discursive connector
	ps	Parenthetical statement
	rt	Tructation repetition
Reparans RA		
Reparans_position_type	***nr***	No restart
	wr	Word restart
	dr	Determinant restart
	pr	Phrase restart
	or	Other restart
Reparans_type	***co***	Continuing the item
	wc	Reparing without change
	rp	Repairing through repeating
	rc	Repair with change in the truncated word
	rm	Repair with multiple change

Figure 2-19: Annotation of interruptions (from Pallaud, Rauzy & Blache 2013)

Disfluency	Description	Symbol	Disfluency subclasses	
			Symbol	Description
UP	Unfilled Pause (Silence)	u< >u	(none)	(none)
FP	Filled pause (Filler Word)	f< >f	ff<	Utterance initial
PR	Prolongation	p< >p	*(x)* *(x-)* *(-x-)* *(-x)* *#* *(x)*	Segment Word initial Word medial Word final Lexeme border Suppressed segment
EET	Explicit Editing Term (Self-correction)	eet	Eet1 Eet2 . . . eet*n*	First word Second word . . . *N*th word … in eet
TR	Truncation	/	(none)	(none)
MP	Mispronunciation	~	(none)	(none)
REP	Repair	[+]	[	Beginning of repair
			+	Interruption point
			]	End of repair
			r*n*	Repeated word *n* in *Reparans*
			d*n*	Deleted word *n* in *Reparandum*
			s*n*	Substituted word *n* in *Reparans*
			i*n*	Inserted word *n* in *Reparandum*

Figure 2-20: Labelling symbols (from Eklund 2004:212)

Labels	Description	Examples	Annotation
<>	Auto-corrected	Sequences of disfluencies	< … >
.	Interruption point	Moment when the speaker interrupts to repair his/her speech	< n.n >
f	Filled pauses	*Ou pode estar* ***<%aa>*** *trancada* (or it can be <%uh> closed)	<f>
lm	Segmental prolongations	***de=*** (of=) pronounced as [di:]	<lm1.>
r	Repetitions	*e* ***<vocês sabem>*** *vocês sable que* (and <you know> you know that)	<r1 r2.r1 r2>
s	Substitutions	*São* ***<os>*** *o conjunto dos ~X, ~Y* (they are <the> the set ~X, ~Y)	<s1.s1>
d	Deletions	*Vai haver uma série de resultados,* ***<vamos chamar>*** *portanto, nós tínhamos a noção de ~R* (there will be a series of results, <let's call> therefore, we had the notion of ~R)	<d1 d2>
i	Insertions	*<em +que é que>* ***em que*** *<u>midida</u>* ***é que*** *o padrão é útil ?* (in what way is the pattern useful ?)	<r1 r2.r1 il r2>
e	Editing expressions		<s1 e1 e2 f.s1>
–	Word fragments	***<comp->*** *complementar (<addi-> additional)*	<r1-.r1>
~	Mispronunciations	*Pode-nos* ***<servir~>*** *servir* (can <serve~> serve us) pronounced as [ʃir'nir] instead of [sir'vir]	<r1~.r1>

Figure 2-21: Disfluency annotation (from Moniz 2013:31)

Besides these influential annotation systems, many researchers have developed their own annotation method and tags to mark the elements they wanted to analyse in a particular piece of research. It is probably not an exaggeration to say that there are as **many annotation systems** as there are studies on (dis)fluency. Examples include – but are definitely not restricted to: Bear *et al*. (1993); Blackmer and Mitton (1991); Hedeland and Schmidt (2012); Honal and Schultz (2003); Maclay and Osgood (1959); Rodriguez, Torres, Varona (2001), who worked on Spanish disfluencies; and the Penn Treebank annotation[86] (Taylor, Marcus & Santorini 2003). The description of each individual system is, unfortunately, beyond the

[86] The Penn Treebank *includes* inline annotation of some (dis)fluency phenomena (incomplete utterances, fillers, explicit editing terms, discourse markers, coordinating conjunctions, asides and repairs), but is certainly not restricted to them: it also includes POS-tag annotation, "skeletal" parsing and parsing of predicate-argument structure.

scope of this thesis, but I will, however, underline some overall **methodological shortcomings** from the perspective of the goal of this thesis.

As Rehbein *et al*. (2012) pointed out, not all annotation systems are **designed for spoken data** in the first place. In early studies, for example, annotation systems were **developed with written text in mind**, thereby running the risk of failing to capture characteristics of spoken language. Moreover, some concepts cannot easily be transferred to spoken language, such as the notion of sentence. In recent years, however, researchers have become increasingly aware of the danger of this bias and focused on the description and development of spoken units (e.g. AS units (Foster, Tonkyn & Wigglesworth 2000), basic discourse units (Degand & Simon 2005; 2009)) and speech phenomena. Besides, although increasingly more annotation systems may be designed for, and on the basis of, spoken data, they may not be adequate for **time aligned** corpora. The notion of interruption point in Shriberg's system (1994, *cf.* above) is, for example, problematic for time aligned data because it corresponds to nothing in the audio signal. In addition, not all (dis)fluency annotation systems are relevant and effective for **stand-off annotation**: the Penn TreeBank disfluency annotations (Figure 2-22), for example, follow an inline design. Consider also Figure 2-23, which follows and XML notation.

A: he's pretty good. / He stays out of the street / {C and, } {F uh, } if I catch
him I call him / {C and } he comes back. / {D So } [he, + he's] pretty good
about taking to commands [and + –
B: {F Um. } /
A: – and] things. /
B: Did you bring him to a doggy obedience school or –
A: No – /
B: – just –
A: – we never did. /
B: – train him on your own / {C and, } -/
A: [I, + I] trained him on my own / {C and, } {F uh, } this is the first dog I've
had all my own as an adult. /
B: Uh-huh. /

Figure 2-22: Example of Penn TreeBank disfluency annotation (from Taylor, Marcus & Santorini 2003:16)

But then to go back to the to th s something along those things.

But then to go back
<replace>
<RM>to the</RM>
<RS>
to
<sot>th</sot>
<stutter>s</stutter>
something
<RS>
</replace>
along those things.

Figure 2-23: Example of annotation in XML format (from Besser 2006:40)

The second shortcoming is related to the **number of (dis)fluency phenomena** annotation schemes encompass. Because (dis)fluency annotation schemes are often developed in the framework of a particular paper and/or to answer specific research questions, many systems thus tend to concentrate on a **limited range of phenomena at a time**. It is important to stress

that this does not affect the quality of the studies in question in any way: such coding schemes usually capture **very detailed** aspects of the phenomenon (phenomena) under scrutiny. However, the perspective on (dis)fluency adopted in the present study is broader. In this context, coding schemes focussing on a limited number of (dis)fluency phenomena present three potential disadvantages: (1) they do not cover all (dis)fluency phenomena that I plan on analysing; (2) they do not cater for the annotation of complex (dis)fluency patterns (such as an unfilled pause within a repetition); and/or (3) due to their high level of detail, they involve a great deal of manual work, which makes them less suitable for **large corpora**.

Another aspect has to do with the replicability of annotation schemes. Few annotation systems are **sufficiently described** (e.g. in a coding book) with clear **examples** and illustrations of problematic cases. This considerably impedes accurate replication. In addition, and despite Leech's advice (see 2.3.3.1), in many cases, no proper **evaluation of the coding system** is presented either (Dybkjaer & Bernsen 2000).

Fourthly, it is unclear to what extent existing (dis)fluency annotation systems can be **applied to both native and learner data**, when they were originally developed exclusively for one *or* the other. If a "native system" is applied to learner data, researchers run the risk of failing to capture potential L2 specificities. If a "learner system" is applied to native data, researchers might run the risk of magnifying their disfluency bias. Besides, to my knowledge, no existing (dis)fluency annotation system allows for the annotation of native French, native and learner English and Belgian French Sign Language. Given the fact that the ARC fluency project aims to enable some comparison between (dis)fluency phenomena across the aforementioned languages and modalities, the interoperability of the (dis)fluency annotation system was deemed essential.

Keeping in mind the aims of the present study and the properties of the data, it appeared from the above discussion that a new (dis)fluency annotation system should be developed that takes the best out of the previously mentioned annotation schemes. But before concluding this section, three additional **caveats** regarding the use of corpus annotations in general are to be considered.

The first caveat is that annotation **imposes an analysis upon the corpus user**. While it has to be acknowledged that annotation is interpretative in essence (and often closely linked to the research objectives of the researcher), corpus users may very well have their own interpretations (or they may also simply ignore the annotation). Besides, "just leaving a corpus unannotated does not mean that there is no process of interpretation occurring when the corpus is analysed. […] The analysis still happens, it is simply hidden from clear view" (McEnery, Xiao & Tono 2006:31; *cf.* also McEnery 2003). Corpus annotation should thus be recognised as a strength rather than a weakness as it provides "an objective record of an explicit analysis open for scrutiny" (McEnery, Xiao & Tono 2006:31), which, in many cases, also makes the analysis easier to perform and to retrieve.

Another criticism is that annotation may sometimes produce **cluttered corpora**. Hunston (2002:94) claims that "[h]owever much annotation is added to a text, it is important for the researcher to be able to see the plain text, uncluttered by annotational labels". Stand-off annotation, however, does not clutter the text at all, contrarily to some types of inline annotation[87], because the annotations are not interleaved within the text, but on a different layer (or "tier"). For this reason[88], I have chosen to annotate the data in a stand-off design.

A further caveat of the use of corpus annotation pertains to **accuracy and consistency**. As we have seen, neither manual nor (semi-)automatic methods produce error-free results. Human annotators may cause a slight drop in consistency, but, depending on the type of annotation, fully automatic methods are either impossible or not as accurate as manual annotations. McEnery and colleagues (2006:32) claim that "while inconsistency and inaccuracy in analyses are indeed observable phenomena, their impact upon an expert human analysis has been exaggerated" and they advise that "the human analyst and the machine should complement each other, providing a balanced approach to accuracy and consistency that seeks to reduce inaccuracy and inconsistency to levels tolerable to the research question that the corpus is intended to investigate". Bearing in mind their advice, I have opted for a semi-automatic method of annotation, where most (dis)fluency features are first annotated automatically, with a manual post-correction.

The (dis)fluency annotation system that was used for this dissertation (including examples and an inter-rater reliability analysis) is set out in Chapter 4, Section 4.2.

[87] As pointed out by McEnery et al. (2006), even if inline annotations are used, they do not necessarily obscure the patterning of words either since most corpus tools (e.g. WordSmith Tools) make it possible to suppress annotation tags in concordance lines (i.e. only the plain text is visible in the search results).

[88] Another reason for choosing stand-off annotation is that this type of annotation is better suited for aligned data than inline annotation.

2.4 (DIS)FLUENCY ASSESSMENT AND SPOKEN CORPORA

The previous sections of this chapter introduced spoken (learner) corpora, these large databases of recorded and transcribed language that may be exploited to obtain empirical measurements of a panel of (dis)fluency features. Such corpora are generally used to investigate the *productive* side of (dis)fluency – i.e. (dis)fluency *on the part of the speaker*. The *perceptive* side of (dis)fluency – (dis)fluency *on the part of the listener* – has long been the prerogative of the field of language testing and assessment. Quite recently, however, bridges have been built between the two research communities, with increasingly more researchers attempting to align rater perception of speakers' (dis)fluency with observable and quantifiable aspects of their performance, as captured by corpus-based measurements.

Structurally, the last section of this chapter consists in two thematic sections. I first discuss some general aspects related to the assessment of learner and native (dis)fluency, including rating scales, and the number and experience of the raters (Section 2.4.1). Then, in Section 2.4.2, I review some of the main findings that have emerged from the alignment of (dis)fluency assessment scores with learner corpus measurements.

2.4.1 Testing and assessment

Language testing and assessment (LTA) is a subfield within applied linguistics that "is concerned with measuring the language proficiency of individuals" (Barker 2010:633) and that subsumes a wide range of testing and assessment contexts. Because both *testing* and *assessment* refer to "the systematic gathering of language-related behaviour in order to make inferences about language ability and capacity for language use on other occasions" (Chapelle & Plakans 2013:241), the two terms are often used interchangeably. Nevertheless, *testing* is generally restricted to institutional contexts while *assessment* tends to be used in a more general sense, referring to the process of data collection and interpretation (Callies & Götz 2015a; Chapelle & Plakans 2013). As explained by Barker, the **general aim of LTA** is "to measure a latent trait in order to make inferences about an individual's language ability. Language tests allow us to observe **behaviours which can be evaluated by attaching test scores** which provide **evidence for an individual's ability in a specific skill or their overall language competence**" (Barker 2010:633; my emphasis).

2.4.1.1 Testing and assessment of speech and (dis)fluency

Testing speaking is the youngest subfield within LTA: it was not until the Second World War that the development of speaking tests became a focus of interest (Fulcher 1996). In his monograph, Fulcher devotes the first chapter to the history of testing second language speaking and shows how early developments in the assessment of L2 speaking are, in fact,

intimately connected to political and military language needs, and how these needs have had "a deep impact upon the form and scoring of many modern speaking tests" (*ibid*.:1). For example, the use of a native-speaker norm dates back to the Foreign Service Institute (FSI) Oral Proficiency Interview (OPI)[89] (Fulcher 2003).

Rating scales provide the framework within which human raters score language performances[90]. They constrain raters' responses, often through scale descriptors that are associated with a fixed number of scale bands. There is, however, something of a tension between the "simplified orderliness of the rating scale" (Lumley 2005:248), which necessarily underrepresents the complexity of the linguistic performance, and raters' reactions to that performance. The challenge for raters is thus to "reconcile their possibly idiosyncratic, intuitive, or nonlinear impression of an L2 performance with rating scale specifications" (Isaacs & Thomson 2013:135).

Assessment of (dis)fluency has thus far mostly, if not exclusively, been aimed at **non-native speakers**. As underlined by Bosker *et al*. (2014:580), "[n]ative speakers are supposedly perceived as fluent by default even though they, too, produce disfluencies such as uhm's, silent pauses and repetitions"[91]. It is probably also for this reason that learners' (dis)fluency level is typically assessed in language tests with rating scales ranging from zero mastery to an end-point representing a well-educated native-speaker (Davies *et al.* 1999:153–154). This **idealised native speaker norm**, however, has long been the object of criticism: Bosker *et al*. (2014:609) for example claim that "a single ideal native fluency standard does not exist", and raters in Isaacs and Thomson's (2013) study reported being "uncomfortable" with the NS standard. In their study, one rater in particular argued that scales should allow successful non-native speakers to be at the top-end of the scale, i.e. that the scale should reflect raters' judgement about success and not about a speaker's first language (*ibid.*:154-155).

2.4.1.2 Spoken corpora and LTA

Corpora began to make inroads into language testing and assessment in the 1990s as a reference resource for test developers (Alderson 1996; Park 2014; Taylor & Barker 2008) and, since then, corpora have attracted increased attention from the LTA field. The potential of **learner corpora to increase transparency and consistency in the assessment of L2 performance** has recently become the topic of several publications, such as the edited

[89] According to Fulcher (2003:8), the OPI was the first published test of speaking. The FSI was set up following military needs in the 1950s (see Fulcher 2003:1-19 for a history of testing second language speaking).

[90] "Performance" is to be taken as an indicator of the underlying ability of a learner (Isaacs & Thomson 2013:135).

[91] A study conducted by Bosker et al. (2014) on the perception of (dis)fluency in learner and native speech, however, suggests that there is no significant difference in the way listeners weigh the (dis)fluency characteristics of native and non-native speakers.

volume *Learner Corpora in Language Testing and Assessment* (Callies & Götz 2015b). In the introduction, Callies and Götz (2015a:2–3) explain that:

> The use of learner corpus data and the application of methods and tools developed in corpus linguistics enable researchers and test-developers to take more data-driven approaches to the assessment of proficiency [and fluency] that is partially independent of human rating, thereby resolving the tension between the expertise of trained individuals, whose holistic ratings are inevitably influenced by subjectivity and variability, and more fine-grained text-centred descriptors.

More specifically, Callies *et al*. (2014) advocate a **threefold distinction of how learner corpora can be used in LTA**. They suggest the use of three criteria – (1) the way corpus data are actually put to use, (2) the aims and outcomes for LTA, and (3) the degree of involvement of the researcher in data retrieval, analysis and interpretation – to classify the use of learner corpora in LTA as "corpus-informed", "corpus-based" or "corpus-driven". Callies and Götz (2015a:1–2) summarise these three approaches as follows:

> In CORPUS-INFORMED applications, learner corpora can be used "throughout the cycle of planning, developing, delivering, and rating a language test" to inform test content or to validate human raters' claims in order to "reveal what language learners can do, which informs both what is tested at a particular proficiency level and how this is rated" (Barker 2013:136of). In CORPUS-BASED approaches, learner corpus data are explored to provide empirical evidence confirming or refuting a researcher's hypothesis. [...] Finally, CORPUS-DRIVEN approaches rely exclusively on computer techniques for data extraction and evaluation in that the questions and conclusions formulated by a researcher will be derived from what corpus data reveal when subjected to statistical analysis.

Most of the analyses provided in Chapter 7, which explores learner corpus data from the perspective of the CEFR fluency levels, qualify as corpus-based (and some as corpus-driven).

2.4.2 Relating assessed (dis)fluency levels and objective (dis)fluency measures

In **L2 (dis)fluency research**, a still relatively small, but growing, number of studies have tried to set quantitative findings on (dis)fluency production in relation to ratings of the perceived level of (dis)fluency (or proficiency) of the learner (e.g. Cucchiarini, Strik & Boves 2000; Cucchiarini, Strik & Boves 2002; Derwing *et al.* 2004; Freed 2000; Ginther, Dimova & Yang 2010; Kormos & Dénes 2004; Mora 2006; Rossiter 2009; Wennerstrom 2000). All these studies involved relating corpus-based measures of utterance fluency (generally temporal (dis)fluency measures), with measures of perceived (dis)fluency (i.e. listener ratings) in order to assess the relative contributions of different (dis)fluency measures to (dis)fluency perception.

Before reviewing the main results emerging from previous studies relating assessed (dis)fluency levels with corpus measurements, the following section addresses the issue of the rating scales used, as well as the number and experience of the raters called upon in corpus-based assessments of L2 (dis)fluency.

2.4.2.1 Rating scales, number of raters and rater experience

In L2 (dis)fluency research, **numerical rating scales** (i.e. Likert-type scales) are becoming increasingly entrenched to measure (dis)fluency. Isaacs and Thomson (2013:136) explain three advantages of using Likert-like numerical scales:

- **Versatility**: numerical scales can be used with learners from virtually any L1 background or proficiency level, on any L2, and on any type of task. Moreover, numerical scales can be used to assess the quality of production of small stretches of language (even single words or phonemes) to extended stretches of language.
- **Reliability**: even untrained raters can use numerical scales and make reliable judgements (reported Cronbach's alpha coefficients are almost always high using numerical scales).
- **No middleman**: numerical rating scales cut out the middleman of more detailed descriptor scales that may not always reflect the rater's impressions.

A major **drawback** of numerical scales, however, is that even when raters assign the same score to a speech sample, their motivation for doing so may be different: "quantitatively equivalent ratings do not preclude *qualitative* differences in raters' approach to the decision-making task or interpretation of the construct" (*ibid.*; emphasis original).

A second drawback pertains to the contentious issue of the **optimum scale length**. While nine-level scales seem to be common (e.g. Bosker *et al.* 2013; De Jong & Bosker 2013; Derwing *et al.* 2004; Derwing *et al.* 2009; Pinget *et al.* 2014; R. Rose 2015), five (Iwashita *et al.* 2008; Kormos & Dénes 2004), and ten-level (Cucchiarini, Strik & Boves 2000) numerical scales are also sometimes be used. In theory, including more scale levels should allow finer-grained distinctions in rating L2 performances, but, to make reliable judgements, raters must be able to reliably distinguish the different scale levels (Bachman 1990; Isaacs & Thomson 2013). Moreover, "[t]he number of scale categories a rater is able to distinguish is not only constrained by his or her ability to detect differences between stimuli but also the discriminability inherent in the speech samples" (Garner 1960; in Isaacs & Thomson 2013:138). For example, in Isaacs and Thomson's study (*ibid.*), 5-point scales were reportedly too constraining for some raters, but some raters also reported difficulties making meaningful nine-level distinctions. Too few, and too many, scale levels should thus preferably be ruled out for reliable scoring.

Surprisingly few studies in L2 (dis)fluency research make use of the **CEFR scales**[92] (Council of Europe 2001) for (dis)fluency assessment. Osborne (2011a) is a notable exception. In his study, he examined the extent to which different measures of fluency are reflected in raters'

[92] See Chapter 7 for a detailed overview of the CEFR scales and descriptors.

perception of oral proficiency as assessed by the CEFR. He found that it is more meaningful to relate CEFR proficiency bands to fluency measured as a bundle of features than to single measures taken in isolation. The fact that a very limited number of studies make use of the CEFR scales is probably due to the **different purposes** of the CEFR and numerical, Likert-like scales: whereas the latter are used to examine perceptions of different aspects of speech, the CEFR is more generally used in (high-stake) exams, and to make decisions about the test-taker's abilities. Another explanation for the rare use of the CEFR in (dis)fluency research might be the place of (dis)fluency in the Common European Framework: in fact, it is not prominently presented, and only considered as one of the two "qualitative factors which determine the functional success of the learner" (Council of Europe 2001:128). More generally, however, a growing number of studies has started to tackle the relationship between CEFR levels and other aspects of learner language, thereby contributing to the well-known research desideratum concerning the validity of the CEFR scales (e.g. Hulstijn 2007). For example, Wisniewski (2017) looked at the tri-dimensional relationship between the contents of the B2 level description for vocabulary control, empirical learner language, and human ratings. In Chapter 7, I will likewise investigate the relationship between the CEFR fluency scale contents, CEFR fluency ratings, and learner data.

Other types of assessment also exist such as qualitative assessments, or introspective research, but these are very peripheral in L2 (dis)fluency research, and will not be further discussed in this dissertation.

With respect to the **number of raters**, previous studies have shown that each rater tends to have a different pattern of rating (e.g. Mullen 1980). For this reason, a **minimum of two raters** are generally required in rating speech to avoid the possible effect that a single rater may have. Sometimes, a **third rater** is also called upon in case of disagreement (i.e. the "2+1 principle"; Alderson *et al*. (2001)), as was the case in Baker-Smemoe *et al*. (2014) or Ginther *et al*. (2010). In actual fact, the number of raters may range from as few as three (Préfontaine & Kormos 2016) to as many as 43 (Préfontaine 2013b), 60 (Susca & Healey 2002) or even 80 raters (Bosker *et al.* 2013).

Cumming stresses that "[p]eople assess what they believe, have learned, and value" (Cumming 2007:289). Likewise, in his book on assessing second language writing, Lumley (2005) emphasised the **centrality of raters' experience**. For example, he discusses the tendency for judges to include criteria from their own experiences rather than the scales or rubrics they are supposed to follow. Although the author's primary focus is on the assessment of writing, the issues he addresses are definitely also valid for the assessment of speech. In L2 (dis)fluency research, "**expert**" ("experienced" or "trained") and/or "**novice**" ("untrained") raters have been called upon. A good example of expert rating is Cucchiarini *et al*. (2000; 2002), who asked phoneticians and speech therapists to make expert judgements. Novice raters have been called upon in, e.g., Kormos & Dénes (2004) and Lennon (1990). For practical reasons, however, novice raters tend to be recruited more often: not only are novice raters

generally more easily accessible, but expert rating, contrarily to novice rating, is often an expensive task. Note also that language teachers are sometimes considered to belong to the expert category (Rossiter 2009), and sometimes to the novice category (Préfontaine, Kormos & Johnson 2015). In addition, it ought to be underlined that, while **native speakers** of the target language have conventionally been called upon, some studies also make use of ratings by non-native speakers of the L2 (e.g. Kormos & Dénes 2004; Rossiter 2009). These inconsistencies inevitably make **cross-study comparisons difficult**. For example, while Rossiter (2009) and Kormos and Dénes (2004) found no substantial difference between expert and novice raters in the assessment of (dis)fluency, other studies did find that experienced raters were more lenient than foreign language teachers or more novice raters (Gilquin, Bestgen & Granger 2016; Thompson 1991 [for pronunciation]).

Lastly, it should be emphasised that, despite its increasingly recognised importance, reporting **inter-rater reliability** is not yet systematic in studies assessing (dis)fluency. Although some studies do report high inter-rater reliability (Cronbach's alpha in Derwing *et al*. (2004) and in Pinget *et al*. (2014) reaches 0.95 or over), it remains uncertain whether the majority does not report such statistics because of the mathematical difficulties involved or because the inter-rater reliability was actually (worryingly) low. It is also important to bear in mind that the nature of the rated excerpt can affect inter-rater reliability: as underlined by Cucchiarini *et al*. (2000:996), read speech material might lead to higher inter-rater reliability coefficients. Inter-rater reliability being a property of a specific sample of testees performing a specific task and rated according to a specific scale by specific raters, it is thus advisable not to rely on published estimates and to measure alpha in each study to add validity to the interpretation of the data (Tavakol & Dennick 2011).

2.4.2.2 Relating perceived and utterance (dis)fluency

A number of studies have tried to set ratings of the perceived level of (dis)fluency (or proficiency) of learners in relation to corpus measurements of (dis)fluency production. As hinted to in the previous sections of this chapter, these studies considerably differ with respect to the raters (especially their number and expertise), the learners (e.g. their number and proficiency level), the rated samples (the length of the excerpts, the speaking task etc.) and the rating procedure (the rating criteria, the rating scale etc.). However, despite those important methodological differences, "there is consensus among researchers that there are clear and significant correlations of the learners' productive fluency and the native speakers' assessments of these learners' perceived fluency" (Götz 2013a:90).

Evidence is accumulating that many **temporal features** of learner speech are **correlated with perceived (dis)fluency level**. More precisely, findings in several studies (e.g. Cucchiarini, Strik & Boves 2000; Derwing *et al.* 2004; Kormos & Dénes 2004; Rossiter 2009) have been consistent in showing that **speech rate, pausing phenomena, and length of runs are primary factors correlating with (dis)fluency ratings**. For example, Préfontaine *et al*. (2015)

investigated the relationship between (dis)fluency measures and raters' perception of L2 (dis)fluency in the speech of 40 English-speaking learners of French at varying levels of proficiency. Eleven judges rated the learner performances, using two different instruments: the fluency descriptors of the CEFR *Qualitative Aspects of Spoken Language Use* scale, as well as a *Fluency perception semantic scale*. The authors reported that three of the four investigated measures were negatively correlated with raters' scores, namely the mean length of speech runs (in syllables), the articulation rate (in syllables per second of phonation time), and the frequency of pauses (of 0.25 second and above) – the first two being the most influential factors in raters' judgements. They also underlined that the relative importance of each measure in predicting (dis)fluency ratings did vary across tasks.

With respect to **repair (dis)fluency**, the literature suggests a **weak relationship** between repair phenomena and perceived (dis)fluency. Although Cucchiarini *et al*. (2002) did not find a relationship between perceived (dis)fluency and "number of disfluencies" (which included repetitions and corrections), Bosker *et al*. (2013) found that the number of repetitions and of corrections (per second of spoken time) did contribute a small, but significant, amount to perceived (dis)fluency. Likewise, Pinget *et al*. (2014) showed that measures of repair (dis)fluency could explain a small, but non-negligible, proportion of the variance in (dis)fluency ratings.

A major issue with many of the aforementioned studies is that they do not take into account **learners' (dis)fluency in their L1**. Previous research has highlighted that many (dis)fluency measures are highly correlated in learners' L1 and L2, particularly the temporal variables (Cox & Baker-Smemoe 2013; Derwing *et al.* 2009; Guz 2015; Rose 2013; 2015; Towell, Hawkins & Bazergui 1996). Fast L1 speakers are, for example, very likely to speak at a fast pace in their L2 too. As underlined by Rose (2015), if many measures of (dis)fluency are related in a speaker's L1 and L2, this leads to a "perceptual quandary". In other words, this raises the question whether perceptual differences in (dis)fluency (which are greatly affected by temporal variables, *cf*. *supra*) can accurately be correlated with developmental changes or should rather be related to individual differences in speech patterns. One possibility is to "**correct**" **L2 measures with L1 data** from the same speakers performing the same task. De Jong *et al*. (2015) have, for example, shown that articulation rate (operationalised as syllable duration) is a better predictor of L2 (dis)fluency when the measure is corrected by observations in the learners' first language. However, correcting measures is often not an option because many learner corpora simply do not include L1 data from the same speakers. In the absence of such types of data, the exploration of **(dis)fluency profiles** (see Section 1.3) also seems promising because, although they cannot account for L1-L2 (dis)similarities, they do challenge the assumption of a monolithic (dis)fluency pattern per CEFR level.

2.5 CONCLUSION

This chapter has given an overview of spoken (learner) corpora and their use in (dis)fluency research.

It has first been shown that, over the past few decades, the analysis of speech has considerably evolved: from experimental, constrained, or impromptu examples, the field steadily moved on to large databases of naturally occurring language from large populations of speakers, and in diversified communicative situations. The review of the properties of spoken learner corpora showed that the field is bubbling with life. However, a closer inspection of the corpora through (dis)fluency glasses revealed that results might not always be directly comparable: what makes the **richness of the field** of spoken learner corpora (i.e. its diversity in terms of speakers and communicative situations), indeed, also makes the **weakness** of the field of L2 (dis)fluency research. Mother tongue background, proficiency level as well as speaking task, which are core defining properties of a spoken corpus, all affect learner (and native) speech to a greater or a lesser extent and comparisons of figures across studies should, in many cases, be regarded as indicative (and not as conclusive). Failure to do so might lead the researcher to attribute differences to, for example, mother tongue background or (non-)nativeness, when they might in fact simply be due to differences in the nature of the corpora. In this respect, the field of learner (and native) (dis)fluency could certainly also gain from **meta-analyses** that aggregate the evidence from multiple studies in view of improving estimates and of analysing inconsistencies across studies.

It would benefit the field of spoken learner corpora if future data collections – and subsequent analyses – could **take L1 characteristics into consideration**. With some exceptions (e.g. Derwing *et al.* 2009; De Jong *et al.* 2015; Larsson Aas & Rørvik 2017), L2 (dis)fluency research has rarely considered the (dis)fluency of the learners in their mother tongue, but such comparative analyses invariably point to a strong relationship between the two. Although databases such as LINDSEI and LOCNEC, where learner and native speaker data have been collected following exactly the same design criteria, do not make it possible to take the learners' L1 speaking pattern into consideration, they do make it possible for researchers to carry out reliable comparative L1-L2 studies.

In a similar vein, to make possible thorough examinations across all proficiency or (dis)fluency levels, there is a strong need for more spoken corpora with data from learners belonging to the **whole range of proficiency levels**. Besides, although increasingly more learner corpora do offer some indication of the level of proficiency of the learners (such as "intermediate" or "advanced"), it would greatly benefit the field if such corpora could be rated in a more systematic manner, possibly even at the time of collection (see also Paquot & Granger 2018 in this respect). In addition, despite the obvious value of longitudinal data, there is still a dearth of studies examining the development of learner (dis)fluency over an extended period of time. Initiatives to **collect longitudinal learner data** should definitely be encouraged.

Lastly, the collection of **multimodal corpora**, i.e. with video recordings of the speaker, should also be promoted and could definitely prove to be a goldmine for future (dis)fluency research. To give but one example, verbal and non-verbal (dis)fluency could be analysed conjointly, and analyses could reveal how they interact or complement each other.

This chapter has also emphasised how the field of spoken corpora has greatly benefited from technological advances: an increasing range of **sophisticated tools** have become available for the recording, transcription, and annotation of speech so that, today, it is possible to explore large databases of recorded speech in ways that would hardly have been conceivable before. For example, **time alignment** not only increases the reliability of speech annotations, but also opens the way to more accurate analyses of temporal phenomena such as the use of pauses or speech rate. Tools such as Praat or EXMARaLDA, within which corpora can be approached in their time aligned version, also affect the way spoken data can be visualised, and reduce the biases traditional transcriptions may have. More generally speaking, the use of automated methods has the potential to open up interesting avenues for the field of (dis)fluency research.

Moreover, this chapter has offered a succinct overview of (dis)fluency research from the point of view of **language teaching and assessment** (LTA). It has been shown that corpora and LTA methods complement each other in the sense that, while corpora stimulate analyses of measures of utterance (dis)fluency (i.e. productive (dis)fluency), assessments of the (dis)fluency of learners as recorded in corpora enable investigations of listeners' perception of the learners' (dis)fluency. Several issues have, however, been raised concerning practices in (dis)fluency assessment such as rating scale length and number or experience of the raters.

First, whilst the **Common European Framework** is not the typical option for L2 (dis)fluency assessment, it does offer several advantages for the present study. Contrary to numerical scales, the CEFR offers a point of reference for cross-study comparisons of (dis)fluency ratings. While the same point on a numerical scale (say, level 4) may not correspond to the same actual (dis)fluency level (in one study, it might in fact correspond to a higher (dis)fluency level than in another), the levels of the CEFR are, in principle, more directly comparable from study to study because they are based on the same descriptors. Moreover, in spite of the criticism that could be expressed against them (see also Chapter 7), the CEFR scales and descriptors *are* used internationally and strongly advocated in a wide range of language learning settings. The conclusions drawn from studies using the CEFR could thus potentially more directly benefit language practitioners, if only by a reappraisal of the CEFR descriptors.

A second weakness of the LTA literature is the fact that rating scales (numerical scales and the CEFR scales) seem to equal the highest level of fluency with an **idealised native speaker** able to produce flawless, disfluency-free discourse. This is problematic at (at least) three levels: (1) all native speakers produce (dis)fluency features; (2) (dis)fluency features can positively affect the listener's cognitive fluency, for example by segmenting the speech flow into more easily processable chunks; (3) native speakers' (dis)fluency (like that of learners) is

affected by task factors, so that the discourse by the same speaker in one task may be more fluent than in another.

To conclude, several gaps have been identified that I would like to address in this study. I would like to contribute to the relatively modest body of literature investigating L1-L2 (dis)fluency contrastively. The CEFR descriptor scales will be used to assess learners' (dis)fluency level. Lastly, while previous studies have generally related temporal measures with CEFR levels, I will investigate the nature of the relationship of a larger range of (dis)fluency features, as well as combinations of them (in the form of (dis)fluency profiles) with CEFR levels.

In the next chapter, I will describe the methodological frame that will be applied in this dissertation.

PART II

Chapter 3
METHODOLOGY

Methodology is intuition reconstructed in tranquillity

Paul Lazarsfeld

Chapter 1 offered a broad overview of the (dis)fluency construct, the way it is defined and the concrete features that allow its empirical measurement. Chapter 2 then examined the contributions of learner and native-speaker spoken corpora to (dis)fluency research.

The present chapter provides an overview of the methodology, which is an integrated corpus-driven methodology rooted in learner corpus research (Section 3.1). The two corpora used for the analyses are introduced in Section 3.2. Following this is a presentation of the fourteen (dis)fluency variables under investigation (Section 3.3), and, finally, the statistical procedures adopted for the analyses of Chapter 5 to 7 are discussed in Section 3.4.

3.1 LEARNER CORPUS RESEARCH

As apparent from its focus and methodology, this thesis is deeply rooted in the tradition of **learner corpus research** (LCR). LCR is a research strand that emerged "in the late 1980s as an offshoot of corpus linguistics, a field which had shown great potential in investigating a wide range of native-language varieties [...] but had neglected the non-native varieties" (Granger, Gilquin & Meunier 2015b:1). As the name suggests, LCR subsumes the range of studies that aim at gaining a better understanding of the mechanisms of second and foreign language acquisition *using learner corpora* (which is not the case for SLA studies). In addition to designating a **field of research**, learner corpus research thus also refers to a **methodology**.

A. An integrated corpus-driven analysis

In learner corpus studies (as in corpus linguistic studies), corpora may be used with either a *hypothesis-based* or a *hypothesis-finding* perspective (Granger 1998), or, in other words, they may be used with a *corpus-based* or a *corpus-driven* approach. What underlies the distinction between the two is "how pre-corpus theoretical premises and intuitions should be incorporated in corpus research" (Xiao 2009:993). In a nutshell[93], while **corpus-based** studies build on pre-existing hypotheses "generated through introspection, SLA theories, or as a

[93] See e.g. Xiao (2009:993) for an in-depth comparison of the two approaches.

result of the analysis of experimental or other non-corpus-based sources of data" (Barlow, Ellis & Barkhuizen 2005:344), **corpus-driven** studies are supposedly free from theoretical premises and "evidence from the data takes precedence over theoretical constructions" (Müller 2005:27). While a corpus-driven approach is "potentially very powerful since it can help us gain totally new insights into learner language" (Granger 1998:16), it has been argued that no study can be perfectly free from initial hypotheses, and that corpus-driven studies are inevitably coloured by previous research in the investigated domain, if only when classifying concordances, or when annotating the data. Gries even claims that "truly corpus-driven work seems a myth at best" (Gries 2010:330).

In this thesis, the open-mindedness towards the data of **corpus-driven** approaches is largely adopted. At the same time, however, I acknowledge the inevitable strong theoretical grounding of this thesis in previous theories: the underlying hypotheses, the annotation of some (dis)fluency features, the interpretation of results etc. do rely on previous literature. In fact, I will try to take advantage of the respective strengths of corpus-driven and corpus-based approaches by combining the two into an "integrated corpus-driven analysis".

B. Contrastive interlanguage analysis

In 1996 and 2015, Granger proposed a new comparative framework for the analysis of learner language, the *Contrastive Interlanguage Analysis* (CIA). CIA includes two types of comparisons, namely a **comparison between learner and native language** (for example L2 and L1 English) and a **comparison between the interlanguages of two (or more) learner groups** (for example the L2 English of French-speaking and of German-speaking learners). Studies that use native speaker corpora as a benchmark for the analysis of learner corpora provide evidence for the nature of learners' interlanguage (e.g. patterns of over- and underuse). Studies that compare different learner groups can be used to highlight particular aspects of language use shared by learners "as a whole". Such analyses have typically focused on learners from different mother tongue backgrounds with a view to distinguishing L1-dependent features from those that are due to the acquisition of a language itself.

In Granger's 2015 article, some of the tenets of this two-pronged approach have been reconsidered. Specifically, in keeping with the variationist trend, CIA² promotes (among others) the notion of "varieties": reference language varieties on the one hand, and interlanguage varieties on the other.

In **L2 (dis)fluency research**, there is a growing consensus that there is a need to evaluate second and foreign language speech with respect to native (and first language) speech. Many (dis)fluency researchers have embraced this approach (e.g. Campillos Llanos & González Gómez 2014; Nivja H. De Jong 2016; De Jong *et al.* 2012b; Derwing *et al.* 2009; Götz 2013a; Guz 2015; Osborne 2011b; Tavakoli 2011), claiming that significant insights can be derived from research designs involving the comparison of learner and native speech. For example, it is essential to contrast L1 and L2 data when investigating the effects of speaking task on

(dis)fluency because "a NS baseline teases out which performance features are the result of the task and which arise from limited language resources" (Skehan, Foster & Shum 2016:110). A key requirement for L1-L2 comparative studies is obviously that the learner and the native data need to be maximally comparable.

Despite their great value, L1-L2 comparisons in learner corpus research have sometimes been argued to fall prey to the so-called ***comparative fallacy***. This notion, coined by Bley-Vroman (1983), refers to the fact that learner analyses can actually be sidetracked by a concern for the target language and that it easily becomes difficult to see learners' interlanguage as anything but deficient (Larsen-Freeman 2014). Several arguments have been brought forward to counter this criticism (see especially Granger 2009, 2015), but, as stressed by Granger (*ibid.*:14), such criticism should act as a reminder that learners' interlanguage should also be studied in its own right, and not necessarily in a strong normative perspective.

In addition to NNS-NS comparisons, CIA also includes **comparisons of different groups of learners**. While some (dis)fluency studies have indeed included learner groups from different mother tongue backgrounds in their analyses (Baker-Smemoe *et al.* 2014; Raupach 1980; Wisniewski 2015), sometimes together with L1 data (Bilá & Džambová 2011; Osborne 2011b), these remain very few. However, another type of NNS-NNS comparison seems to have gained popularity in L2 (dis)fluency research, namely the comparison of learner groups from different proficiency or perceived fluency levels. A typical example is the study by Kormos and Dénes (2004), who have compared the speech produced by eight low-intermediate and eight advanced Hungarian learners of English.

In this study, I will compare (1) learner and native speech (Chapter 5 and Chapter 6) and (2) learner groups from two different levels (Chapter 7). The comparison of L2 and L1 speech is made possible thanks to the availability of two perfectly comparable corpora of learner and native speech, namely the *Louvain International Database of English Interlanguage* (LINDSEI; Gilquin, De Cock & Granger 2010)) and the *Louvain Corpus of Native English Conversation* (LOCNEC; De Cock 2004). This comparison will enable me to tease out those aspects that are due to the act of speaking from those that are due to speaking in a foreign language. As advised by Granger (2015), I will keep the comparative fallacy warning in mind, so as not to get sidetracked in the analyses of learner language.

3.2 THE TWO SPOKEN CORPORA

The investigation of learner and native speaker (dis)fluency is based on two spoken corpora: the French component of the *Louvain International Database of Spoken English Interlanguage* (Gilquin, De Cock & Granger 2010)[94] providing the learner data, and the *Louvain Corpus of Native English Conversation* (De Cock 2004) offering a perfect British English native speaker counterpart to LINDSEI. Both corpora were developed at the Centre for English Corpus Linguistics (Université catholique de Louvain, Belgium).

3.2.1 The learner database LINDSEI

The ***Louvain International Database of Spoken English Interlanguage*** (LINDSEI) contains spoken data produced by university students of English as a foreign language from 20 mother tongue backgrounds. It was built – and is still being developed – under the direction of the Professors G. Gilquin, S. Granger, and S. De Cock. To date, eleven components have been published in the first version of LINDSEI, namely the Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish, and Swedish components. Nine more components will be included in the second version of LINDSEI, namely the Arabic, Basque, Brazilian Portuguese, Czech, Finnish, Lithuanian, Norwegian, Taiwanese, and Turkish components. Each component consists in 50 transcribed informal interviews[95] of about 15 to 20 minutes each between a learner of English and an interviewer, which corresponds to 10 to over 14 hours of recorded speech per subcorpus and over 130 hours in total in the first version of LINDSEI (Gilquin, De Cock & Granger 2010:30). Each interview is made up of three speaking tasks: a warm-up activity based on a set topic, a free discussion and a picture description task (see 3.2.4 for further details on the speaking tasks). It is important to point out that sixteen variables have also been recorded in a "profile": these include information about the interviewer, the interviewee and the setting of the interview (see Figure 3-1, from Gilquin *et al.* (2010:7)). As stressed by Gass and Selinker (2008:33) such metadata is most crucial for the accurate interpretation of the data.

---

[94] Henceforth, "LINDSEI" will refer to the LINDSEI database as a whole (http://www.uclouvain.be/en-cecl-lindsei.html), and "LINDSEI-FR" to the French component of the database. Likewise, "LINDSEI-GE" will refer to the German component of the learner corpus.

[95] Note, however, that some components do have slightly more than 50 interviews to better match with the total number of words of other components. The Chinese component includes 53 interviews [82,536 words; Gilquin, De Cock, and Granger (2010, 23)], and the Japanese 51 [56,239 words; *ibid.*]. By contrast, due to the very limited number of speakers available, the forthcoming Basque component includes only 30 interviews [284,257 words].

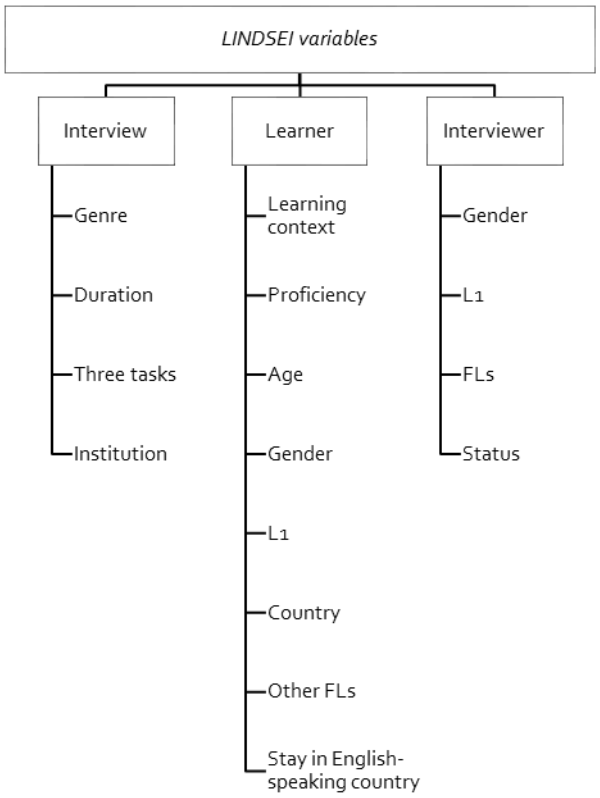


Figure 3-1: LINDSEI variables (taken from Gilquin, De Cock & granger 2010:7)

The learners in the LINDSEI database were aimed to have an **advanced level of proficiency**, which was defined based on an external criterion: the number of years they had been learning English. The learners in the database are mostly undergraduates of English in their third or fourth year at university. With a view to identifying potential differences in proficiency across sub-corpora, five random excerpts of each of the 11 components of the database were submitted to a professional rater who assessed the learners' level based on the *Common European Framework of Reference for Languages* (CEFR; Council of Europe (2001)). The results of this rating show that, while some components do qualify as advanced, the majority of the samples were rated at B2 level, i.e. an upper intermediate level. Within one and the same sub-corpus too, excerpts were sometimes rated at different levels.

On a more practical note, the transcriptions of the eleven components of the first version of LINDSEI are available via a specific interface, accessible via CD-ROM[96]. The interface makes it possible to **customise the learner corpus** on the basis of the various criteria present in the metadata. It is, for example, possible to automatically extract the transcriptions of all the Chinese speakers, or of the learners who spent more than *n* months abroad. Whilst researchers may have access to all the transcriptions and metadata from the interface, at this stage, the recordings remain the property of the national teams.

[96] The forthcoming second version of LINDSEI will, however, be web-based.

The **French component of LINDSEI** (LINDSEI-FR), which is the source of learner data in this study, was published in the first version of LINDSEI. It was collected at the Université catholique de Louvain (Belgium) between November 1995 and December 1997.

Like the other components of the database, it contains **fifty interviews** of Belgian French-speaking university students who were studying English at Master's level, aged between 20 and 33 (average: 22.09). Thirty learners are female and 20 are male. On average, at the time of recording, they had been learning English for 3.76 years at university (min. 3 and max. 6), with an additional average of 4.6 years of English at school (min. 3, max. 7). Whereas 10 learners had never spent time in an English-speaking country, 36 had (on average 1.94 month)[97].

LINDSEI-FR totals 91,402 words of learner language (1,828 learner words on average per interview)[98]. Timewise, LINDSEI-FR contains 14 hours 23 minutes and 48 seconds of recorded speech, i.e. just above 17 minutes on average per interview. As regards the proficiency level of the learners, the five samples of the French learners have been assessed at B2 level (i.e. higher intermediate) (Gilquin, De Cock & Granger 2010:11–31).

3.2.2 The native corpus LOCNEC

LINDSEI's British English native speaker counterpart, the ***Louvain Corpus of Native English Conversation***, was compiled by Prof. S. De Cock at the University of Lancaster, United Kingdom, in 1995-1996. It functions as a mirror benchmark of the non-native LINDSEI with 50 British native speakers as interviewees. Most of the informants are undergraduates in their first or second year, but some are postgraduates. The majority of the interviewees are English Language or Linguistics students, but some of them read in French, Chemistry or Management. The 30 female and 20 male native speakers are aged between 18 and 30 (average: 21.6).

LOCNEC was collected by Professor S. De Cock using the **same design criteria as in LINDSEI**: each interview in LOCNEC is also made up of three speaking tasks and is linked to a profile with the same metadata about the interviewer, the interviewee and the setting (see Figure 3-1 above).

[97] Four learners did not specify the number of months they had spent in an English-speaking country in their profile.

[98] Including the interviewer's turns: 143,887 words in total, 2,878 words on average per interview (*ibid.*). Note that the word counts include filled pauses and backchannels.

Despite their great similarity in terms of corpus collection scheme, LINDSEI-FR and LOCNEC do differ on several aspects. First, whereas it was essential for the learners to be majoring in English and to be in their third or fourth year to qualify as advanced, those criteria were not as relevant for the native speaker counterparts (see *supra*). Second, in LOCNEC, the interviewer is a non-native researcher rather than a native speaker tutor. Third, the LOCNEC transcription scheme slightly differs from that of LINDSEI (see De Cock (2003) for further details). Because I aimed at comparing the two corpora, I applied the transcription conventions of LINDSEI on LOCNEC[99]. Also, I paid particular attention to the cases of uncertainties in the transcription (also in LINDSEI-FR transcripts), which were marked either by a tag (<?> or <x[x][x]>), or in the form of an explicit comment in the LOCNEC doc file. If I was certain that I understood the speech uttered, I allowed myself to modify the transcription. If the sound was inaudible or if I was still not certain of the words that were uttered, the tag used for marking uncertainty was kept. The figures for LOCNEC that are mentioned in the rest of the thesis are based on this revised version of the transcripts.

The next subsection describes the original transcription procedure and scheme that was used for LINDSEI-FR and that I applied to LOCNEC.

3.2.3 Transcription procedure and scheme

The LINDSEI-FR and LOCNEC interviews were originally recorded on cassette tapes and were later digitised. Based on the audio recordings, the orthographic transcriptions (in .txt format) were carried out manually, first by transcribing the data proper and then by carrying out post-transcription checks. In addition, the transcriptions also include interlinear mark-up and a number of annotations that serve three major purposes:

1. the identification and delimitation of the interviews, tasks and speaker turns;
2. the transcription of typical features of speech such as filled pauses, overlapping speech or specific phonetic realisations;
3. the transcription of contextual comments such as voice quality or non-verbal vocal sounds.

The complete account of these **transcription guidelines** are available in the LINDSEI booklet (Gilquin, De Cock & Granger 2010) and in Appendix 9.1, but Table 3-1 offers a synthetic overview of the main aspects of the mark-up in LINDSEI and LOCNEC.

[99] See Appendix 9.1 and 9.4.

1. Identification/delimitation	
Speaker turns	• <A> and </A>* for the beginning and end of the interviewer's turns; • <B> and </B>* for the beginning and end of the interviewee's (the learner in LINDSEI and the native speaker in LOCNEC) turns.
2. Features of speech	
Empty pauses	• . for short pauses (< 1 second); • .. for medium pauses (1-3 seconds); • ... for long pauses (> 3 seconds). e.g. <B> (erm) .. it's a British film there aren't many of those these days </B>
Filled pauses and backchannelling	• They include: eh [brief], er, em, erm, mm, uhu and mhm; • They are transcribed between brackets*. e.g. <B> yeah . well Namur was warmer **(er)** it was **(eh)** a really little town </B>
Truncated word	• Truncated words are immediately followed by an "=" sign. e.g. <B> it still **resem=** resembled the theatre </B>
Foreign words and pronunciation	• Foreign words are indicated by <foreign> (before the word) and </foreign> (after the word)*. e.g. <B> we couldn't go with (er) knives […] **<foreign>** enfin **</foreign>** we were (er) </B> • As a rule, foreign pronunciation is not noted, except in the case where the foreign word and the English word are identical. If in this case the word is pronounced as a foreign word, this is also marked using the <foreign> tag. e.g. <B> I didn't have the (erm) . **<foreign>** distinction **</foreign>** </B>
Phonetic features	• Syllable lengthening: a colon is added at the end of a word to indicate that the last syllable is lengthened. Colons are not be inserted within words. e.g. <B> that's something I'll I'll plan **to:** to learn </B> • Articles: when pronounced as [ei], the article *a* is transcribed as "a[ei]"; e.g. <B> and it's about (erm) . life in **a[ei]** (eh) public school in America I think </B> • Articles: when pronounced as [i:], the article *the* is transcribed as "the[i:]". e.g. <B> and **the[i:]** villa we were staying in was in one of the valleys </B>
Overlapping speech	• The tag <overlap />* indicates the beginning of overlapping speech in both turns. The end of overlapping speech is not indicated. e.g. <B> yeah I went on a bus to London once and I'll never **<overlap />** do it again </B> <A> **<overlap />** that's even worse </A>
3. Contextual comments	
Voice quality	• If a particular stretch of text is said laughing or whispering for instance, this is marked by inserting <starts laughing> or <starts whispering> immediately before

	the specific stretch of speech and <stops laughing> or <stops whispering> at the end of it. e.g. <B> **<starts laughing>** I don't have to assess it I only have to write it **<stops laughing>** </B>
Non-verbal vocal sounds	• Nonverbal vocal sounds are enclosed between angle brackets. e.g. <B> I hope so I've I've got some **<coughs>** friends out there </B>

Table 3-1: LINDSEI and LOCNEC mark-up
Note: asterisks indicate that the format of the mark-up slightly differs in LOCNEC interviews as transcribed by De Cock (2003).

It is important to stress that, as can be noticed from Table 3-1, the mark-up in LINDSEI and LOCNEC is in an **xml-like format**, which has important implications for further processing, especially as regards time alignment and to a lesser extent, fluency annotation. The adoption of this type of format in the original transcriptions means that annotations occur either:

1. between angle brackets:
    - in pairs with an opening and a closing tag with the same "content", such as the tag indicating a foreign word (<foreign> </foreign>), or with different "contents" such as <starts laughing> and <stops laughing>;
    - alone, such as the tag indicating overlapping speech (<overlap />);
2. between square brackets, such as [i:] and [ei], which shows the marked pronunciation of the articles *the* or *a*;
3. between parentheses, used for filled pauses: (eh), (erm) etc.;
4. without bracketing, such as the signs indicating a truncated word (=), a vowel lengthening (:) or a silent pause (., . . and . . .).

Example 3-1 below illustrates a typical LINDSEI-FR or LOCNEC transcription after homogenisation of the transcription conventions (the mark-up is in bold type). Note, among others, the use of <A>, </A>, <B> and </B> to show speaker turns, the dots to show unfilled pauses, the filled pauses between brackets, inaudible passages (<X> or <XX>) and overlapping speech (<overlap />).

3-1: FR008-S[100]

<A> <laughs> </A>

<B> so it was about **(eh)** three years ago **(erm)** I I had the**[i:]** opportunity to go in the States **..** in fact **(em)** in Belgium I am **(em) <X>** of staff I'm very **.** I I **.** spare a lot of time of my free time for in scouting **</B>**

<A> oh right **</A>**

[100] In what follows, LINDSEI-FR and LOCNEC examples will be referred to as follows: "FR" and "EN" refer to the learner and the native corpus, respectively, the three figures (008 in example 3-1) to the number of the interview (here: the eighth interview in LINDSEI-FR), and the final capital letter (S, F or P) to the speaking task the example is taken from.

<B> and so **(er) .. every:** every month we get **(em) ..** like **a:** a newspaper for for for leaders **..** and there was **<XX>** so **.. w=** which says **..** okay if you are more than eighteen and if you can speak quite a a little bit English and you are available **(er) ..** it was six to eleven weeks during the **.** summer holiday **.** okay you can **wri=** write **(eh)** to have some **<X>** […] and **they:** they sent my **let=** to **the[i:]** other federation **<overlap />** to the to the boy scouts of America **</B>**

<A> <overlap /> oh I see **.** yes **</A>**

3.2.4 Speaking tasks

As previously stressed, LINDSEI and LOCNEC were developed according to the same design criteria so as to ensure perfect comparability between the learner and native data. The two most notable criteria are that the interviews should (1) be informal (i.e. the interviewee should not feel unduly restrained by the presence of the interviewer) and (2) elicit different types of data (e.g. monologue and dialogue, spontaneous and prepared speech, free and controlled tasks). For these reasons, each interview in LINDSEI and LOCNEC follows a particular pattern made up of three speaking tasks.

The interviewees are first presented with a list of **three topics** and are asked to select one of them. They are given some time to think about what they want to say (but they cannot take written notes) before starting to talk. The three topics are:

1. *An experience you have had which has taught you an important lesson. You should describe the experience and say what you have learned from it.*
2. *A country you have visited which has impressed you. Describe your visit and say why you found the country particularly impressive.*
3. *A film/play you have seen which you thought was particularly good/bad. Describe the film/play and say why you thought it was good/bad.*

Although this activity is not entirely spontaneous and slightly constrained topic-wise, the interviewees can organise and formulate their speech freely. It is aimed to be more or less monologic, but the interviewer may intervene at times and ask a question or two to help the interviewee produce more speech. This activity is sometimes referred to as the "warm-up activity" because it is also aimed to help the interviewee feel at ease with the situation.

After the first task, the interviewer goes on with a number of questions about various subjects such as life at university, hobbies or travels. The aim of those questions is to stimulate a natural, **spontaneous dialogue** between the interviewer and the interviewee.

In the last task, the interviewee is presented with a four-picture cartoon (Figure 3-2) and is asked to **describe the story** that is pictured (he/she may have some time to understand the

story). He/she can use the words of his/her choice, i.e. language is not constrained, but the structure of the talk and the content are pre-defined.

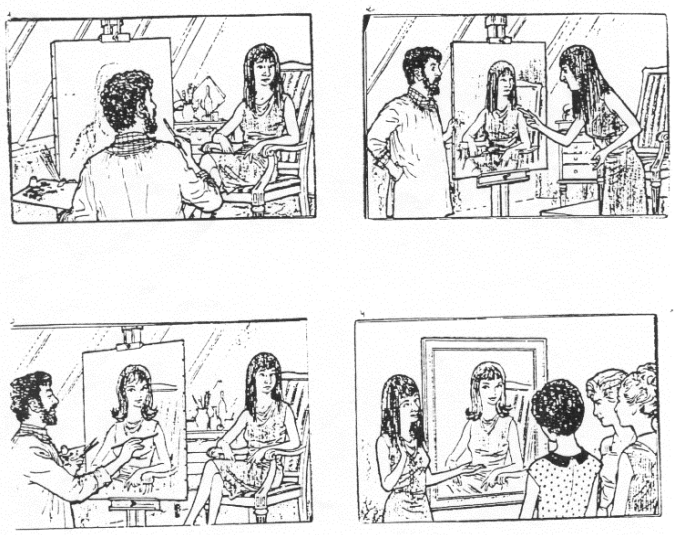


Figure 3-2: The picture description task

3.2.5 Corpus size and duration

The general figures of the two corpora are summarised in Table 3-2, namely the **number of tokens** (including filled pauses and backchannels, but excluding all elements between angle brackets) per corpus and per interview (A+B turns and B turns only each time) as well as the **corresponding durations**. The data for LINDSEI-FR are taken from Gilquin *et al.* (2010) and the data for LOCNEC from the revised version of the corpus.

The number of tokens, both for A+B and for B turns only, is slightly higher in LOCNEC (176,381 and 129,780, respectively) than in LINDSEI-FR (143,887 and 91,402). The average length of the interviews, however, is slightly longer in the learner corpus (c. 17 minutes) than in the native LOCNEC (c. 15.5 minutes) which amounts to more than 14 hours of recorded speech in LINDSEI-FR and just over 13 hours in the native corpus.

Corpus	No. of tokens (A & B turns)	No. of tokens (B turns only)	Total duration (A & B turns)
LINDSEI-FR (Mean per interview)	143,887 (2,878)	91,402 (1,828)	14h 23m 48s (c. 17m)
LOCNEC (Mean per interview)	176,381 (3,527)	129,780 (2,595)	13h 02m 04s[101] (c. 15,5m)
Totals	***320,268***	***221,182***	***27h 25m 52s***

Table 3-2: Tokens and durations in LINDSEI-FR and LOCNEC

Table 3-3 and Table 3-4 reveal the **breakdown of the number of tokens per task** in LINDSEI-FR and LOCNEC. At this stage, it is not possible to measure the total duration per task because the interview is recorded as one audio file (and not one audio file per task).

	Set topic	Free discussion	Picture description	Total
A & B turns	65,061 (45.2%)	69,374 (48.2%)	9,452 (6.6%)	**143,887 (100%)**
B turns only	46,076 (50.4%)	38,836 (42.5%)	6,490 (7.1%)	**91,402 (100%)**

Table 3-3: Number of tokens per task in LINDSEI-FR

	Set topic	Free discussion	Picture description	Total
A & B turns	58,697 (33.3%)	107,883 (61.2%)	9,801 (5.5%)	**176,381 (100%)**
B turns only	46,741 (36%)	75,345 (58.1%)	7,694 (5.9%)	**129,780 (100%)**

Table 3-4: Number of tokens per task in LOCNEC

The above tables show that the picture description task elicits the least speech from the interviewees (7.1 % in LIDNSEI-FR and 5.9 % in LOCNEC) as compared to the other two tasks. But, whereas in the learner corpus, it is the set topic that appears to trigger slightly more speech than the free discussion (50% vs 42%), the most productive task in terms of number of tokens produced by the native speaker interviewees is the free discussion: about 58% of the native speakers' speech is produced in the second task and "only" 36% in the set topic. Looking more closely at the figures for the corpora as a whole (i.e. A & B turns), it appears that the set topic and the free discussion account for nearly the same proportion in LINDSEI-

[101] The total duration was computed by adding up the lengths of the 50 audio files.

FR (45% vs. 48%) whereas a larger discrepancy can be found for the native speakers (33% vs. 61%).

Speaking task is well-known for its great influence on language production (see Section 2.2). Thanks to this 3-tier pattern in the corpora, it becomes possible to analyse learner and native-speaker language at a micro-level by analysing each task separately or contrastively, or at more macro-level by considering the three tasks as a whole.

3.2.6 The metadata

3.2.6.1 Situational variables

This study into learner and native (dis)fluency is inscribed within the frame of a large-scale project entitled "Fluency and disfluency markers. A multimodal contrastive perspective". As presented in the Introduction, this Concerted Action Research project (ARC) groups a large team of researchers who investigate fluency and disfluency markers in two languages (French and English) and four modalities (spoken and sign language; native and learner language). One of the initiatives that was undertaken to **ensure the comparability of the results** between the languages and modalities involved in the project addresses the issue of the standardisation and categorisation of "situational features". The PhD theses involve a dozen corpora and an impressive panel of communicative situations ranging from political interviews, sermons or class debates to humorous sketches, sports commentaries, or scientific presentations at conferences. Given this impressive variety, data comparisons and generalisations become particularly hazardous without a cross-genre specification of the variables involved.

Text type specifications abound in the literature and many have attempted to identify the major variables that could distinguish one text type from another. In the second language testing literature more particularly, special attention has been paid to the definition of tasks in terms of lists of characteristics such as the channel (aural, visual), the language (native and/or learned), the number, roles or statuses of the participants, the topic or the degree of interaction (Wright 1987:49; Pica, Kanagy & Falodun 1993; Weir 1993:39; Bachman & Palmer 1996:49; Lewkowicz 1997; Lewkowicz 2000) (see Fulcher 2003 for a review). In reality, many systems overlap for some of the core features or are specific to second language testing settings and can thus not easily be extended to L1 settings or to the large panel of situations in our ARC (dis)fluency project.

Drawing on previous tentative categorisations, a collaborative study (Crible *et al.* 2014) was thus carried out in order to establish a **parametric standardisation of the major variables** involved in our spoken corpus situations. Six core "situational variables" were highlighted, namely the degree of elicitation, the number of interlocutors, the degree of preparedness,

the degree of interactivity, the degree of media coverage and the media-orientation. Each situational variable is gradual (each has three levels, with the exception of "professional aspect" which only has two) and each level is precisely glossed (see Appendix 9.2).

The **degree of elicitation** is defined as the presence and weight of the experimental protocol as a constraint on the interaction. It can be *supervised*, *semi-supervised* or *natural* depending on whether speech production is dictated by the input, guided (with a reasonable degree of flexibility in how the speaker expresses him-/herself) or dependent upon the speakers themselves.

The question of the number of speakers who are actively taking part in the interaction is addressed in the second variable: the **number of interlocutors**. It can be a *monologue*, a *dialogue* or a *multilogue*.

The variable **degree of preparedness** seeks to evaluate the extent of (spoken and/or written) preparation of the main speaker's discourse. It can be considered as *prepared* when both content and form have been carefully pre-planned, *semi-prepared* when only the general frame of the discourse has been thought of but the precise wording is spontaneous, or *spontaneous* when the discourse has not been prepared at all.

The **degree of interactivity** refers to the speakers' ability to adapt their speaking behaviour to the other interlocutor(s) with respect to what is expected from their status in the interaction. If the situation allows all the speakers to speak and hold the floor, it is *interactive*; if one speaker holds the floor more than the others, it is labelled *semi-interactive* and if one speaker keeps the floor nearly continuously, it is *non-interactive*.

The **degree of media coverage** is defined as the extent of broadcasting as the main aim of the interaction. It can be *media-oriented* such as a TV show or a radio interview; *semi media-oriented* (e.g. a sermon which is also broadcast on TV for example) or *not media-oriented* when the interaction is not broadcast, such as a class at school.

The **professional aspect** refers to whether or not the situation is due to one of the speakers' professional activity. It can be *professional* or *non-professional*. An impromptu telephone conversation between friends would, for instance, be considered as belonging to the non-professional category, whereas an interview between a journalist and a celebrity would be classified as professional because the communicative situation was set up in the framework of the journalist's job.

Following these six situational variables, the three speaking tasks in LINDSEI-FR and LOCNEC, which are perfectly similar in nature, are characterized as follows (Table 3-5):

Task	Elicitation	Interlocutors	Preparedness	Interactivity	Media coverage	Professional aspect
Set topic	semi-supervised	dialogue	semi-prepared	semi-interactive	not media-oriented	professional
Free discussion	semi-supervised	dialogue	spontaneous	interactive	not media-oriented	professional
Picture description	supervised	monologue	spontaneous	non-interactive	not media-oriented	professional

Table 3-5: Situational features of the 3 tasks in LINDSEI-FR & LOCNEC

As regards the **first situational variable**, the picture description lies at the end of the elicitation continuum: the interviewee has to speak about a precise topic (i.e. the story shown in the pictures), using specific vocabulary and following the same discourse structure (the plot of the story). It is thus considered as a "supervised" task. The set topic and the free discussion are both qualified as "semi-supervised" because the interviewee's speech is less constrained by the task design: for the set topic, the interviewee has to speak about one of the three topics, and in the discussion, the interviewer asks questions on a restricted number of topics such as university life or plans for the future. In each case, the vocabulary, grammar, structure and length of the discourse is totally up to the interviewee.

Secondly, a similar grouping seems to apply for the **second variable**. The third task is clearly deemed monologic while the second (the free discussion) is dialogic because the floor is more or less equally distributed between the two speakers. The classification is however less straightforward for the first task. Although it was aimed to be on the monologic side (though admittedly to a lesser extent than the picture description), in practice, the set topic task in the French component of LINDSEI and in LOCNEC is more on the dialogic end: typically, the learner or native speaker begins talking about his/her chosen topic, but the interviewer soon and regularly steps in for clarifications or questions to stimulate the exchange, which inevitably renders the task dialogic. This is indeed supported by a formal analysis of the transcriptions: the repeated and quick alternation of speaker turns in the set topic indicates that the two speakers are actively taking part in the interaction, though it is true that the floor belongs slightly more to speaker B than to the interviewer[102]. In addition, without a close look at the very content of the speech, it is most generally impossible to spot the boundary between the first two tasks. For these reasons, the set topic is classified as a dialogue.

As far as **preparedness** is concerned, the interviewee is given some time to prepare for the first task (but he/she could not take notes), which is not the case for the other two tasks that are spontaneous in nature (although learners do spend a few seconds to understand the story in the cartoon, they do not actively prepare their speech).

[102] The equality (or lack of it) in sharing the floor is taken into account in the situational feature "degree of interaction".

The three degrees of **interaction** are illustrated in the corpus: whereas the picture description is non-interactive (the interviewee holds the floor nearly exclusively), the free discussion is interactive (the floor is more or less equally distributed) and the set topic has a medium position (the interviewee has the upper hand).

Lastly, the variables **"media coverage"** and **"professional aspect"** are identical for the three tasks because they reflect the way the corpora are designed: they are not aimed to be broadcast ("not media-oriented") but emerged from the professional activity of the interviewer, i.e. research ("professional").

It is hoped that this typology of situational variables will prove useful in comparing speaking communications and that similarities (or unexpected differences) between situations that may not look similar at first sight but are characterized similarly according to this system will be discovered in later analyses. For example, in Crible *et al.* (2017), we used the typology of situational variables to identify similar communicative situations in four corpora of learner English, L1 English, L1 French and Belgian French sign language, and to compare some (dis)fluency features across these four modalities.

3.2.6.2 Homogenisation of the metadata

A second initiative within the ARC project aimed at homogenising the metadata of all corpora involved in the project using similar codes. The codes and labels (available in Appendix 9.3) used are not discussed here as they do not affect the present study but will be useful for future comparisons between the results of the four theses involved in the project.

3.2.7 Potential and limitations of LINDSEI-FR and LOCNEC

Compared to previous spoken corpora, LINDSEI-FR and LOCNEC undoubtedly have many major advantages that are of prime importance for investigations of learner and native spoken language.

Firstly, the two corpora contain data on 50 learners and 50 native speakers. A hundred speakers may arguably not be a lot compared to written corpora, but it is rare for spoken corpora to contain as many speakers. Although generalisations of the results might be hazardous, it is important to stress that twice fifty speakers is sufficient for **reliable statistical analyses** (see also Section 3.4). Secondly, whereas many corpora contain (very) little data from each speaker (often only a couple of minutes per speaker), the interviews in LINDSEI-FR and LOCNEC are 15 to 20-minute long, which is arguably more **representative** of a speaker's speaking competence. In LINDSEI-FR and LOCNEC, the speakers have sufficient time to develop their speech and show more than one facet of their speaking competence. Besides, the fact that the interviews are made up of **three speaking tasks** is also an

undeniable plus. These tasks offer a good panel of more or less spontaneous, more or less prepared, and more or less interactive speech and, despite the fact that (semi-)interactive and less constrained tasks are acknowledged to be more difficult to analyse, I believe that they open up research perspectives into how (dis)fluency is shaped by authentic speaking situations. Fourthly, not only does the learner corpus come together with a native counterpart, but the two corpora were also collected following exactly the **same design criteria**: this makes them perfectly comparable in terms of speaker variables as well as in terms of speaking tasks. Lastly, as the learner corpus is actually a subcomponent of a larger database containing data from learners from various mother tongue backgrounds, possibilities of comparing the (dis)fluency of French-speaking learners with that of **other learner populations** (esp. the German component as used in Götz (2013), but also the Czech, Swedish or Taiwanese components whose coordinators also work on (dis)fluency-related issues) can be envisaged.

These plus points notwithstanding, the **authenticity** of the data might still be questioned. The speaking tasks were chosen to offer a wide panel of speaking situations but they may arguably not be the most representative of the speakers' (especially the learners') actual use of English. It might also be argued that, despite the fact that the interviewees knew they were being recorded, the unfamiliar environment and situation might have influenced the way they spoke. The speakers might for example have been more careful in the way they formulated their utterances, or have paid particular attention to avoiding blanks and hesitations that could have been negatively interpreted as a lack of competence or self-confidence for example. However, when listening to the contents of the interviews, one soon notices that many speakers talk about (very) personal details or casually joke with the interviewer, which could be interpreted as them being at ease and having forgotten they were being recorded.

The **familiarity with the speaking tasks** may also not have been equal for the two speaker groups. Whereas describing cartoon pictures is a task learners could have got more familiar with in foreign language classrooms, it may not be as typical an activity for native speakers, and this could in turn affect the way they speak. Monologic tasks such as the set topic are also likely to be more familiar to learners than to native speakers, who are potentially more used to casual conversations. Whereas there are reasons to think that there might be a slight bias in the speakers' familiarity with the three speaking tasks, it is partly counterbalanced by the fact that (1) the learners were university students who had already learned English for many years and had presumably used this language to communicate both in more monologic and in more dialogic situations and that (2) the majority of the native speakers knew at least one foreign language and had thus also gotten the opportunity to become more familiar with typical learner speaking tasks.

Additionally, researchers always face the issue of availability of material for their research, and it has to be acknowledged that the **date of collection** of the learner corpus constitutes a

limitation of the data at two levels. LINDSEI-FR and LOCNEC were collected between 1995 and 1997, and, at the time when LINDSEI-FR was being collected (i.e. between 1995 and 1997), the focus of foreign-language classes used to be on **accuracy** (i.e. form) more than on fluency. Outside school and university, pupils and students had **little exposure to English**: newspapers, radio shows and TV programmes used to be either in French or dubbed, and the internet was obviously not yet so widespread in Belgian homes. English books, magazines or newspapers were not easily available either. For LOCNEC too, it might be that some linguistic changes have taken place since the date of collection (e.g. Leech *et al.* 2009; Leech 2010).

The situation has changed at several levels since then. Exposure to English outside foreign language classrooms has increased drastically, especially through the use of the **Internet**: it is now easy to have access to online English-speaking newspapers, blogs or websites, podcasts and films in their original language. Online games that include (generally, but not exclusively) English chat rooms have also become popular. Besides, Belgian citizens regularly travel to foreign countries and more students embark on an **Erasmus** stay (though not always in an English-speaking country) or take a gap year before going to university.

More importantly perhaps, major structural changes have been implemented in Belgian secondary education shortly after the collection of the LINDSEI-FR data, with considerable implications for foreign language classes. The notion of "skill" (FR. *compétence*), which had become increasingly popular in the 1990s, was officially decreed by law in 1997, and new learning programmes came into force shortly afterwards. Practically, this means that each school subject is now taught and learned in an action-oriented perspective. For foreign language classes, this implies that grammatical aspects, exercises, or questions during exams, for example, must be situated in a meaningful context for the learner.

It is out of the scope of this thesis to discuss the details of these reforms (all the more so because, to my knowledge, their concrete impact on foreign language skills has not been comprehensively and empirically evaluated), but, as the **learner data are anterior to these reforms**[103], they could admittedly not ideally reflect current French-speaking learners' knowledge. Besides, over 20 years, **exposure to English** (and especially to spontaneous English through the Internet and television) has skyrocketed in Belgium, which, for some learners, is likely to have had a (most likely positive) impact. Bearing in mind that the learners were university students majoring in English in their third or fourth year, it could be argued that the influence of these two factors is hopefully limited by the fact that the language curricula at university have been less affected by the aforementioned reforms. Consequently, the most important part if the learners' training, that is, the part where they *specialise* in English, has remained largely identical since the date of collection of the corpus. These two

[103] The French learner data was collected between November 1995 and December 1997.

elements should however still be borne in mind when interpreting the results in the next chapters and when drawing conclusions.

As regards the **format** of the corpora, several important limitations for precise investigations of (dis)fluency ought to be pointed out.

The two corpora can be classified as **spoken corpora**: they contain written transcriptions of spoken data. The national teams possess the original recordings and use them if the audio file is needed to disambiguate a specific case, but the **use of the audio recordings** (e.g. to listen to a specific passage) proves to be very time-consuming when no systematic temporal links are made between the two files. Generally, analysing a spoken corpus actually means analysing a "collection of transcriptions" (Leech, Myers & Thomas 1995:6). Also, the corpora do not contain **visual data**, which renders impossible the study of gestures, for example, though they also contribute to (dis)fluency.

The lack of systematic links between the audio recordings and the transcriptions also adds complexity to (or sometimes makes impossible) the investigation of some core spoken features such as the study of pronunciation or prosody. **Temporal phenomena** in particular, which are of prime importance for the study of (dis)fluency, cannot be reliably measured: calculation of speech rate or mean length of runs, for example, can only be approximate. The analysis of **unfilled pauses** is also problematic on account of the low degree of objectivity and precision in their detection. Silent pauses in LINDSEI-FR and LOCNEC were transcribed perceptively using one, two, or three dots depending on their length. There are three issues with this method of transcribing silent pauses (see e.g. Arlington *et al.* 1992; Duez 1985; Edwards 1992; Megyesi & Gustafson-Capkova 2002; Pye, Wilcox & Siren 1988).

- At the time of transcription of LINDSEI-FR and LOCNEC, not many tools existed that could help visualise the audio file with a spectrogram, for example. This could have helped in the identification of silent pauses: transcribers had to rely solely on their hearing abilities. However careful they may have been, transcribers may not have perceived all the unfilled pauses in speech: some of them could have escaped the ears of the transcribers because of their "natural" location in the utterance or because there was already much to transcribe and mark up in the surrounding context. It is thus likely that some unfilled pauses slipped through the net.

- When they were perceived, silent pauses were transcribed in three different ways depending on their length: a short pause of less than 1 second was transcribed with one dot; two dots were used for medium pauses (between 1 and 3 seconds) and three dots for long pauses of more than 3 seconds. Again, this transcription is based on the

perception of the length of pauses. No empirical measurements were made[104]. It would be more accurate to say that the pauses were transcribed on what was *perceived* to be less than a second, 1 to 3 seconds, or more. Studies (see Section 2.3) have however shown that the perception of the length of pauses depends on many elements such as the location of the pause in the utterance or the speech rate. Consequently, it is to be expected that the classification into short, medium and long pauses (the lengths of which, by the way, are arbitrary) may not be very sound.

- Lastly, inter-turn pauses were transcribed only exceptionally (for example when they were very long, as in the picture description, when the interviewee discovers the pictures and tries to make sense of them). Inter-turn pauses might however reflect the speed of the exchange between the two speakers and can greatly affect temporal measurements depending on whether they are taken into account or not.

The transcription guidelines include inline annotations of typical spoken phenomena and these are undeniably a very valuable asset: it is, for example, possible to investigate filled pauses automatically, as well as truncated words. However, the **mark-up was not intended for an analysis of the speakers' (dis)fluency** but was aimed to capture generic properties of spoken language so as to increase the reusability of the data. Some more specific (dis)fluency-related phenomena are thus not annotated, such as restarts, false starts, or repetitions. Besides, the spatial display of the transcriptions (i.e. a vertical transcription, *cf*. Section 2.3) is of critical importance, but also tricky, as it also influences the perception of the data. Consider for example the two following different ways of transcribing the same excerpt. Example 3-2 is the preliminary transcription of an interview from the Turkish component of LINDSEI. Example 3-3 is the revised transcription of the same excerpt. In the preliminary version of the transcription, it seems as if the interaction is very limited. This inappropriate spatial display corresponds to what Edwards (1992) refers to as format-based bias. It can easily be adjusted by reformatting. Example 3-3 is a suggestion of such reformatting. Note the quick alternation of A and B turns which gives off a very different impression than 3-2.

3-2: TR019 - transcriber version

```
<B> my teacher is there <overlap /> and ... my (eh) in my primary school (eh) I (eh)
my teacher .. call me and say she (eh) she's .. she (eh) is . giving cour= course for
<overlap /> the <overlap /> other students she say me to help me I will be happy
<overlap /> maybe if I (eh) return to my hometown I will <overlap /> help her </B>

 <A> <overlap /> (mhm) </A>

 <A> <overlap /> (mhm) </A>
```

[104] Note, however, that some of the national teams, especially those responsible for the new components of LINDSEI, may have used tools for the visualisation or the detection of unfilled pauses. For LINDSEI-FR and LOCNEC, this was not the case.

```
<A> <overlap /> (mhm) </A>

<A> <overlap /> (mhm) </A>

<A> <overlap /> (mhm) </A>
```

3-3: TR019 - adaptation of the transcriber version

```
<B> my teacher is there <overlap /> and </B>

<A> <overlap /> (mhm) </A>

<B> my (eh) in my primary school (eh) I (eh) my teacher .. call me and say she (eh)
she's .. she (eh) is . giving cour= course for <overlap /> the </B>

<A> <overlap /> (mhm) </A>

<B> <overlap /> other students </B>

<A> <overlap /> (mhm) </A>

<B> she say me to help me I will be happy <overlap /> maybe </B>

<A> <overlap /> (mhm) </A>

<B> if I (eh) return to my hometown I will <overlap /> help her </B>

<A> <overlap /> (mhm) </A>
```

As we have seen, LINDSEI-FR and LOCNEC have many advantages, but they also come with some limitations. While some of them will have to be borne in mind at later stages of the analysis, it is possible to deal with the lack of time alignment and the limitations in terms of (dis)fluency annotations. The methodology adopted for overcoming these two aspects is discussed in Chapter 4.

The next section provides an overview of the variables under investigation in this study.

3.3 THE VARIABLES

3.3.1 The productive (dis)fluency variables

Prior to any analysis, a principled decision must be made concerning the variables to be used. As presented in Chapter 1 and Chapter 2, previous literature was surveyed to identify a set of **fourteen (dis)fluency variables**. These are displayed in Table 3-6.

Ten features from Table 3-6 (i to x) were directly annotated in the time aligned version of LINDSEI-FR and of LOCNEC with the annotation tool EXMARaLDA (Schmidt & Wörner 2014) (see Chapter 4 for the annotation scheme and procedure) and frequency counts were subsequently extracted. Four temporal (dis)fluency measures (xi to xiv) were calculated automatically based on the time aligned corpus data. The table also provides the **formula and unit for the measurement of each variable**. As can be seen, variables i to x are normalised **per hundred words** (phw) of interviewee speech. Normalisation is crucial because the length (in terms of words) of each speaker's speech can vary widely.

Frequency counts per minute are a popular alternative to frequency counts per hundred words, perhaps more particularly in the L1 literature. It is, however, yet unclear which implications the choice of the former versus the latter may have on research findings, and no study has extensively compared, on the same dataset, the impact that adopting frequency counts per hundred words or per minute has on subsequent research findings[105]. Moreover, it appears that what is exactly meant by "minute" and "word" is itself liable to some degree of variability: what is a spoken word is not always precisely defined (what about contractions for example?), and the handling of unfilled pauses in time counts also tends to greatly differ (is it per minute of phonation time?).

In the present study, I chose the measure per hundred words mainly for reasons of **comparability**, as it appears that frequency counts in the English L2 literature tend to be more often normalised per hundred words than per minute. The number of words of interviewee speech used to measure the frequency of (dis)fluency features is based on the number of words resulting from the segmentation process (which was a prerequisite for the time alignment of the corpora). A word is thus here defined as a letter or a sequence of letters

[105] In Dumont (2017a), I tried to make a first step towards addressing this methodological issue. I compared the use of a few (dis)fluency variables measured both per hundred words and per minute, and showed that there are, indeed, clear implications in terms of research findings. This small-scale attempt obviously needs to be complemented by further analyses to better understand the consequences of favouring one measure, and, perhaps, to advise the researcher for, or against, using one or the other.

surrounded by blank spaces[106]. The word count is **unpruned,** i.e. it includes all the words uttered by the interviewees, including repetitions, reformulations etc., as well as filled pauses. Only (inter-turn and intra-turn) **unfilled pauses** as well as **nonverbal sounds** such as laughter or coughing were **excluded** from the interviewees' word counts.

	(Dis)fluency measure	**Definition/formula**	**Unit**
i.	Unfilled pauses [UP (phw)]	Number of intra-turn unfilled pauses divided by the total number of words*100	phw
ii.	Filled pauses [FP (phw)]	Number of filled pauses divided by the total number of words*100	phw
iii.	Repetitions [Rep (phw)]	Number of repetitions divided by the total number of words*100	phw
iv.	Restarts [RS (phw)]	Number of restarts divided by the total number of words*100	phw
v.	False starts [FS (phw)]	Number of false starts divided by the total number of words*100	phw
vi.	Truncations [T (phw)]	Number of truncations divided by the total number of words*100	phw
vii.	Vowel lengthenings [L (phw)]	Number of vowel lengthenings divided by the total number of words*100	phw
viii.	Discourse markers [DM (phw)]	Number of discourse markers divided by the total number of words*100	phw
ix.	Conjunctions [C (phw)]	Number of conjunctions divided by the total number of words*100	phw
x.	Foreign words [W (phw)]	Number of foreign words divided by the total number of words*100	phw
xi.	Mean length of unfilled pauses [Mean UP length (s)]	Total length of intra-turn unfilled pauses divided by the number of intra-turn unfilled pauses	sec.
xii.	Speech rate [SR (wpm)]	Total number of words divided by the total speaking time (including pausing time)*60	in words per minute

[106] Illustrations and more details about the segmentation process, including a more precise definition of "word", can be found in Section 4.1. Note that contractions are counted as two words.

xiii.	Mean length of runs [MLR]	Mean number of words between unfilled pauses, calculated by dividing the total number of words by the number of runs (a run is defined as a segment of speech that occurs between two UPs)	in words
xiv.	Phonation time ratio [PTR]	Speaking time without inter-turn UP time divided by speaking time including inter-turn UP time*100	%

Table 3-6: The (dis)fluency measures used in the analyses of Chapter 5, 6 and 7
Notes: (1) phw = per hundred words; (2) sec. = second

For **speech rate**, I first considered using the number of syllables per minute as this measure has been claimed to be more accurate (see e.g. Griffiths 1991). I experimented with some open-source software such as Praat scripts (e.g. De Jong & Wempe 2009; 2007; Easy Align from Goldman 2011), SPPAS (Bigi 2015), as well as online tools such as Syllable Counter[107] (as in Kahng 2014), but these proved either highly inaccurate or particularly demanding in terms of time or technical knowledge. Given the size of the corpora and my untrustworthy knowledge of phonetics and syllabification rules in spoken English, I excluded the option of manually segmenting and counting syllables, and, following *inter alia* Lennon (1990), Götz (2013a), and Gráf (2015), I decided to measure speech rate in **words per minute**. Incidentally, the difference in terms of accuracy between speech rate in syllables per minute and speech rate in words per minute might not be as great as has sometimes been claimed: in his PhD thesis, Gráf (2015) compared the precision of the count of words vs. of syllables per minute and calculated the ratio of syllables per word in English speech. He showed that the average length of a word is 1.29 syllables, and that the standard deviation (.05) is "very low", which indicates that, in spoken English, "the differences in word length produced by individual speakers is negligible" (Gráf 2015:92–93).

To calculate the **mean length of runs**, I first needed to obtain the number of runs in each speech sample. A run is here defined as a word or a sequence of words that occurs between two unfilled pauses (Götz 2013a; Grosjean 1972; Tavakoli 2016). Note that the interviewer speech may also mark the end of a run. Illustrations of speech runs are provided in Figure 3-3 and Figure 3-4 (*cf.* tier 3 – "FR045 [runs]").The number of runs was obtained semi-automatically by manually marking the beginning and end of each run in each file within EXMARaLDA, and then automatically exporting the exact resulting number.

[107] http://www.syllablecount.com (last accessed 10/01/2017).

	[(	137 [00:39	138 [(	139 [00:4(	140 [00:4(	141 [(	142 [00:	143 [00:41.4]	144 [00:42.(	145 [(	146 [(	147 [00:42	148 [00:43	149 [00:43	150 [00(	151 [(	152 [(	153 [(	154 [(	155 [00:45.1]	156 [00:4(	157 [00:46	158 [00:4(	159 [00:47	160 [(	161 [00
A [TR]															yeah										
FR045 [TR]		(0.490)	I	choose	to	to	study	economics	because	it	was	(0.280)	general	(0.420)	you	had	a	lot	of	mathematics	and	(0.710)	and	history	and	more
FR045 [runs]		*(0.490) I choose to to study economics because it was*										*(0.280) general*		*(0.420) you had a lot of mathematics and*								*(0.710) and history and more general courses*				
FR045 [anno-1]		<UPL>			<L>+<R0	R1>						<UPL>		<UPA>							<C>+<R0	<UPL>	<C>+R1>			
FR045 [anno-2]																						<N>				
FR045 [anno-3]																										
A [POS]															RB											
FR045 [POS]		<pause>	PP	VVP	TO	TO	VV	NNS	IN	PP	VBD	<pause>	JJ	<pause>	PP	VHD	DT	NN	IN	NNS	CC	<pause>	CC	NN	CC	JJR

Figure 3-3: Segmentation into speech runs (FR045-S)

	0	25 [00:05.0]	26 [0(	27 [00:06.(	28 [0(	29 [0(	30 [0(	31 [0(	32 [00:(	33 [00:07.8	34 [00:	35 [00:10.:	36 [0(	37 [00	38 [0(	39 [00:12.:	40 [00:12.6	41 [00:13.:	42 [0(	43 [00:14.:	44 [00:1	45 [0(	46 [0(	47 [00:15.4	48 [0(	49 [00:16.8
A [TR]		pictures		(0.250)	and	tell	me	the	story	(1.950)											mhm					
FR045 [TR]			yes	(0.250)						(1.950)	so	(0.890)	a	man	is	(0.270)	painting	(0.760)	a	woman		and	er	(0.750)	the	woman
FR045 [runs]			*yes (0.250)*							*(1.950) so*		*(0.890) a man is*				*(0.270) painting*		*(0.760) a woman*				*and er*		*(0.750) the woman*		
FR045 [anno-1]											<DM>	<UPL>				<UPL>		<UPL>				<C>	<FP>	<UPA>		
FR045 [anno-2]																										
FR045 [anno-3]				<UPA>						<UPL>																
A [POS]		NNS		<pause>	CC	VV	PP	DT	NN	<pause>											NP					
FR045 [POS]			UH	<pause>						<pause>	IN	<pause>	DT	NN	VBZ	<pause>	VVG	<pause>	DT	NN		CC	VVG	<pause>	DT	NN

Figure 3-4: Segmentation into speech runs (FR045-P)

3.3.2 The CEFR fluency ratings

As previously mentioned, a second aim of the present study is to examine the nature of the relationship between quantitative data on (dis)fluency features (*cf.* above) and CEFR fluency ratings.

Three native and professionally-trained raters were asked to assess the French learners from LINDSEI-FR according to the **CEFR scales and descriptors**. Five subskills were assessed, namely accuracy, range, fluency, phonological control and coherence, as well as the learners' general speaking proficiency. The CEFR descriptors cover **five bands**, ranging from A2 (basic user), B1 & B2 (intermediate), and C1 & C2 (advanced), but the raters could further distinguish sublevels by using + or - increments (such as "C1-", indicating a weaker performance within the C1 band).

The rating was based on a c. **5-minute excerpt from the free discussion task**, and the raters were not provided with the corresponding written transcripts so as not to influence them. No training session was organised prior to the rating. The raters were provided with the CEFR descriptors for spoken skills, worked independently, and had **no contact with each other** before or during the rating.

The raters' **grades for fluency** (i.e. "B2","C1", "C2") were converted into numerical values and a **mean was calculated** to obtain a final CEFR fluency score per learner. Two variants of the CEFR measure for fluency are used in Chapter 7, depending on the type of analysis: the CEFR fluency grade (or level), and the CEFR fluency score (see Table 3-7). The former is a categorical variable, and the latter is a numerical (continuous) variable.

More details on the assessment procedure, inter-rater reliability and the calculation of the CEFR fluency score are provided in Chapter 7.

	Measure	**Definition**	**Example**
i.	CEFR fluency grade	The mean CEFR grade provided for fluency by the 3 raters	B1, B2, C1, C2
ii.	CEFR fluency score	The mean CEFR score provided for fluency by the 3 raters	Numerical scale from 1 to 10

Table 3-7: The (dis)fluency measures used in the analyses of Chapter 7

3.4 STATISTICAL PROCEDURES

The ultimate goal of this study is to determine which aspects of (dis)fluency characterise intermediate to advanced French-speaking learners of English, as compared to British native-speakers of English. This objective involves **three main steps**. First, I set out to identify which (dis)fluency variables distinguish L1 from L2 speakers. Secondly, I adopt a multivariate perspective and seek to identify the relationships between (dis)fluency measures with a view to underlining possible underlying dimensions of L1 and L2 (dis)fluency as well as "(dis)fluency profiles". Finally, I zoom in on the learner data, and relate the L2 (dis)fluency measures with the learners' assessed CEFR fluency level.

Throughout this thesis, the **fourteen (dis)fluency variables** displayed in Table 3-6 (Section 3.3.1) are investigated. These are obtained from the time aligned and annotated version of the **50 LINDSEI-FR and the 50 LOCNEC interviews.**

Before using statistical tests, it is crucial to confirm that the data matches the assumptions of the statistical tests I plan on using in terms of level of measurement (categorical, ordinal or ratio) and distribution. These aspects are examined in Sections 3.4.1 and 3.4.2, respectively. Then, Section 3.4.3 sets out to provide an overview of the statistical tests used in Chapter 5, 6 and 7.

3.4.1 Screening the variables

As explained by Howell (2013:4), variables are the "property of an object or event that can take on different values". Variables can be either *dependent* or *independent*. Dependent variables (DV) correspond to the phenomena under investigation and are sometimes also referred to as *outcome variables.* Independent variables (IV) are thought to affect the DV and are also sometimes referred to as *predictor variables* (Field 2013:7–8).

In our data, independent (or predictor) variables include "Nativeness" (is the speaker a NNS or a NS?), the speakers' identification code, and the CEFR fluency grades and scores. The **fourteen (dis)fluency variables** are dependent (or outcome) variables.

3.4.2 Levels of measurement

The level of measurement of a variable refers to "the relationship between what is being measured and the numbers that represent what is being measured" (Field 2013:8). Variables can be categorical or continuous.

A **categorical variable** assigns the same code to same entities: when two things are equivalent, they are given the same name or number. From a statistical point of view, gender is traditionally considered a categorical variable, and it is binary, because there are only two categories to choose from – male or female. Another example of a binary variable is coin toss (the two categories are heads or tails). When the (two or more) categories are ordered in a meaningful way (though the differences or intervals between the categories may be unequal), the categorical variable becomes ordinal. Another level of measurement includes **continuous variables**. One type of continuous variable is an interval variable, where the intervals between individual points on the scale are equal. The other type of continuous variable is called the ratio variable. Ratio variables go a step further than interval data "by requiring that in addition to the measurement scale meeting the requirements of an interval variable, the ratios of values along the scale should be meaningful" (Field 2013:10).

For the present study, the following variables qualify as categorical variables: **nativeness** (learner or native speaker, i.e. a binary variable); **speaker identification code** (100 different codes, one for each of the 50 learners and 50 native speakers); **CEFR fluency grades** (B2 or C1, i.e. an ordinal variable); **CEFR fluency score** (from 1 to 10, ordinal variable), and the fourteen (dis)fluency variables enter the analyses as continuous ratio variables.

3.4.3 Choosing statistical tests

To choose statistical tests, researchers first need to establish that their data meet the assumptions of the tests they plan on using.

First, each (dis)fluency variable in the learner and the native corpus was visually represented through boxplots to spot **outliers** (Field 2013:163–212). Where outliers were identified, the data was checked to determine whether they possibly resulted from encoding errors (such as very long pauses due to recording issues), or whether they corresponded to real outliers. In the former case, they were deleted from the data, but not in the latter.

One of the major assumptions relates to the **normality of the data**. Parametric statistical tests such as *t*-tests (used to compare two means) or ANOVAs (i.e. analysis of variance, used to compare several means) assume that the data is normally distributed within each condition. Following Howell (2013) and Field's (2013) recommendations, I used the following methods to investigate the normality of the data: the Shapiro-Wilk and Kolmogorov-Smirnov normality tests (*cf.* Appendix 9.6) and the visual representation of each (dis)fluency variable (per corpus) through histograms, P-P plots and Q-Q plots (quantile-quartile plots).

The Shapiro-Wilk and the Kolmogorov-Smirnov tests indicate that normality should not be assumed for – generally one of the conditions of – some (dis)fluency variables. For example, while normality should not be assumed for filled pauses in LOCNEC, the same variable

perfectly meets the assumption of normality in LINDSEI-FR. After a visual inspection of the data, it appeared that departures from normality were not substantive. In addition, according to Howell (2013:658–659), proponents of parametric tests argue that "the assumptions normally cited as being required of parametric tests are overly restrictive in practice and that the parametric tests are remarkably unaffected by violations of distribution assumptions" (see also Rietveld, Hout & Ernestus 2004:360).

Besides, the central limit theorem states that a sampling distribution of the means approaches the normal distribution as the sample gets larger. It is generally suggested that a number of observations of 25 to 30 is sufficient to produce a normal sampling distribution (Field 2013:169–172; Howell 2013:178–181). In both LINDSEI-FR and LOCNEC, there are 50 observations per variable, which implies that the concerns for normality should not be too serious and that the data do not substantially depart from normality.

Lastly, although it is increasingly acknowledged that the assumption of normal distribution is invalid unless sufficiently large corpora are used (*cf*. the central limit theorem), non-parametric tests have only been used exceptionally in L2 (dis)fluency research. Kormos and Dénes (2004) are a notable exception, but they analysed the speech samples of 16 learners, which is lower than the recommended number of observations by the central limit theorem for assumptions of normal sampling distribution.

While acknowledging that there is a dire need for more comparisons of parametric and non-parametric statistical tests on the same corpus data, for the reasons stated above, **parametric tests** will be used for the statistical analyses.

To analyse the data, several different statistical analyses were used, as described in the following sections. For those statistical analyses, I mainly used IBM Statistics **SPSS 23.0** (IBM Corp. 2013), but also made sporadic use of **R**[108] for some advanced functions.

3.4.3.1 Descriptive statistics (Chapters 5, 6 and 7)

To provide a descriptive overview of the data, means, standard deviations, boxplots and scatterplots are presented.

In order to determine whether there is a difference between two group means (learners vs. native speakers, or B2 vs. C1 learners), *t*-tests for independent samples were used. Effect sizes (Cohen's *d*) were calculated online using the online calculator from the Social Science Statistics[109]. For multiple comparisons, ANOVAs were used, using a Bonferroni procedure to adjust the level of significance.

[108] Available at: https://www.r-project.org/ (last accessed 12/01/2018).

[109] http://www.socscistatistics.com/effectsize/Default3.aspx (last accessed 18/01/2018).

Pearson's *r* is used to measure the strength of the relationship between two continuous variables. To better visualise the correlations, the data are plotted in scatterplots.

3.4.3.2 The relationship between (dis)fluency variables (Chapter 6)

In view of uncovering the latent structure underlying learner and native (dis)fluency features, a **Principal Components Analysis** (PCA) was run on the **NNS data** and on the **NS data separately**.

PCA is used to **analyze interrelationships** among a large number of variables and to summarise these correlations in a concise fashion by building a small set of common **underlying dimensions** (called "components" or "factors") that are "conceptually clearer than the many linguistic measures considered individually" (Biber 1988:64). Components are linear combinations of interrelated variables derived from a correlation matrix and they empirically summarise the correlations among the original observed variables (Biber 1988; Crawley 2007; Field 2013; Hair *et al.* 1995; Loewen & Gonulal 2015; Tabachnick & Fidell 1989; Walker 2013)

There are two major types of factor analyses: **exploratory** and **confirmatory**. The former is appropriate in "searching for structure among a set of variables" by "not set[ting] any a priori constraints on the estimation of components or the number of components to be extracted" (Hair *et al.* 1995:367). By contrast, if researchers have hypotheses on the structure of the data (e.g. based on theoretical grounds), they may use confirmatory factor analysis to "assess the degree to which the data meet the expected structure of the analyst" (*ibidem*).

The key steps in conducting a principal components analysis include:

- **Selecting the variables to be included in the analysis**. In this study, the fourteen (dis)fluency variables are included in the original analysis, for learners and for native speakers independently.
- **Determining the appropriateness of PCA:**
    - Screening the correlation matrix between the variables: a visual inspection of the matrix should reveal a substantial number of correlations greater than .30. Also, any variable that does not correlate with any other variable should be eliminated, and the analysis re-run (Field 2013:685–686).
    - Examining the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and the Bartlett test of sphericity: the KMO statistic varies between 0 and 1 and can be calculated for the set of variables and for individual variables. It should reach a bare minimum of 0.5 (Field 2013:684). Barlett's measure tests for the presence of correlations among the observed variables and should be significant (i.e. $p < .05$).

- **Determining the method of extracting factors**: Principal Components Analysis (PCA) or common Factor Analysis (FA). The difference between PCA and FA is in the variance that is analysed: in FA, only *shared variance* is analysed. In PCA, *all the variance* in the observed variables is analysed. The factors in PCA are called "components". In the study, Principal Components Analysis was selected.

- **Determining the number of components/factors**: although it is possible to obtain as many components/factors as there are variables, not all of them are retained and only those with large "eigenvalues" (a measure of explained variance) are kept. Complementarily to a scree plot (which is used to visualise the inflexion point of the eigenvalues of each factor), Kaiser's criterion may be adopted and only components with eigenvalues greater than 1 retained.

- **Rotating the factors to increase interpretability**: there are two classes of rotation. The first is orthogonal rotation (if the components are assumed to be uncorrelated), and the second is oblique rotation (the components are allowed to correlate). There are different methods for each type of rotation: varimax is the most frequently used method for orthogonal rotation, and direct oblimin for oblique rotation. Field (2013) recommends the use of varimax "because it is a good general approach that simplifies the interpretation of factors" (2013:644).

- **Examining the factor loadings**: factor loadings indicate the strength of the association between each independent variable and each component. Ideally, each variable should have a high loading on only one factor, and small loadings on the remaining factors. A variable with low loadings on all factors indicates that the variable is not strongly associated with any of the other variables (and should perhaps be excluded from the analysis).

- **Calculating factor scores**: factor (or component) scores are "a composite score for each individual on a particular factor" (Field 2013:673), based on his/her scores for the constituent variables (i.e. the variables that load highly on the factor or component). There are several sophisticated techniques to calculate factor scores (such as the regression technique, the Bartlett method, and the Anderson-Rubin method – see DiStefano *et al.* (2009) or Uluman and Doğan (2016) for a discussion) but they basically utilise the original measurements (i.e. the values obtained for the (dis)fluency variables) and the factor analytic results to calculate a new score. Hair *et al.* (1995:390) explain that "an individual who scores high on the several variables that have heavy loadings for a factor will surely obtain a high factor score on that factor. The factor score, therefore, shows that an individual possesses a particular characteristic represented by that factor to a high degree".

The fourteen (dis)fluency variables (*cf.* Table 3-6) were integrated in the original PCA. Oblique rotation was used (with direct oblimin) because there was no sound reason to assume a priori

that components do not correlate. Factor scores were calculated with the Anderson-Rubin method (see Field 2013:673).

Given that "we should consider whether the techniques provide different solutions to the same problem" (Field 2013:676), I also ran a **common Factor Analysis** to compare the results. The two solutions **differed only minimally,** which is in line with Guadagnoli and Velicer's (1988:266) conclusion that "the solutions generated from principal components analysis differ little from those derived from factor-analytic techniques". In practice, because the PCA is conceptually less complex and more commonly used in variation analysis (Walker 2013:455), only the results from the Principal Components Analysis are presented.

3.4.3.3 Profile analysis (Chapter 6)

The objective of the Principal Components Analysis is to analyse a large set of *variables* to identify a small number of potential underlying dimensions of (dis)fluency. The analysis of (dis)fluency profiles takes a slightly different perspective by examining a large number of *respondents* (i.e. speakers) and combining them into a small number of **distinctly different and mutually exclusive groups based on their similarities**.

Cluster analysis, which is a multivariate exploratory procedure, is used to group cases (here speakers) that are very similar with regard to some variables and is useful when there is extensive variation among the individual cases. Groups resulting from cluster analysis are devised "based on a distance measure between the respondents' scores on the variables being analysed" (Hair *et al.* 1995:372). In other words, based on their shared similarities across (dis)fluency variables, cluster analysis classifies the speakers into groups with minimized inner variance as compared to the total variance in the sample. Cluster analysis is also "particularly relevant where there is evidence to suggest that different subgroups of learners may utilise different pathways to language learning, including different strategies, aptitudes, motivational profiles, or different linguistic features to produce successful spoken or written language" (Staples & Biber 2015:244).

The key steps in conducting a cluster analysis include:

- **Selecting and transforming the variables to be included in the analysis**. Staples and Biber (2015:253–254) advise to use standardised variables (e.g. *Z*-scores) if these use different scales and/or have different ranges as this affects the outcome of the clustering (i.e. variables with larger values contribute more to the distance measure than variables with smaller values).

- **Selecting a clustering algorithm**. There are **two main types of cluster analysis,** generally depending on whether the researcher has decided on the best number of clusters before the analysis or not. If the number of clusters has been determined in advance, *disjoint (or non-hierarchical) clustering* is usually recommended. If not,

hierarchical clustering is advised. Hierarchical cluster analysis proceeds through several iterations: at the beginning stage, each speaker constitutes a cluster; the two closest speakers are then combined into a new aggregate cluster. Eventually, in the final iteration, the last two clusters are aggregated into a single cluster. This procedure is also known as the agglomerative hierarchical cluster analysis (Baayen 2008:148–180; Crawley 2007:738–744; Gries 2013:336–349; Hair *et al.* 1995:437–438; Manning & Schütze 2000:497–514; Staples & Biber 2015:246). When the clustering process proceeds in the opposite direction, that is, the splitting off of one large cluster, the cluster analysis is said to be *divisive*, but, in fact, "divisive methods act almost as agglomerative methods in reverse" (Hair *et al.* 1995:438).

There are several agglomeration methods (or "rules"), such as the nearest or furthest neighbour, but "[b]ased on a review of the literature, Ward's method is the most commonly used measure within HCA [Hierarchical Cluster Analysis]" (Staples & Biber 2015:252; see also Aldenderfer & Blashfield 1984 for more technical aspects of this method).

- **Selecting a similarity measure**. The choice of the measure of similarity that estimates the distance between pairs of individuals and is then used to develop clusters of closely similar speakers is of central importance in cluster analyses. It can be measured with either correlational measures, distance measures and association measures (for nonmetric data). Distance measures represent similarity as the proximity of observations to one another across the variables. Several options are available, but the most commonly used is the Euclidean distance (see Hair *et al.* 1995:432 for more details). According to Staples and Biber (2015:253), the **squared Euclidean distance** should be used with Ward's method. The squared Euclidean distance is chosen because it is advised with Ward's method.

- **Plotting the clusters**. The hierarchical tree structure can be visualised with a dendrogram plot.

- **Determining the number of clusters**. A major issue with hierarchical cluster analysis is how to select the **final number of clusters** that best represents the number of groups in the data. Although there are many guidelines to approach this issue, there is no standard and objective procedure. The examination of the **dendrogram** plot is a first indicator. A useful, more qualitative, criterion is the **relative distance between the clusters** formed at each step: the larger the distance, the less similarity there is between cases that have been clustered together. In other words, large distances indicate that two dissimilar clusters are combined. It is thus "common to select as the optimal solution the number of clusters that one finds just before a large jump in the relative distance coefficient (or fusion coefficient)" (Jarvis *et al.* 2003:385). Based on the examination of the distance coefficients, it is also often advised to compute a

number of different cluster solutions (e.g. 3, 4 and 5-cluster solutions) and select the best option among these after manual examination.

- **Interpreting the composition of each cluster**. This interpretation is based on the comparison between the mean values of the (dis)fluency variables per cluster using one-way ANOVAs.

The fourteen (dis)fluency variables are included in the cluster analysis, for learners and for native speakers independently[110]. Because there is no reason to assume a priori that there is an optimal number of clusters in the data, **agglomerative hierarchical algorithm with Ward's method** is used (Staples & Biber 2015:252; see also Götz 2013a). The dendrogram plot as well as the fusion coefficients (in the agglomeration schedule in the output) are used to determine the final number of clusters to be retained.

3.4.3.4 Predicting CEFR fluency scores (Chapter 7)

A **multiple linear regression analysis** was used to determine the extent to which L2 (dis)fluency variables may predict CEFR fluency ratings.

Multiple regression analysis (MRA) refers to a family of statistical methods involving the prediction of an outcome variable from several predictor variables (i.e. individual independent variables and their interactions). To put it another way, MRA is a means to explain variance in the outcome variable as a function of one or more predictor variables (Hee Jeon 2015:131). Depending on the purpose of the study and the nature of the variables under investigation, different MRA can be chosen. With continuous predictor variables, as is the case here, linear regression is used.

The key steps in conducting a multiple linear regression analysis include:

- **Checking for multicollinearity**. The assumption of multicollinearity can first be checked by running bivariate correlations on all predictor variables. *R* values equal to, or higher than, .90 or -.90 indicate multicollinearity. Together with these *r* values, the Tolerance statistic or the variance inflation factor (VIF) should also be examined. As a rule of thumb, a Tolerance statistic lower than .40 and a VIF higher than 2.50 indicate multicollinearity. If multicollinearity is detected, it is advised to eliminate the most intercorrelated variable(s) from the analysis, unless there is strong theoretical motivation for not doing so (Hee Jeon 2015:137–140).

[110] Components indentified through a Principal Components Analysis may also be used in Cluster Analysis intead of the primary observed variables. While this may solve some issues, it also leads to several other problems, as underlined by Dolnicar & Grün (2008). In this thesis, the balance between the pros and cons tipped in favour of using the observed variables (i.e. not the components).

- **Ensuring linear relationship** by examining bivariate scatterplots of variables and residual plots.
- **Choosing a feature selection algorithm.** There are three main approaches to select the optimal subset of predictor variables (Field 2013:322–323; Gries 2013:260):
    - the backward selection, which starts with a maximal model containing all predictors, and predictors that do not contribute (or contribute very little) to the model are successively discarded;
    - the forward selection, which starts with a minimal model and successively adds predictors until no addition of a predictor improves the model (or when all available predictors are already in the model);
    - the bidirectional selection, which is a combination of the backward and forward selection.
- **Checking for potential influencial cases**. Three statistics may be used to gage the influence of individual cases on the model: Cook's distance, leverage, and Mahalanobis distances (*cf.* Field 2013:306–307; Hee Jeon 2015:137).

The fourteen learner (dis)fluency variables as well as all their interactions are used in a stepwise multiple regression with forward selection.

3.5 CONCLUSION

This chapter discussed the methodology adopted for this thesis. The five distinctive methodological characteristics of the present study are:

- an integrated corpus-driven approach to the data;
- a CIA design including a comparison between learner and native language as well as a comparison between two learner groups of different CEFR fluency levels;
- a wide panel of (dis)fluency measures extracted from time aligned, (dis)fluency annotated corpora
- a strong statistical basis;
- the use of CEFR ratings of learner speech.

A few precautionary words are in order with respect to the statistical tests described in 3.4.3.2 through 3.4.3.4.

It is very important to keep in mind that **different methodological choices can, and do, impact on the test results**. The choice of the variables to include in (or exclude from) the analysis, their level of measurement, whether they are standardised or not, the rotation method, the measure similarity etc. all impact on the subsequent findings to a lesser or greater extent.

Besides, although many guidelines exist to guide the researcher through statistical meanders, there is always **some level of subjectivity** involved and the process relies heavily on the interpretation of the researcher. To give but one example, the choice of the optimal number of clusters in the Cluster Analysis is definitely not entirely objective as it partly relies on the subjective examination and appreciation of different cluster solutions.

As with any study, **replication studies** (Porte 2012) are very much needed to corroborate and refine the results and conclusions drawn in the following chapters.

Prior to presenting the findings proper, Chapter 4 turns to two technical manipulations, namely the time alignement and the (dis)fluency annotation of LINDSEI-FR and LOCNEC. Those two manipulations are pre-requisites to obtain frequency counts of ten (dis)fluency features and reliable measurements of four temporal (dis)fluency measures.

Chapter 4
ALIGNING AND ANNOTATING LINDSEI-FR AND LOCNEC

William coughed politely. "Er... hm..." he said. This is what he did when he wanted to introduce a new subject. He managed to do it gracefully because it was his habit — and I believe this is typical of the men of his country — to begin every remark with long preliminary moans, as if starting the exposition of a completed thought cost him a great mental effort. Whereas, I am now convinced, the more groans he uttered before his declaration, the surer he was of the soundness of the proposition he was expressing.

The Name of the Rose (Eco 1984:145)

As established in Chapter 2, two techniques can be used to explore and maximise the potential of spoken corpora, namely time alignment and linguistic annotation. These prove to be crucially important in the frame of (dis)fluency research, not only to broaden the scope of the investigated features, but also to objectify their identification and measurement as much as possible.

This chapter is divided into two main parts. It firstly offers a complete account of the transformations operated on LINDSEI-FR and LOCNEC to undergo time alignment (Section 4.1). Then, in Section 4.2, the (dis)fluency annotation scheme and procedure is presented.

4.1 TIME ALIGNMENT

As defined in Chapter 2, time alignment refers to the synchronisation, or **mapping, between text and sound** in spoken corpora. It is a process by which virtual anchors (also called timestamps) are inserted in the transcription files to mark the temporal beginning and end of predefined units of talk. In other words, alignment makes each unit in the transcription match with its acoustic equivalent in the recording. This procedure transforms a "mute" spoken corpus into a "speaking" spoken corpus (Ballier & Martin 2015).

From a technical point of view, the procedure aims to align the LINDSEI-FR and LOCNEC written transcriptions with their corresponding audio recordings in an output format that is **compatible with the EXMARaLDA software** (Schmidt 2001), a system that is designed to

create, annotate, manage and analyse spoken corpora. The alignment procedure was also designed to **preserve the original (inline) LINDSEI-FR and LOCNEC mark-up** and to convert it as **stand-off annotations**.

After a summary of the main aspects of the original format of the corpora (Section 4.1.1), the alignment procedure is described in four successive sections: the pre-alignment manipulations (Section 4.1.2), the segmentation phase (Section 4.1.3), the phonetic transcription phase (Section 4.1.4) and the alignment proper (Section 4.1.5). Then, the post-alignment checks are presented in Section 4.1.6. In the last two sections, the limitations of the alignment procedure are briefly discussed (Section 4.1.7) and, finally, the precise durations of the corpora (as a whole and per task) are provided (Section 4.1.8).

4.1.1 Initial format

LINDSEI-FR and LOCNEC consist of two types of files: the recordings proper in .wav format and the orthographic transcriptions (which contain some mark-up inspired from the XML format) in .txt files. In addition to the corpus proper, LINDSEI-FR and LOCNEC also have a corresponding database in Access format where all the metadata is stored (see Figure 3-1 for the list of variables).

Each recording/transcription pair is identified by a **code** consisting of two capital letters followed by a three-digit number. The **two letters** identify the mother tongue of the interviewee (FR for French in this case, but also DU for Dutch or GE for German etc. in the other components of LINDSEI; and EN for English in the native LOCNEC) and the three-digit number corresponds to the **interview number** (aka 001 to 050). FR003, for example, refers to the third interview ("003") in the French component of LINDSEI ("FR") and EN040 to the fortieth interview ("040") of a native English speaker in LOCNEC ("EN"). Note that in what follows, I will use these codes to identify corpus examples, with an additional "-S", "-F" or "-P" to specify the task (set topic, free discussion or picture description, respectively) when relevant.

As mentioned earlier, the transcriptions in both corpora also involve some **mark-up** in pseudo-XML format. A set of letters identifies:

- speaker turns: "A" for the interviewer's turn, and "B" for the interviewee's;
- speaking tasks: "S" for set topic, "F" for free discussion and "P" for picture description.

More specific tags represent typical phenomena (pauses as dots, interruptions of words as equal signs etc.) – a summary of the phenomena and their corresponding marking up has been presented in Table 3-1 in Chapter 3.

4.1.2 Preliminary checks

Before aligning the corpora, a number of technical manipulations proved to be necessary. A full account of the checks and modifications can be found in Appendix 9.4, but the following explains the four most important ones.

A. Cleaning the transcriptions

The transcription files were converted to standard XML files and **small slips** in the transcriptions were corrected, such as missing (or double) angle brackets or a missing end of turn tag. Moreover, in order not to significantly increase the number of tags, similar phenomena were homogenised. For example, <makes a noise>, <makes the sound>, <makes noise> and <tapping noise> were grouped under <noise>; unclear words and unclear word endings (<X> and <?>) were grouped under <unknown/>.

B. The timing of audio files

The major constraint of the alignment tool is that it requires that the length of each recording be roughly 5 minutes (300 seconds). Each recording, however, lasts for about 15 to 20 minutes. The recordings were split into three sub-files, each corresponding to one of the three tasks (i.e. set topic, free discussion, or picture description). This also has the advantage of enabling distinctions between speaking tasks in further analyses. If a task was still far longer than 5 minutes (typically the free discussion task), it was further divided into sub-parts of about 5 minutes and a specific set of tags was added in the transcription files, representing the new cut-off points (e.g. <F1> and </F1>, <F2> and </F2> etc. for the first, second etc. 5-minute section of task F).

Practically, the precise temporal boundaries of the beginning and end of each task were measured using the Audacity software and written down in an Excel file in the format hh:mm:ss.mmm, i.e. in hours, minutes, seconds and milliseconds.

Note that in the final output, the different sub-parts of a given task (e.g. the two sub-parts of task F) are glued back in order to have three aligned files per interview, i.e. one per speaking task.

C. One transcription file per task

In parallel to the timing of audio files, each transcription file was split into three sub-files so as to have one transcription file per task. These were named simply by adding the letter S, F or P after the interview ID. FR003_F thus corresponds to the free task from the third interview in the French component of LINDSEI and EN049_P to the transcription of the picture description in the 49th interview of LOCNEC.

D. Overlapping speech

The last pre-alignment manipulation pertains to overlapping speech. When the speech of the two speakers overlaps, the tag `<overlap />` is used to show the beginning of the overlapping speech in both speakers' turns (following the LINDSEI transcription guidelines, the end of the overlap is not indicated). The tags should thus **always come in pairs**, one in each speaker's turn. In Example 4-1, the word *after* uttered by B overlaps with the beginning of *look have you got any sort of e= examples* as uttered by the interviewer and *cos that's quite interesting* (also uttered by A) overlaps with *oh . I well so there were only boys there for example* from speaker B.

4-1: FR008-S

```
<B> […] and then <overlap /> after<?> </B>

<A> <overlap /> look have you got any sort of e= examples as <overlap /> cos that's
quite interesting </A>

<B> <overlap /> oh . I well so there were only boys there for example […] </B>
```

Despite the guidelines provided in LINDSEI's booklet (Gilquin, De Cock & Granger 2010), two types of problems occurred with these overlapping tags.

While the tags were supposed to always occur in pairs, some did occur without their twin in the other speaker's turn. It is however an essential prerequisite for the alignment tool to have both: if there is only one overlapping tag, the transcription is either not aligned correctly, or the alignment of the file stops altogether. I thus checked all the tags one by one and corrected the transcriptions where there was an uneven number of `<overlap />` tags by listening to the audio file. Example 4-2 illustrates typical problematic cases.

4-2: FR004-S

```
<B> (er) when I was sixteen . my mum just dropped me at the station in Ostend and </B>

<A> <overlap /> and you went </A>

<B> off I yeah . and (erm) I kn= didn't know anybody cos (er) I w= I arrived in
Nottingham cos it was Nottingham . I arrived at Nottingham station I I had to call
the people . who and they they . they picked me up at the station and (er) I didn't
know anybody I was like .. <overlap /> quite strange </B>

<A> <overlap /> what so this had been arranged […] </A>
```

The first overlap in A's turn is problematic because it does not have a matching tag in the preceding B turn or at the beginning of the next turn. When listening to the recording, I could make sure that there was indeed an overlap, and noted that *and you went* overlapped with *off I yeah*. Consequently, the transcription was corrected as follows (Example 4-3):

4-3: FR004-S - corrected (part 1)

```
<B> (er) when I was sixteen . my mum just dropped me at the station in Ostend and </B>
```

<A> **<overlap />** and you went </A>

<B> **<overlap />** off I yeah . and (erm) I kn= didn't know anybody […] </B>

The second overlap in 4-2 (*<overlap /> quite strange </B>*) nicely matches the second overlap in A (*<A> <overlap /> what so this had been arranged […]*).

Example 4-4 below illustrates the second problem with the transcription of overlapping speech.

While, according to the transcription guidelines, the **end of the overlapping speech** is not marked by a specific tag, it appeared that the transcriber sometimes felt the need to indicate it, either by using an overlap tag and/or an unfilled pause. In the third turn of Example 4-4, two overlaps occur very close to each other in A's turn (*<overlap /> and that <overlap />*) while there is only one in the following B turn. A close listening to the recording revealed that there is only one overlap (A's *and that* overlaps with B's *yeah (erm)*) and that the transcriber used the second tag in A's turn to show where A's speech did not overlap any more with B.

4-4: FR014-S

<A> but (er) . no I think it was good fun and did you go to the Planetarium as well </A>

<B> no **<overlap />** no (mm) . (mm) </B>

<A> **<overlap />** or you didn't you just went to Madame Tussauds and what about (er) did you visit any of the gardens and **<overlap />** and that **<overlap />** sort of thing </A>

<B> **<overlap />** yeah (erm) Hyde Park </B>

<A> yeah </A>

A more accurate transcription that respects the transcription conventions while also showing where the overlapping speech ends in both speakers' turns is shown in Example 4-5.

4-5: FR014-S - corrected

<A> but (er) . no I think it was good fun and did you go to the Planetarium as well </A>

<B> no **<overlap />** no (mm) . (mm) </B>

<A> **<overlap />** or you didn't you just went to Madame Tussauds and what about (er) did you visit any of the gardens and **<overlap />** and that </A>

<B> **<overlap />** yeah (erm) </B>

<A> sort of thing </A>

<B> Hyde Park </B>

<A> yeah </A>

In similar cases where an <overlap /> tag was used to indicate the end of the overlapping speech, the extra tag was removed manually (but I always came back to the audio file in case of hesitation).

Alternatively, **unfilled pauses** were also sometimes (mis)used to **indicate the end of the overlapping speech**, as can be seen in Example 4-6, where B's overlapping speech in the third turn is only *oh* and the unfilled pause actually signals that the speaker resumes speaking alone, after A has finished uttering *cos that's quite interesting*. The corrected transcription is shown in 4-7.

4-6: FRoo8-S

<B> […] and then **<overlap />** after<?> </B>

<A> **<overlap />** look have you got any sort of e= examples as **<overlap />** cos that's quite interesting </A>

<B> **<overlap />** oh . I well so there were only boys there for example […] </B>

4-7: FRoo8-S - corrected

<B> […] and then **<overlap />** after<?> </B>

<A> **<overlap />** look have you got any sort of e= examples as **<overlap />** cos </A>

<B> **<overlap />** oh </B>

<A> that's quite interesting </A>

<B> I well so there were only boys there for example […] </B>

Example 4-8 below illustrates the various ways the transcriber used the overlap tags both correctly and erroneously in consecutive turns of the same transcription. The first two overlaps are correctly transcribed (see above) but, shortly after, two overlap tags occur very close to each other in A's turn (*<overlap /> the <overlap />*) while there is only one in B's following turn. There is actually only one overlap (A's the overlaps with B's *yeah yeah*) and the transcriber (1) used the second tag in A's turn to show where A's speech did not overlap any more with B's, and (2) indicated the end of B's overlapping speech by inserting an unfilled pause. The transcription was corrected as shown in 4-9.

4-8: FRoo4-S

<B> […] and (erm) I kn= didn't know anybody cos (er) I w= I arrived in Nottingham cos it was Nottingham . I arrived at Nottingham station I I had to call the people . who and they they . they picked me up at the station and (er) I didn't know anybody I was like .. **<overlap />** quite strange </B>

<A> **<overlap />** what so this had been arranged with **<overlap />** the **<overlap />** family but </A>

<B> **<overlap />** yeah yeah .. and (er) I didn't know who I was waiting for I was just standing at .. (er) at the[i:] entrance of the station and . but it was quite all

right cos (em) in the beginning I was supposed to look after the children (erm) .. and then . they they would say they would go to parties and said oh well c= come on with us we'll we'll we'll (er) take a baby-sitter so I was supposed th= I was supposed to: to be to look after the children but . you know they would take **<overlap />** somebody else </B>

4-9: FR004-S - corrected (part 2)

<A> […] what so this had been arranged with **<overlap />** the </A>

<B> **<overlap />** yeah yeah </B>

<A> family but </A>

<B> and (er) I didn't know […] </B>

The cases where the transcriber used unfilled pauses to indicate the end of overlapping speech actually proved to be more complex than I first thought because unfilled pauses might also indicate a short pause in the speaker's speech while A and B are both speaking at the same time (such as B's overlap in Example 4-6). After several tests, however, it proved to be far more efficient to include a script in the alignment tool which interprets unfilled pauses within a span of 3 words after an overlap tag as the actual end of overlapping speech and to manually correct the cases where such pauses did indicate a pause within overlapping speech in the post-alignment checks.

Last but not least, on a more technical note, the interviews were recorded in **monophonic** sound reproduction (and not stereophonic sound, where the two voices are kept separate). This has practical consequences for the time alignment of overlapping speech: with monophonic sound, the alignment tool can only **time align the words uttered by one of the two speakers during an overlap**, the other speaker's speech being not accurately aligned with the recording. With stereophonic sound, it would have been possible to time align both speakers' speech at the same time. Several possibilities were thus considered. Time alignment of overlapping speech could be systematically based on the learner's speech (i.e. the interviewer's speech would not be aligned accurately during all overlaps). It could also be based on the interviewer's speech, as interviewers regularly "have the upper hand" during an overlap. Alternatively, the time alignment of overlaps could be based on the speaker who is "dominant" during the overlap (i.e. who has the longest turn). Better results proved to be obtained using this last option.

4.1.3 Word segmentation[111]

As explained previously, various units can be adopted for the time alignment: the audio recordings and the transcriptions can be aligned at the level of the turn, the word, the phoneme etc. depending on the level of precision needed with regard to the research objectives. Bearing in mind that the unit of alignment also greatly influences the way annotation may be implemented at later stages, it was decided to **time align at the level of the word** so as to offer a high level of granularity to the annotation of (dis)fluency features.

The segmentation of transcriptions into words-to-be-aligned first involves defining what a word actually is. A "word" is here defined as a letter or a sequence of letters, a symbol or a sequence of symbols, or a combination of letters and symbols separated by blank spaces. Items such as *he*, *does* or *and* are counted as words, and so are filled pauses, the dots used for unfilled pauses and paraverbal tags such as `<laughing/>` (see illustrations in Table 4-1). Although filled and unfilled pauses as well as paraverbal tags are not words in the conventional sense (although there is some debate going on, especially with respect to filled pauses, e.g. Clark and Fox Tree (2002), see also Tottie (2011; 2014)), they were given this status nonetheless purely for practical reasons.

Hyphens, even though they are not immediately followed by a blank space, were also used as a word separator. **Contracted verbal forms** were split into two "words". For example, *I'm* is split into *I* + *'m* and *she'll* is divided into *she* + *'ll*.[112] **Contracted negations** were also segmented before the apostrophe.

		XML transcription	Word segmentation
Full forms	1.	he does not	he | does | not
	2.	and I was	and | I | was
Genitives	3.	Students' Union	Students' | Union
Unfilled pauses	4.	I .. go to Los Angeles	I | .. | go | to | Los | Angeles
Paraverbal information	5.	very <laughing/> I found it funny	very | <laughing/> | I | found | it | funny
	6.	there <whistles> I think	there | <whistles> | I | think

[111] I am deeply grateful to Sophie Roekhaut (CENTAL, Université catholique de Louvain) for her expertise and for helping me with the technical aspects of word segmentation, phonetic transcription, and time alignment. Her help and patience were invaluable.

[112] Contracted verb forms were also split with a view to facilitating the automatic part-of-speech tagging procedure with Treetagger, given that the Treetagger tagset does not include a specific tag for contracted forms.

Filled pauses	7.	around (er) fifteen	around | (er) | fifteen
Truncations	8.	any sort of ex= examples	any | sort | of | ex= | examples
Lengthenings	9.	I had the[i:] opportunity	I | had | the[i:] | opportunity
	10.	every: every month	every: | every | month
Unclear words	11.	I am (erm) <unknown/> of staff	I | am | (erm) | <unknown/> | of | staff
Uncertain word endings[113]	12.	the small wig <unknown/> was	the | small | <unknown/> | wig | was
Contracted verb forms	13.	I'm allergic to	I | 'm | allergic | to
	14.	she'll be pleased	she | 'll | be | pleased
	15.	that's easy	that | 's | easy
Contracted negations	16.	he doesn't	he | doesn |'t
	17.	won't it	won | 't | it
Hyphens	18.	about twenty-five	about | twenty | five

Table 4-1: Segmentation examples (1)

Figure 4-1 to Figure 4-4 illustrate various segmentation scenarios involving not only pauses but also contracted forms (in the coloured boxes).

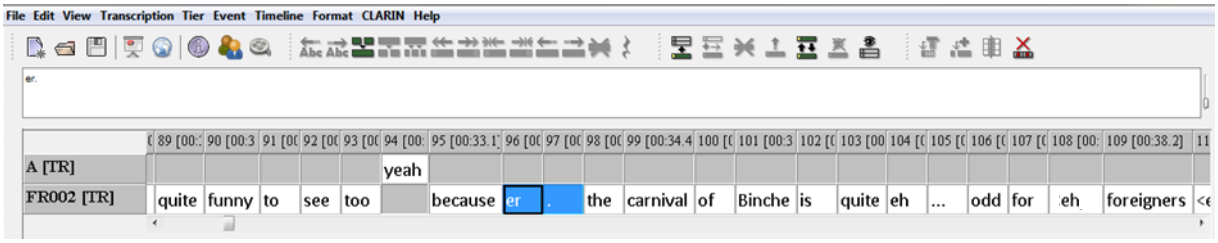


Figure 4-1: Segmentation example (1) - FR002-F

Figure 4-2: Segmentation example (2) - FR002-F

[113] Uncertain word endings were automatically converted into <unknown/> tags in the preliminary manipulations, *cf. supra*.

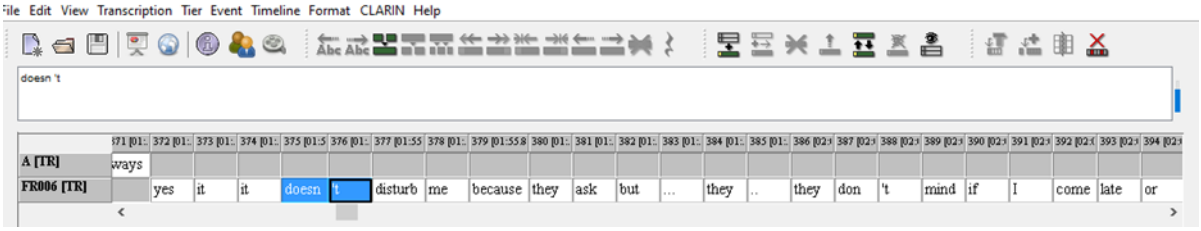


Figure 4-3: Segmentation example (3) - FR006-F

Figure 4-4: Segmentation example (4) - FR006-S

4.1.4 Phonetic transcription

The last phase before the alignment proper is the phonetic transcription of the written transcriptions. It is automated using the eLite software (Beaufort & Ruelle 2006; Roekhaut *et al.* 2014), which produces a TextGrid file as output. With a view to increasing the accuracy rate of eLite, **two sets of phonemes** were used: the "A set" for the interviewer and the "B set" for the interviewee. This way, the speech of the two speakers can be phonetically transcribed (and aligned) separately.

The phonetic transcription has not been manually checked, but the output was deemed of sufficient quality to obtain a reliable alignment at the level of words.

4.1.5 Time alignment and output

The TextGrid files resulting from the automated phonetic transcription and the audio recording were aligned by the Train&Align program (Brognaux *et al.* 2012a; Brognaux *et al.* 2012b). The advantage of this tool is that it can train an alignment model specific to the data and use it to align the transcriptions.

Two devices were used in order to **increase the accuracy rate of the alignment,** namely a bootstrap procedure and a script for the automatic detection of unfilled pauses. The **bootstrap** consists in the first LINDSEI-FR interview which was aligned automatically using the procedure explained above as well as a 60-second excerpt that I manually aligned at the

level of the phoneme. The script allows the **automatic detection and mark-up of unfilled pauses** which were not perceived/transcribed in the original transcription files.

The output after those manipulations consists of two types of files:

- **Audio files**: there is one audio file per speaking task for each interview. Each file is re-named accordingly (e.g. FR043_P for the picture description part of the 43rd interview of LINDSEI-FR, EN005_F for the free discussion part of the fifth interview in LOCNEC). In total, there are thus 300 files (3 tasks in 50 interviews in 2 corpora).
- **TextGrid files**: there is one file per speaking task for each interview. Each file is re-named accordingly (FR012_S, EN048_S etc.) and contains seven tiers, as displayed in Table 4-2.

Tier no.	Description
1	Segmented transcription of the interviewer's speech
2	Segmented transcription of the interviewee's speech
3	Annotation of the (dis)fluency features in the interviewee's speech (1)
4	Annotation of the (dis)fluency features in the interviewee's speech (2)
5	Annotation of the (dis)fluency features in the interviewee's speech (3)
6	POS tagging of the interviewer's speech
7	POS tagging of the interviewee's speech

Table 4-2: Overview of the 7 tiers in the alignment output files

Tiers 1 and 2 contain the segmented and time aligned transcriptions of the interviewer's and interviewee's speech. The original mark-up (i.e. parentheses, equal sign, colon, `[i:]`, and `[ei]`) is removed from those tiers but the interviewee's symbols are automatically converted into (dis)fluency annotation tags in tier 3, e.g. `<FP>` or `<T>` for filled pause and truncation, respectively (see Section 4.2 for further details on annotation). The **length of unfilled pauses**, which was automatically measured during the alignment, is inserted in the two transcription tiers instead of the dot(s).

Tiers 3 to 5 are dedicated to (dis)fluency annotations. While tiers 4 and 5 are empty at this stage, tier 3 already contains some annotations that were automatically converted from the original LINDSEI-FR and LOCNEC mark-up.

Tiers 6 and 7 contain part-of-speeech tags of the interviewer's and of the interviewee's speech.

For greater clarity, the aligned (and annotated) versions of LINDSEI-FR and LOCNEC are henceforth referred to as LINDSEI-FR+ and LOCNEC+.

4.1.6 Post-alignment corrections

The time alignment of LINDSEI-FR+ and of LOCNEC+ was followed by post-alignment checks whose purpose was to **check the quality and reliability of the alignment**, and increase them by **manually correcting** thealignment where necessary. These post-alignment corrections were actually made at the same time as the annotation of (dis)fluency features, but, for the sake of clarity, annotation will be described in a separate section (Section 4.2).

Each aligned transcription in .TextGrid format and its audio counterpart was first uploaded in EXMARaLDA and saved as a new .exb file (e.g. FR001-S.exb). By selecting a word or a sequence of words in the transcription tiers (see the blue boxes in Figure 4-5) and clicking on the play button in the interface ( ), it is possible to listen to the part of the audio recording that corresponds to this (these) word(s).

I then listened to the audio recording by chunks of 4 to 6 words to check whether the alignment quality was sufficient. I aimed at not having **alignment gaps** of more than one word after the chunk I was listening to. For example (see Figure 4-5), when playing the chunk *I was studying there er*, if I heard only *I was studying there er*, the alignment was considered correct and I went on with the next chunk. If, however, I could also hear the following word (i.e. *I was studying there er classical*) or more, I corrected the alignment of each word within the chunk as well as the following word(s) until the word-alignment was accurate again.

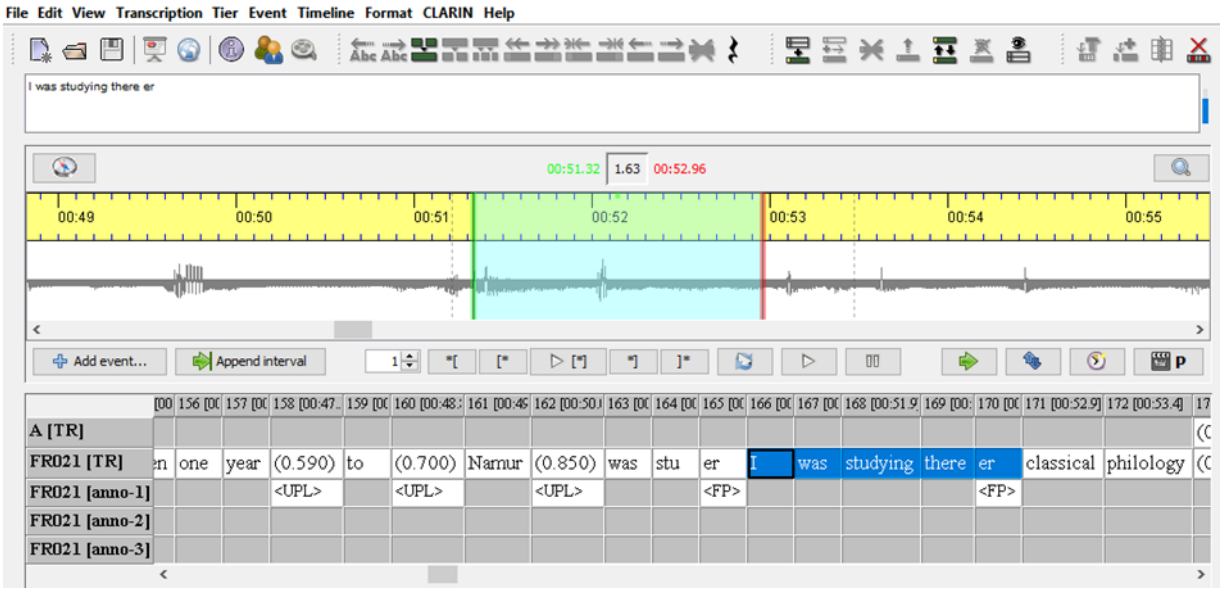


Figure 4-5: Post-alignment corrections (1) - FR021-F

I paid specific attention to the alignment of **unfilled pauses**. More specifically, I manually corrected their temporal boundaries if these were erroneous and corrected their length in the transcription tiers (as in Figure 4-6 and Figure 4-7). Listening by small chunks of words at a time also allowed me to detect unfilled pauses that still managed to escape the script (e.g. when there was background noise or when the sound quality was poor). I also checked the

alignment of **overlaps** to see if they faithfully represented the actual discourse, and corrected them when necessary (as in Figure 4-8 and Figure 4-9).

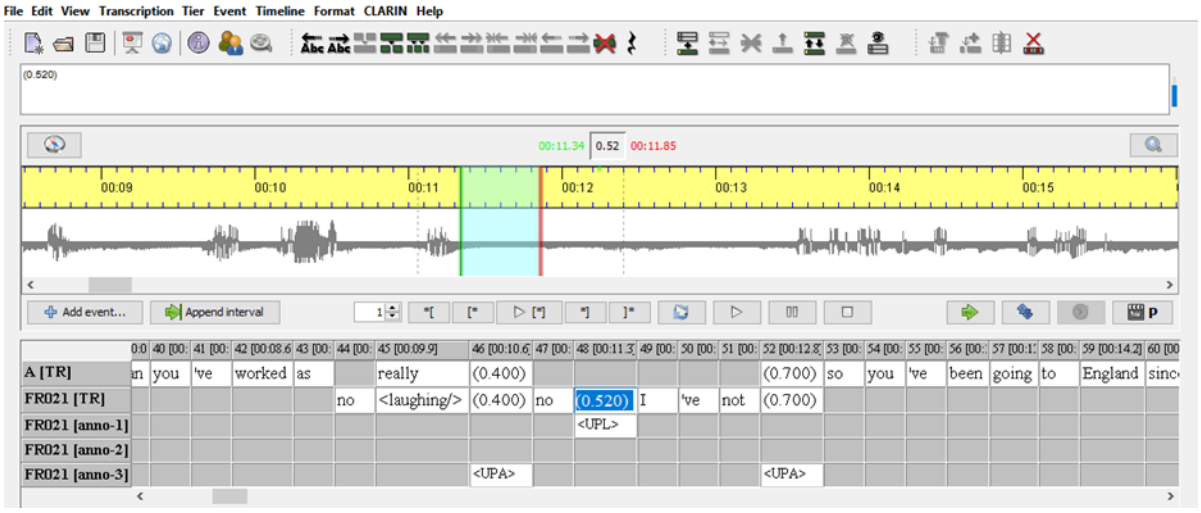


Figure 4-6: Post-alignment corrections (2) - FR021-F, raw aligned file

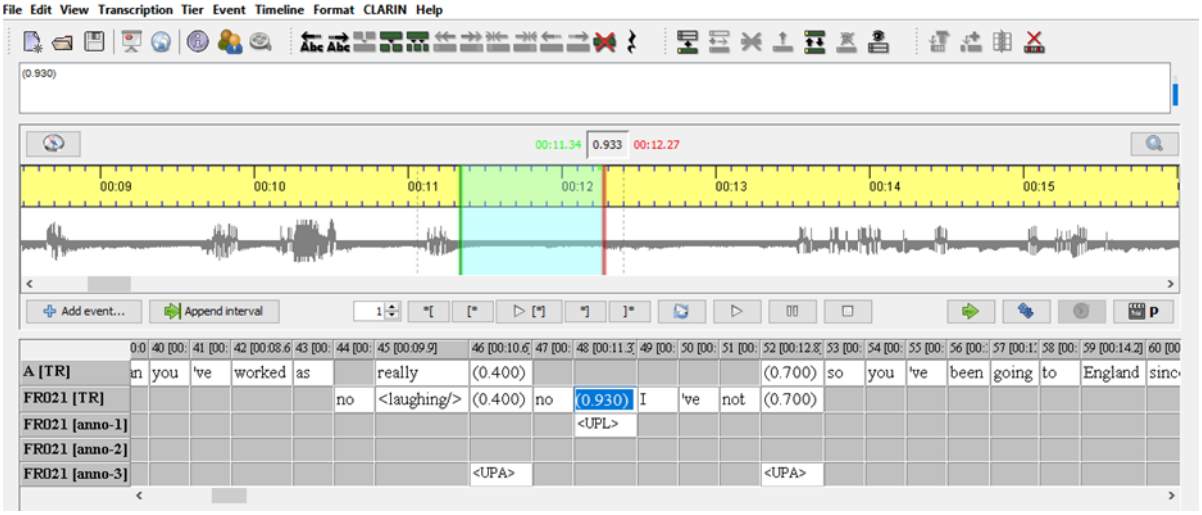


Figure 4-7: Post-alignment corrections (3) - FR021-F, with corrected alignment of the unfilled pause

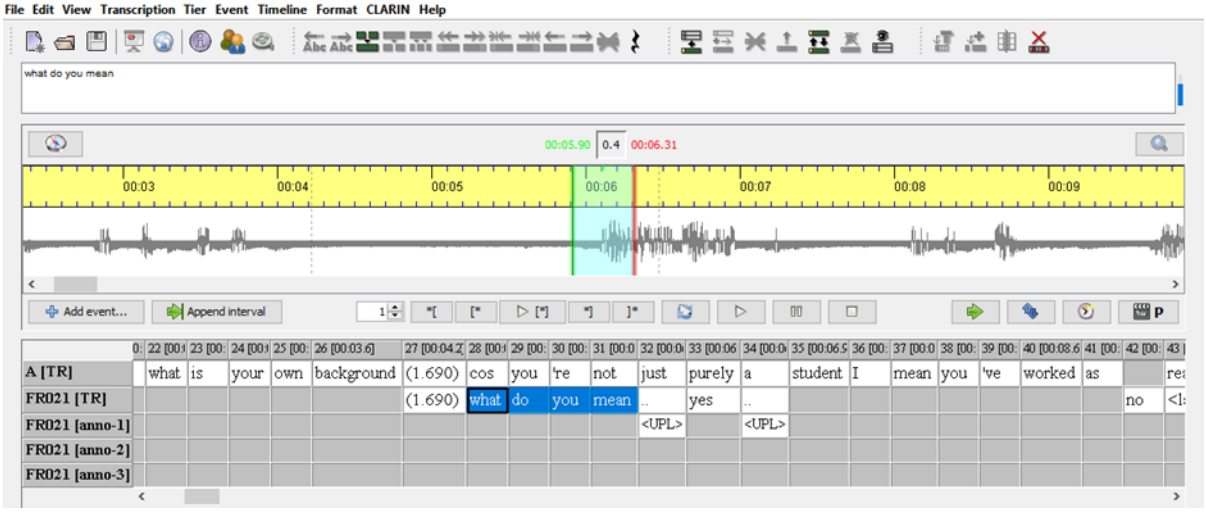


Figure 4-8: Post-alignment corrections (4) - FR021-F, raw aligned file

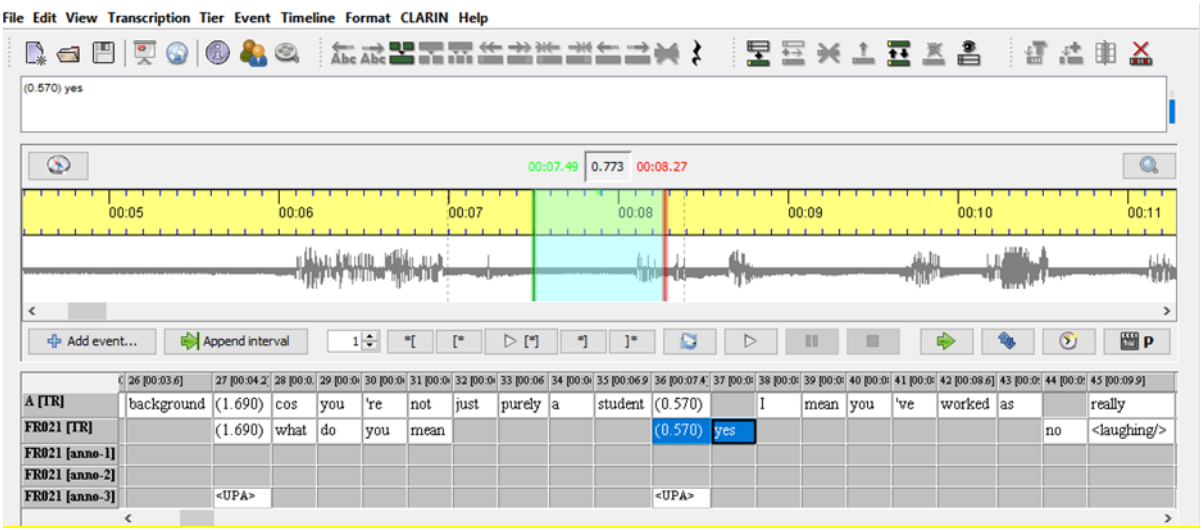


Figure 4-9: Post-alignment corrections (5) - FR021-F, with corrected alignment of the overlap

Another problematic area appeared around **paraverbal tags** (such as laughs or sighs). Paraverbal information could not be transcribed phonetically, which implies that the alignment of laughter, sighs etc. (as well as the alignment of the words following them) was not always accurate and needed correction. In addition, laughs usually caused wide differences in pitch and volume. As a consequence, the microphone decreased its sensitivity the next few seconds, and the speech just after the laughing was rendered less clear (or even sometimes barely audible), which also decreased the quality of the alignment at those places.

In some (fortunately infrequent) cases, the inaccurate alignment of unfilled pauses, overlaps or paraverbal information caused the alignment to become increasingly worse in the following seconds or even minutes because it could not "fall back on its feet" any more. When this happened, each word had to be re-aligned manually, one by one.

4.1.7 Limitations of the alignment procedure

Whereas the manipulations described in Sections 4.1.3 to 4.1.5 were performed in collaboration with Sophie Roekhaut and mainly automatically, the pre- and post-alignment corrections were made manually by myself.

The procedure for **preliminary checks** needed several revisions as is illustrated in Figure 4-10. I first performed the preliminary checks on a sample of three interviews and submitted them for segmentation, phonetic transcription and alignment. Then, I checked the alignment quality in the output and went back to the transcriptions to refine the corrections until satisfactory results were obtained. Only then did I make the preliminary corrections on the other 47 LINDSEI-FR interviews. I did the same for LOCNEC interviews a few months later.

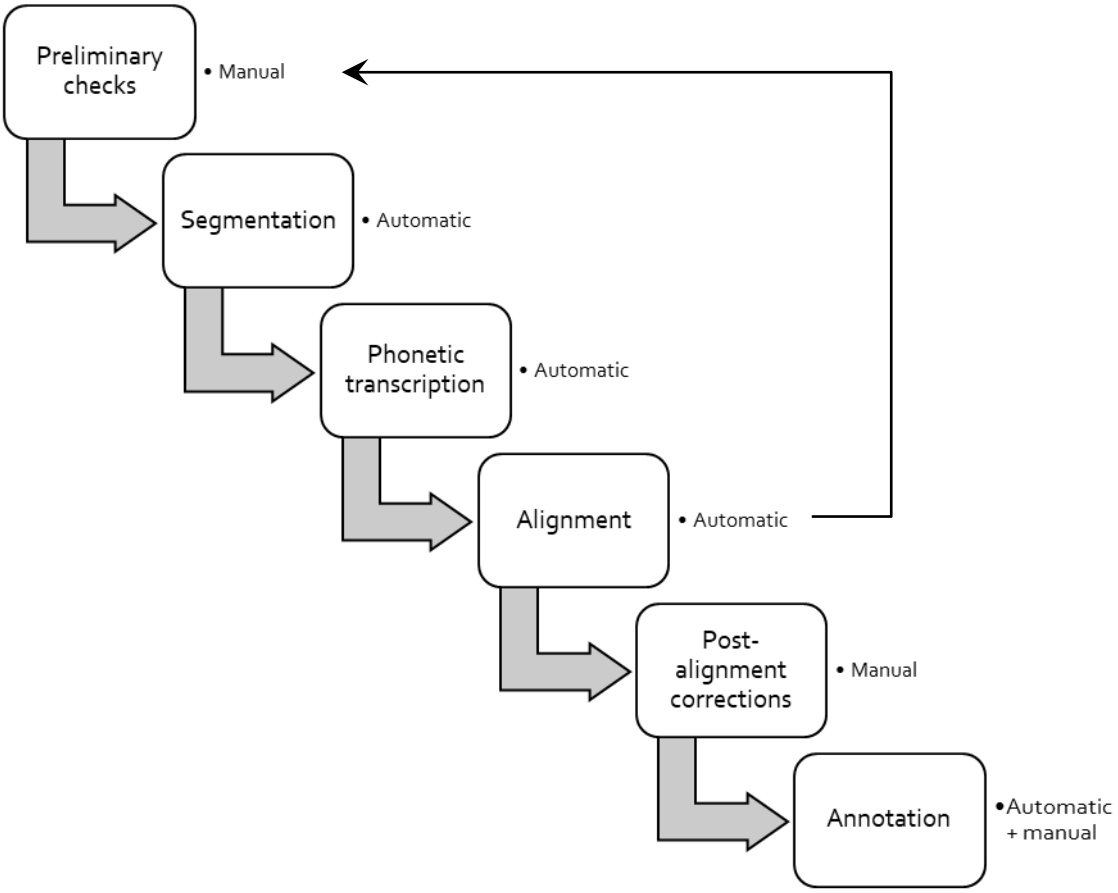


Figure 4-10: Overview of the alignment procedure

Even though much care had been taken in pre-alignment checks, the manual **post-alignment checking and correction** of each interview turned out to be **extremely time-consuming**, sometimes taking up to 8 hours to correct the time alignment of a 15-minute excerpt when a major alignment gap occurred. Fortunately, this did not happen very often and on average, I managed to correct at least one, up to two interviews, in a day. The time required depended on several factors such as the quality of the original recording (generally speaking, the higher the quality of the recording, the better the subsequent alignment), the quality of articulation of the interviewee and his/her accent, the level of interaction and the number of overlaps etc. The picture description task was the quickest to correct: it was always nearly perfectly aligned as there was little interaction and overlaps, the speakers tended to speak more slowly and the task was short (i.e. if an alignment gap did occur, it had few consequences). By contrast, the free discussion task was the longest and hardest to correct: this task generally includes a large number of overlaps, laughter, speakers talk faster and the task is very long (i.e. when an alignment gap occured, the consequences often grew exponentially).

The final result of all these manipulations and corrections, however, has many valuable advantages, including easier access to the sound file and the specificities of speech, and availability of reliable temporal data, as will hopefully be illustrated in the following section.

Before going on to the presentation of the temporal data of the aligned LINDSEI-FR+ and LOCNEC+, I would like to stress two points for the benefit of researchers considering time aligning their corpus:

- If the quality of the audio recordings is poor, e.g. due to background noise, the quality of the alignment will be greatly impacted. If other LINDSEI components were to be aligned, it would definitely be a worthwhile investment to systematically **clean the audio files** using, e.g., the software Audacity, prior to the alignment. This additional step is very quick to perform and greatly increases the quality of the aligned output.
- Automatic alignment is best followed by manual **post-alignment checks** and corrections to obtain reliable temporal data. The time and manpower needed for such corrections should definitely not be under-estimated.

4.1.8 Sizes and durations in LINDSEI-FR+ and LOCNEC+

This section presents the sizes and durations of the learner and native corpora as measured in their aligned versions. The figures presented in Table 4-3 through Table 4-7 will be used to normalise the frequencies per hundred words and per minute in further analyses.

Table 4-3 displays the unpruned **number of words** resulting from the segmentation process in each corpus. As mentioned above, the definition of "word" adopted here does not entirely correspond to the conventional definition: filled pauses or truncated words are for example counted as single words too. Only unfilled pauses and paraverbal tags are excluded from the word counts presented below (although they were considered as "words" for the purposes of time alignment). The first row in the table includes the number of words produced by the interviewer ("A") and the second the number of words produced by the interviewee ("B"). Inter-turn pauses – unfilled pauses between the two speakers' turns – are presented in a separate row because it is not possible to attribute each of these pauses to one or the other speaker with a fair degree of objectivity and certitude (see also Tavakoli (2016) for a discussion of the status of inter-turn pauses). Table 4-4 below then reveals the number of words (excluding unfilled pauses) produced by the interviewee per speaking task in each corpus.

No. of words	LINDSEI-FR+	LOCNEC+
Speaker A *Mean per interview*	52,839 *1,057*	43,637 *873*
Speaker B *Mean per interview*	94,993 *1,900*	128,857 *2,577*
Inter-turn pauses *Mean per interview*	2,415 *48*	3,272 *65*

Table 4-3: Number of words in LINDSEI-FR+ and LOCNEC+

No. of words	Set topic	Free discussion	Picture description
LINDSEI-FR+ *Mean per interview*	45,023 *901*	43,178 *864*	6,792 *136*
LOCNEC+ *Mean per interview*	46,307 *926*	74,901 *1,498*	7,649 *153*

Table 4-4: Number of words per task in LINDSEI-FR+ and LOCNEC+ (B turns only)

In LINDSEI-FR+, the **learners** produce **1,900 words on average per interview** (about 95,000 words in total); in LOCNEC+, the British native speakers utter quite more: 2,577 words on average per speaker (about 130,000 words in total). As shown in Table 4-4, whereas in the learner corpus, the set topic and the free discussion tasks are of about the same size (c. 45,000 words; c. 900 on average per speaker), in LOCNEC+, the word count is much higher in the second speaking task than in the set topic with c. 75,000 words in the second task (mean per interview: 1,498 words) vs. c. 46,000 in the first task (mean: 926 words). The picture description elicits the lowest number of words in both corpora: the average word count in LINDSEI-FR+ amounts to 136 words and in LOCNEC+ to 153 words.

Being time aligned, the two corpora can now be measured precisely. Table 4-5 through Table 4-7 display the precise **durations of each corpus and of each speaking task**. Inter-turn pauses are not included in any of the speakers' speaking time but their duration is presented in a separate row, for reasons explained above. As can be seen, the **interviewees' total speaking time** is **about 10 hours** in each corpus (c. 11 minutes and a half on average per interview). Again, in LINDSEI-FR+, the figures are very similar for the set topic and the free discussion, where the interviewees speak for about 4 hours and a half in total (c. 5 minutes on average per interviewee). In the native corpus, however, the interviewees speak longer in the free discussion task (5 hours and a half; mean: 6.5 minutes) than in the set topic (3 hours and a half; mean: 4 minutes). Each learner speaks for about 1 minute in the picture description task, which amounts to about 50 minutes of speech for this task in LINDSEI-FR+. The figures for the interviewees in LOCNEC+ are slightly lower, but still very similar, with 49 seconds spent on average for this task (41 minutes in total).

Speaking time	LINDSEI-FR+	LOCNEC+
Speaker A *Mean per interview*	4h 32min 32s *5min 27s*	3h 32min 55s *4min 15s*
Speaker B *Mean per interview*	9h 43min 54s *11min 41s*	9h 37min 04s *11min 32s*
Inter-turn pauses *Mean per interview*	0h 20min 39s *0min 25s*	0h 30min 22s *0min 36s*
Total duration[114]	**14h 23min 48s**	**13h 02min 04s**

Table 4-5: Speaking times in LINDSEI-FR+ and LOCNEC+

Speaking time	Set topic	Free discussion	Picture description
Speaker A *Mean per interview*	1h 36min 01s *1min 55s*	2h 37min 51s *3min 09s*	0h 18min 40s *0min 22s*
Speaker B *Mean per interview*	4h 36min 56s *5min 32s*	4h 15min 30s *5min 07s*	0h 51m 28s *1min 02s*
Inter-turn pauses *Mean per interview*	0h 07min 40s *0min 09s*	0h 10min 25s *0min 12s*	0h 02min 34s *0min 03s*
Total duration	**6h 04min 15s**	**6h 42min 30s**	**1h 10min 29s**

Table 4-6: Speaking times per task in LINDSEI-FR+

Speaking time	Set topic	Free discussion	Picture description
Speaker A *Mean per interview*	0h 54min 24s *1min 05s*	2h 27min 29s *2min 56s*	0h 11min 02s *0min 13s*
Speaker B *Mean per interview*	3h 29min 46s *4min 12s*	5h 26min 16s *6min 32s*	0h 41min 02s *0min 49s*
Inter-turn pauses *Mean per interview*	0h 09min 27s *0min 11s*	0h 16min 36s *0min 20s*	0h 04min 19s *0min 05s*
Total duration	**4h 20min 50s**	**7h 46min 28s**	**0h 54min 36s**

Table 4-7: Speaking times per task in LOCNEC+

[114] The total duration of the corpus (and of each speaking task) is slightly lower than the sum of A's speaking time + B's speaking time + inter-turn pauses. This is perfectly normal as the two speakers sometimes speak at the same time.

Although it might be very tempting to compare LINDSEI-FR+ and LOCNEC+ sizes and durations with other spoken (learner) corpora, this proves to be a particularly tricky enterprise. Corpus durations and what they actually include (or not) are not always clearly mentioned in papers. More specifically, many studies make use of dialogues or other interactive tasks but it is very rare for the authors to report the durations of the learner's speech – usually, only the overall duration of the task or the corpus is reported. In addition, inter-turn pauses are typically overlooked. This being said, compared to the data used in the recent (dis)fluency literature, the figures presented above are remarkable in terms of size: Derwing *et al*. (2009) used 20 seconds of L2 speech per speaker at 3 time points; Préfontaine and Kormos (2015; 2016) used 1 to 5 minutes per speaker in each one of 3 narrative tasks; and Tavakoli (2016) used 1 minute of monologue and 3 minutes of dialogue per speaker. By contrast, de Jong (2016) used a corpus of similar size as the one used in this study: it totals 15 hours for 72 participants – so, about 12 minutes per speaker. One major difference with LINDSEI-FR+ and LOCNEC+, however, is that in her corpus, each speaker performed eight speaking tasks: more tasks are represented, but the speaking time per task is also comparatively shorter than in LINDSEI-FR+ and LOCNEC+.

The second part of this chapter is devoted to corpus annotation: I present the main principles that have underpinned the design of the annotation system and provide a quantitative overview of the (dis)fluency annotations in the two corpora.

4.2 CORPUS ANNOTATION

As specified in Chapter 3, LINDSEI-FR and LOCNEC were collected prior to the beginning of the present research project. The transcription guidelines (see Appendix 9.1) were aimed to capture some specificities of spoken language: filled and unfilled pauses, as well as truncated words were, for example, indicated in the transcriptions by means of specific symbols. In the frame of this thesis, this type of transcription is already a tremendous asset because the mark-up enables the easy retrieval of each filled pause, truncated word etc. in its discursive context. However, not all linguistic phenomena that are considered constitutive of (dis)fluency in the present study are marked in the original LINDSEI-FR and LOCNEC transcriptions: reformulations (i.e. restarts) are, for example, not marked, and neither are repetitions or false starts. Without annotation, instances of these phenomena cannot be retrieved and, consequently, their analysis is particularly difficult, if not impossible. Besides, given that time alignment necessarily also involves a major reshaping of the transcriptions – their mark-up included – it soon became apparent that a new annotation of the LINDSEI-FR and LOCNEC corpora would be a major determinant of this study.

The annotation process was carried out in three steps: the design of the annotation scheme (Sections 4.2.1, 4.2.2, and 4.2.3), the annotation proper, and, finally, the evaluation of intra-annotator reliability (both in Section 4.2.4).

4.2.1 The design of a (dis)fluency annotation scheme

4.2.1.1 Main theoretical and methodological principles

Several theoretical principles have underpinned the design of the (dis)fluency annotation system.

The main hypothesis of the research project out of which this annotation system has arisen is that fluency and disfluency are **the two sides of the same coin**. In other words, the same feature can be used as a means to enhance fluency at one point, and as a marker of disfluency at another, and it is in the recurrence and combination of these features that fluency or disfluency can be established. Consequently, the annotation system attempts to make, as far as possible, no a priori decision as to which elements should be considered fluent or disfluent: **all occurrences of a feature are annotated with the same tag**. For instance, all unfilled pauses are marked using the same annotation tag, whether they are short or long, serve as a structuring device or are a strategy to gain time.

The **integrated componential approach** to (dis)fluency, i.e. (dis)fluency seen as a variety of features contributing to a holistic phenomenon, constitutes the second cornerstone of the system. While the protocol offers the possibility of annotating a dozen distinct (dis)fluency features and analyse them separately (i.e. from a componential perspective), it also makes it possible to draw a holistic picture of individual speakers' (dis)fluency behaviour through correlations of features (i.e. the holistic side of the system).

Moreover, on a methodological level, several additional principles have been observed. First, the system is designed **for and on the basis of spoken data**: the (dis)fluency annotation protocol is solely based on concepts of spoken language such as filled pause, truncated word or false start. Reference to written grammatical concepts is avoided as much as possible. A related principle is that the annotation system combines a **theory-motivated**[115] basis with bottom-up, **induction-oriented**, amendments, which hopefully lessens potential ambiguities or doubts during the annotation process. Practically, this implies that preliminary versions of the annotation scheme were iteratively tested on a sample of LINDSEI-FR+ and amended accordingly to reach the final version. Furthermore, specific attention was paid to the **applicability** of the annotation system. The (dis)fluency annotation system is aimed to be as flexible as possible in terms of global architecture and annotation format so that it can be applied to large corpora, to different speaking tasks, and to both learner and native speaker language.

4.2.1.2 The annotation tool

On a practical level, a last important consideration for the design of the (dis)fluency annotation system pertains to the constraints inherently imposed by the **annotation tool**. Many of the tools that are available for the annotation and analysis of audio and/or video data offer useful features and seem very potent and promising, but, as rightly stated by Rohlfing *et al*. (2006:99), "[f]or a researcher looking for an annotation tool, it is difficult to decide about its usefulness and usability". Furthermore, "[t]o decide about usefulness and usability, it is necessary to know about the ease of use, strengths/weaknesses for specific annotation purposes, and the type of data or analysis the tool is designed for – knowledge that is usually gained only after becoming an expert in the use of a particular tool" (*ibid*.).

Some of the best known tools include ELAN (http://www.mpi.nl/tools/elan.html), EXMARaLDA (Schmidt 2001; Schmidt & Wörner 2014) and Praat (http://www.fon.hum.uva.nl/praat/). Space precludes an exhaustive report of the strengths

[115] Bear *et al*. (1993); Eklund (2004); Heeman, McMillin and Yaruss (2006); Meteer (1995); Rodríguez, Torres & Varona (2001) and Shriberg (1994) have, among others, suggested various interesting and well-thought ways of categorising and annotating (some) (dis)fluency features. Their work forms the theory-motivated part of the system.

and weaknesses of each tool. Suffice it to say that EXMARaLDA was selected for carrying out the annotation because it is more general than Praat[116] and it offers more flexibility than ELAN with respect to import and export functions, annotation and search possiblities.

EXMARaLDA ("the *EXtensible MARkup Language for Discourse Annotation*") was created at the University of Hamburg by T. Schmidt. It is freely available (http://exmaralda.org/en/) and is still being improved based on user suggestions or requests. It consists in three software tools for the creation, management and analysis of spoken corpora, namely:

- The Partitur-Editor, which can be used for inputting and outputting transcriptions of spoken data following the layout of musical scores (i.e. with separate lines for each speaker or modality). It also includes a simple search functionality.
- The Corpus Manager, a tool designed for corpus construction, the management of transcriptions, and the querying of metadata.
- The EXAKT query tool, an elaborate search tool for concordancing.

4.2.2 Global architecture

The (dis)fluency annotations are displayed in a **multi-layered scaffolding** consisting of successive, complementary and interconnected levels (so-called "tiers"). As explained by Gries and Berez (2015; also Eckart 2012), multi-tiered annotation has the advantage of structuring the data by assuming a relationship between the annotations that are displayed on different tiers. Another advantage of having a multi-layered annotated resource is that information encoded in one layer of representation can be used to infer that of another.

In this protocol, the various annotation tiers (the "daughter-tiers") depend either on the interviewee's or interviewer's transcription tier (the "mother-tiers") (*cf.* also Table 4-2). The daughter-tiers include three (dis)fluency annotation tiers and two tiers containing part-of-speech tags.

The three **(dis)fluency annotations tiers**, termed [anno-1], [anno-2] and [anno-3], all depend on the interviewee's transcription tier. In these tiers, the labelling system focuses on the linear distribution of (dis)fluency features following a three-plane architecture. While the first level of annotation ([anno-1]) provides generic tags for 10 (dis)fluency features, the second level ([anno-2]) offers the opportunity to dive deeper into specific characteristics of features annotated at the first level. The third level ([anno-3]) allows for the annotation of additional elements which are not considered as (dis)fluency features per se (e.g. grammatical or lexical

[116] As specified by the main title on the official website ("Praat: doing phonetics by computer"), Praat is firstly aimed at phonetic studies.

mistake, difficulty in articulating a word) or to add occasional remarks or observations (e.g. "reported speech", "paraphrase", "grammatical error").

In addition to these tiers, **two tiers contain part-of-speech tags** (one tier per speaker). These tiers are named [POS]. The POS tagging is automatically generated by Treetagger – the tagset[117] and tagging guidelines of which are described in Santorini (1990).

At a later stage of the analysis, I also included an additional tier containing the interviewee's speech manually segmented into **speech runs**. Following Lennon (1990) and Götz (2011), I considered as a speech run a segment of speech that is surrounded by unfilled pauses and/or the beginning and/or end of a turn (as indicated by the interviewer's speaking). As this tier does not contain annotations in the core sense of the term, it was set as a "description" tier in EXMARaLDA.

Table 4-8 below summarises the eight tiers (i.e. 2 transcription tiers, 1 description tier and 5 annotation tiers) that make up the multi-layered architecture of each LINDSEI-FR+ and LOCNEC+ annotated file. An illustration of the raw architecture in the EXMARaLDA interface is presented in Figure 4-11. The same excerpt, after annotation, is illustrated in Figure 4-12.

Tier no.	Tier name in EXMARaLDA	Type of tier in EXMARaLDA	Description
1	A [TR]	Transcription	Segmented transcription of the interviewer's speech
2	FR/EN*nnn* [TR]	Transcription	Segmented transcription of the interviewee's speech
3	FR/EN*nnn* [runs]	Description	Transcription of the interviewee's speech segmented into speech runs
4	FR/EN*nnn* [anno-1]	Annotation	Annotation of the (dis)fluency features in the interviewee's speech (1)
5	FR/EN*nnn* [anno-2]	Annotation	Annotation of the (dis)fluency features in the interviewee's speech (2)
6	FR/EN*nnn* [anno-3]	Annotation	Annotation of the (dis)fluency features in the interviewee's speech (3)
7	A [POS]	Annotation	POS tagging of the interviewer's speech
8	FR/EN*nnn* [POS]	Annotation	POS tagging of the interviewee's speech

Table 4-8: The 8 tiers in the annotated LINDSEI-FR+ and LOCNEC+

[117] The tagset is also available online at https://courses.washington.edu/hypertxt/csar-v02/penntable.html (last accessed 4/07/2017).

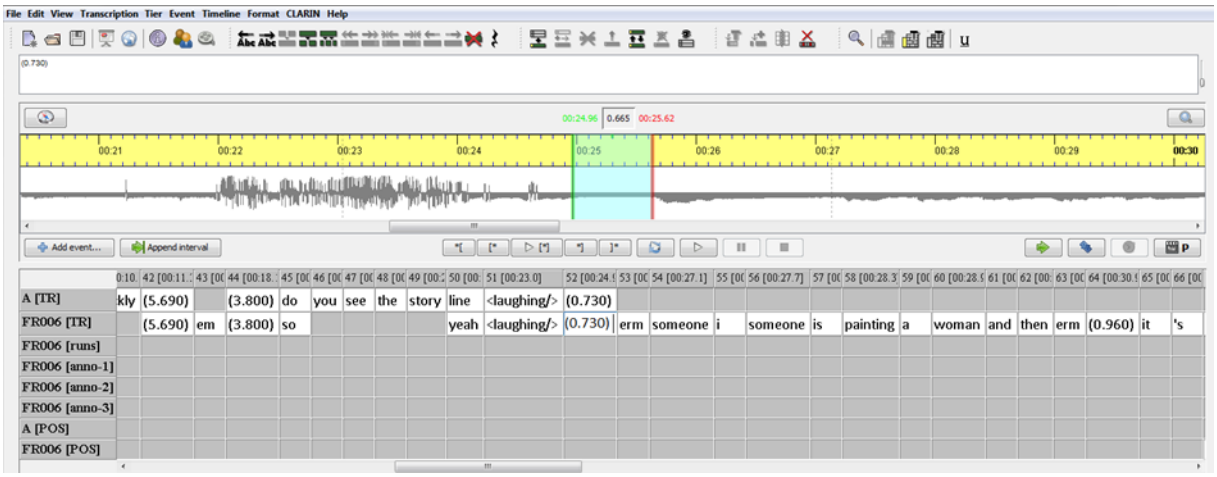


Figure 4-11: The multi-tiered architecture (FRoo6-P; raw version)

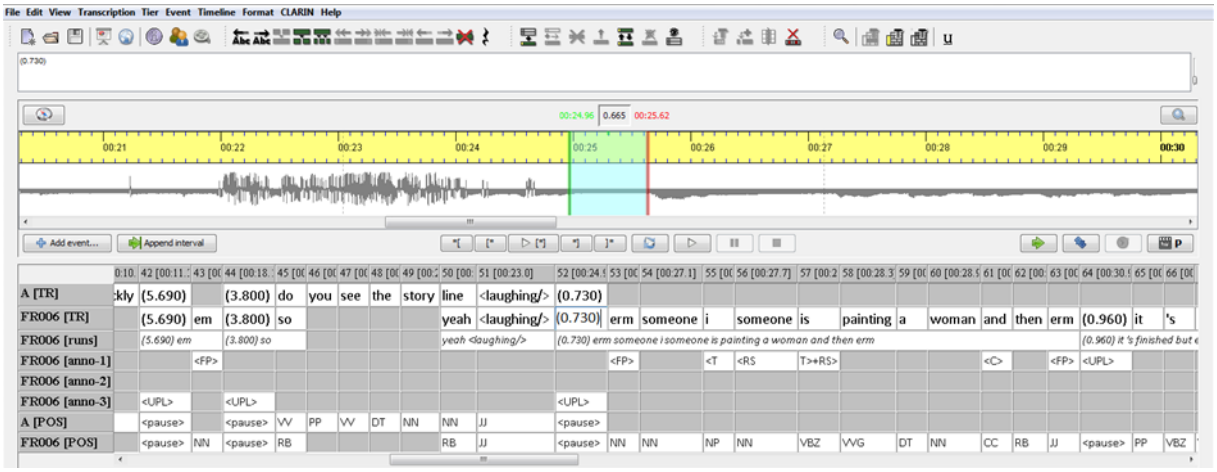


Figure 4-12: The multi-tiered architecture (FRoo6-P; annotated version)

In the next sections of this chapter, the **annotation examples** are displayed in a table-format where each line corresponds to a tier. The interviewee's transcription tier is always included (the tier with the speech runs is not included, unless specified). If there is an overlap with the interviewer, or if there is some interaction between the two speakers, the two transcription tiers are included, first the interviewer, then the interviewee. Anno-1 is always displayed, anno-2 and anno-3 are included when relevant (i.e. when they are not empty). The [POS] tiers are not included either.

4.2.3 (Dis)fluency annotation protocol

Except for tier 3 which contains speech runs, the annotation tiers are automatically segmented at the same level as their mother-tier, that is, at the level of the word (see Section 4.1.3). This implies that the design of the annotation system has to conform to this level of granularity.

The system makes use of three devices, namely letters, brackets, and symbols.

A. Letters

Each (dis)fluency feature is tagged by means of **one or two capital letters**. These letters generally correspond to the initial letters of the name of the feature, e.g. FP for filled pause or DM for discourse marker.

The letter tag is applied **under each word affected by the (dis)fluency feature**. In Example 4-10, the annotation tags of the two unfilled pauses (UPL) are integrated in the annotation tier just below the pauses in [anno-1]. Likewise, in Example 4-11, the first discourse marker has a corresponding tag DM in the first annotation tier, but the second has two, one below each word in the discourse marker (i.e. *I* and *mean*).

4-10: FR021-S

I	don	't	go	for	(0.350)	holiday	(0.250)	at	other	times
					<UPL>		<UPL>			

B. Angle brackets

The letter tags are coupled with a bracketing system indicating the beginning and end of each phenomenon: the **onset** of a (dis)fluency feature is marked by an opening bracket (<) and the **offset** of a (dis)fluency feature is marked by a closing bracket (>).

If the (dis)fluency feature affects a single word, the opening bracket immediately precedes the letter tag and the closing bracket immediately follows it (i.e. without blank spaces). For example, *well*, which is a one-word unit discourse marker, is simply marked by the tag <DM>. If the (dis)fluency feature affects more than one word, the opening bracket is added immediately before the first letter tag, and the closing bracket right after the last letter tag. For example, for two-word discourse markers such as *I mean*, the first word (*I*) is marked with the opening tag <DM and the last (*mean*) is annotated with the closing tag DM>. Consider also Example 4-12, where only the onset and the offset of the restart are indicated by means of opening and closing brackets, with the middle tag consisting of letters only.

4-11: FR021-S

well	yeah	I	mean	I	don	't	go	for
<DM>		<DM	DM>					

4-12: FR005-S

er	if	he	knows	if	he	knew	something
				<RS	RS	RS>	
						<SM>	

C. Symbols

Some words may be involved in more than one (dis)fluency feature at a time: a repetition may include a lengthening, a discourse marker can be uttered in another language, etc. When more than one tag has to be applied on the same word, the **tags are ordered alphabetically** and **linked with a + sign** (without spaces). Example 4-13 is an illustration of a lengthening within a repetition: the lengthening tag <L> precedes the repetition tag <R0 under the word a, and the two tags are linked with a + sign.

4-13: FR002-F

it	's	a	(0.230)	a	round	instrument	with	er
		<L>+<R0	<UPL>	R1>				<FP>
			<N>					

If the tags for restart (RS) and repetition (R*n*) need to be ordered alphabetically, RS is annotated first, then R*n*. If two *Rn* tags are annotated on the same word, they are ordered in increasing order of *n*: R0 first, then R1, R2 etc., as shown in Examples 4-14 and 4-15.

4-14: FR015-F

a	(0.300)	a	loud	(0.230)	a	loud	noise
<R0	<UPL>	<R0+R1>	R0	<UPA>	R1	R1>	
	<N>			<N>			

4-15: FR028-S

it	's	written	in	in	the	in	the	(0.240)	the	Gospel
			<R0	<R0+R1>	R0	R1	<L>+<R0+R1>	<UPA>	R1>	
								<N>		

This combination of letters, brackets, and symbols makes it possible to cover the three possible cases that may occur during the annotation:

- one (dis)fluency feature that corresponds to one word, e.g. a pause, a vowel lengthening, a one-word discourse marker (e.g. *well*);
- one (dis)fluency feature that covers several words, e.g. a repetition, a two-word discourse marker (*you know*, *I mean*);
- one word that is included in several (dis)fluency features, e.g. a vowel lengthening within a repetition (e.g. *the: the*), or a truncation within a restart (e.g. *the man . the big m= the tall man*).

In the following, the annotation of (dis)fluency features is presented level per level, from [anno-1] to [anno-3]. For each feature, a brief definition (see also Section 1.2 for wordier definitions), the annotation tag and corpus illustrations are provided. Possible queries for the extraction of the features in EXAKT, the concordancing tool within EXMARaLDA, are presented in Section 4.2.3.4.

4.2.3.1 First level of (dis)fluency annotation ([anno-1])

The first level of annotation includes generic tags for ten (dis)fluency features (see Table 4-9).

	(Dis)fluency feature	**Annotation tag**
1.	Filled pause	FP
2.	Unfilled pause	UPA & UPL
3.	Vowel lengthening	L
4.	Truncated word	T
5.	Foreign word	W
6.	False start	FS
7.	Repetition	R*n*
8.	Restart	RS
9.	Discourse marker	DM
10.	Conjunction	C

Table 4-9: The ten (dis)fluency features annotated in anno-1

A. Filled pauses

Definition: "A filled pause is occupied not by silence, but by a vowel sound, with or without accompanying nasalization" (Biber *et al.* 1999:1053). Back-channeling (*mm, uhu* and *mhm*) is not included in this category.

Annotation: Filled pauses are marked with the tag `FP`. Filled pauses always involve a single word: they are thus always tagged `<FP>`.

4-16: Annotation of filled pauses (1) - FR002-S

we	love	to	go	to	**eh**	dancings
					<FP>	

4-17: Annotation of filled pauses (2) - FR001-F

women	**er**	[…]	have	**er**	(0.930)	graduated	s	something
	<FP>			**<FP>**	<UPL>		<T	T>

B. Unfilled pauses

Definition: Period of silence or the occurrence of non-speech acoustic events such as breathing (Gut 2009:80; Riggenbach 1991:426; Biber *et al.* 1999:1053).

Annotation: Unfilled pauses are marked with the tag UP. This tag is complemented by an additional letter (L or A[118]) to distinguish between pauses that were transcribed in the original transcriptions on a perceptive basis (UPL) and unfilled pauses that were not perceived by the transcriber but were detected automatically during the alignment process (UPA). It is important to keep track of this difference because it shows how sensitive (or not) transcribers are to unfilled pauses in general and to their length, or context of occurrence in particular. Note that, unlike UPLs, a threshold of 200 ms was set for UPAs. Unfilled pauses always involve one unit: they will thus always be tagged <UPL> or <UPA>.

4-18: Annotation of unfilled pauses (1) - FR015-F

it	makes	erm	a	**(0.300)**	a	loud	**(0.230)**	a	loud	noise
		<FP>	<R0	**<UPL>**	<R0+R1>	R0	**<UPA>**	R1	R1>	
				<N>			<N>			

4-19: Annotation of unfilled pauses (2) - FR002-F

there	are	**(0.420)**	a	**(0.740)**	a	lot	of	different	places
		<UPA>	<L>+<R0	**<UPA>**	R1>				
				<N>					

C. Vowel lengthenings

Definition: The notion of lengthening (also sometimes referred to as "drawls" or "sound stretches") refers to the lengthening of the final vowel and to nonreduced vowels as in *a* ([ei]), *to* ([to:]) and *the* ([ði:]) (Fox Tree & Clark 1997:152).

Annotation: Vowel lengthenings are marked by L. Lengthnings always pertain to a single word and are thus always tagged by <L>.

4-20: Annotation of vowel lengthenings (1) - FR002-F

there	are	(0.420)	**a**	(0.740)	a	lot	of	different	places
		<UPA>	**<L>**+<R0	<UPA>	R1>				
				<N>					

[118] The letter L stands for "LINDSEI-FR/LOCNEC" and A for "added".

4-21: Annotation of vowel lengthenings (2) - FR026-F

I	(0.210)	I	erm	I	go	**to**	a	dance	course
<L>+<R0	<UPA>	R1	<FP>	R2>		**<L>**			
	<N>		<N>						

D. Truncated words

Definition: Truncated words are defined as "midword interruption[s]" (Levelt 1983:57; Brennan & Schober 2001:277) or "une interruption de morphèmes en cours d'énonciation" (Pallaud 2002:79).

Annotation: Truncated words are marked by the tag T. If the truncation is abandoned (i.e. never completed), the tag <T> is used (as in Example 4-22). If the truncation is completed, the tag <T is used for the truncation itself, and T> for the completion (see examples 4-23 and 4-24).

4-22: Annotation of abandoned truncations - FR008-F

and	(0.180)	well	**m**	(0.130)	perhaps	erm	this	(0.390)	this	summer
<C>	<UPL>	<DM>	**<T>**	<UPL>		<FP>	<R0	<UPL>	R1>	
								<N>		

4-23: Annotation of completed truncations (1) - FR002-F

b	**because**	they	er	they	have	to	dress	themselves
<T	**T>**	<R0	<FP>	R1>				
			<N>					

4-24: Annotation of completed truncations (2) - FR033-F

it	is	a	(0.220)	more	or	less	**autobi**	**biography**
			<UPA>				**<T**	**T>**

In case of successive truncations of the same word, the first truncation (or the first truncations) is tagged as an abandoned truncation (<T>). The last truncation can be completed (<T followed by T>) – as in 4-25 – or abandoned (<T>). Identical truncated words are not annotated as repetitions.

4-25: Annotation of successive truncations - FR002-S

I	realized	(0.230)	**o**	**o**	**on**	the	moment	it	was	funny
		<UPA>	**<T>**	**<T**	**T>**					

When a word is uttered in full after having been interrupted, the completion may be preceded by a restart (RS). In those cases, the completion is considered to be part of the restart and consequently receives the two tags (i.e. RS>+T>). Two examples are shown in 4-26 and 4-27.

4-26: Annotation of completed truncations with restart (1) - FR002-F

they	**d**	(0.310)	they	**dance**	around	the	fire
	<T	<UPA>	<RS	RS>+**T>**			
		<N>					

4-27: Annotation of completed truncations with restart (2) - FR021-S

we	**m**	we	**met**	(0.410)	er	we	just	met
	<T	<RS	RS>+**T>**	<UPL>	<FP>	<RS	RS	RS>
							<I>	

E. Foreign words

Definition: "[J]uxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or subsystems" (Gumperz 1982:59).

Annotation: Words in another language than English are tagged W. If they are used in a sequence (e.g. expressions), each word is assigned the <W> tag (as in 4-29). Note, however, that cities that do not have an English equivalent (e.g. *Louvain-la-Neuve* or *Namur*) are not tagged as foreign words (as in 4-31).

4-28: Annotation of foreign words (1) - FR004-F

he	's	doing	a	**régendat**	in	er	(1.240)	erm	(0.690)	modern	languages
			<L>	**<W>**		<FP>	<UPL>	<FP>	<UPL>		

4-29 : Annotation of foreign words (2) - FR002-F

they	er	(2.840)	**vont**	**cherch**	**venir**	**chercher**	(0.330)	they	er
	<FP>	<UPL>	**<W>**	<T+**<W>**	<RS+**<W>**	RS>+T>+**<W>**	<UPA>	<RS>	<FP>
					<SM>			<Del>	

4-30: Annotation of foreign words (3) - FR022-S

there	is	the	famous	camp	with	**Arbeit**	**macht**	**frei**
						<W>	**<W>**	**<W>**

4-31: Annotation of foreign words (city) - FR049-F

I	don	't	really	like	the	(0.290)	architecture	of	er	**Louvain**	**la**	**Neuve**
					<L>	<UPA>			<FP>			

F. False starts

Definition: "[C]ases that can be [...] called 'false starts', or even more precisely retrace-and-repair sequences [...] occur when the speaker retraces (or notionally 'erases') what has just been said, and starts again, this time with a different word or sequence of words" (Biber *et al.* 1999:1062; my emphasis)[119]. The interruption is marked by a grammatical incompletion; the formal aspect of words (not the lemma) is taken as a basis to distinguish "different words".

Annotation: False starts are annotated with the tag FS. Only the word before which the interruption occurs is annotated as a false start. The tag is thus always <FS>.

4-32: Annotation of false starts (1) - FR010-F

Louvain	la	Neuve	is	known	**for**	er	(0.240)	everything	is	(0.140)	close
					<FS>	<FP>	<UPA>			<UPA>	

4-33: Annotation of false starts (2) - FR002-F

it	**'s**	er	(0.650)	they	are	(0.250)	they	form	circle
	<FS>	<FP>	<UPL>			<UPL>	<RS	RS>	
								<SP>	

G. Repetitions

Definition: "Any repetition forms a block in speech that contains at least two elements: a first element which we will call the '*repeatable*'" and a second element, identical to the first, which we will call the '*repeated*'. In theory, any unit produced in speech is in principle a *repeatable*, and it is only the presence of a *repeated* immediately afterwards that makes this repeatable part of a block which we call *a posteriori* a 'repetition'"[120] (Candea 2000:315; my translation; my emphasis).

[119] Repetitions and restarts are also false starts in the sense that the speaker says something, goes back ("retraces") and begins again, but they differ from false starts in the 'strict sense' because the new beginning takes the form of a (partial) repetition while in the case of false starts the speaker starts totally afresh with a different set of words (see also Biber *et al.* (1999:1062) in this respect).

[120] Original quote: "[T]oute répétition forme un bloc dans la parole qui comporte au minimum deux éléments: un premier élément que nous appellerons le *« répétable »* et un deuxième élément, identique au premier, que

Annotation: Repetitions are tagged with the letter R. The adjunction of a number after the initial R distinguishes the different stages of the repetition: the repeated element(s) in the initial commitment is (are) marked by an additional 0 (i.e. R0), by a 1 in the first repetition (R1), by a 2 in the second (R2) etc. Note that a repetition may sprawl over non-propositional elements (typically a filled or an unfilled pause, but also a discourse marker), as can be seen in Example 4-34, where a triple repetition is annotated. If a word is part of two repetitions, the two tags are ordered in increasing order, as in 4-35 and 4-36.

4-34: Annotation of one-word repetitions - FR041-F

in	a	country	town	**where**	**where**	(0.600)	**where**	I	was
				<R0	**R1**	<UPL>	**R2>**		
						<N>			

4-35: Annotation of two-word repetitions - FR038-F

I	mean	**when**	**when**	**you**	(0.260)	**when**	**you**	look
<DM	DM>	**<R0**	**<R0+R1>**	**R0**	<UPA>	**R1**	**R1>**	
					<N>			

at	**the**	(0.790)	**at**	**the**	landscape
<R0	<L>+**R0**	<UPL>	**R1**	**R1>**	
		<N>			

4-36: Annotation of repetitions with nested pauses - FR015-F

you	**you**	go	there	because	**you**	(0.560)	**you**	think	that
<R0	**R1>**				**<R0**	<UPL>	**R1>**		
						<N>			

there	**are**	**there**	**are**	lots	**of**	erm	(0.970)	**of**	language
<R0	**R0**	**R1**	**R1>**		**<R0**	<FP>	<UPL>	**R1>**	
						<N>	<N>		

cour	er	lessons
<T>	<FP>	

nous appellerons le *« répété »*. Il va de soi qu'en théorie toute unité produite par la parole est en principe un *répétable* et ce n'est que la présence d'un *répété* immédiatement après qui fait que ce répétable va entrer effectivement dans la composition d'un bloc que nous appelons *a posteriori* une « répétition »".

Repetitions with a rhetorical purpose like emphasiser, for example, are also annotated. Additional remarks, such as the tag `emphasis`, may be added in tier anno-3 (Examples 4-37 and 4-38).

4-37: Annotation of emphasising repetitions (1) - FR035-F

the	level	of	eh	the	pupils	was	**really**	**really**	bad
			<FP>				**<R0**	**R1>**	
							emphasis		

4-38: Annotation of emphasising repetitions (2) - FR018-S

oh	like	right	up				
				yes	**yes**	**yes**	**yes**
				<R0	**R1**	**R2**	**R3>**
				emphasis			

Cases where the close co-occurrence of an identical unit is due to language constraints and rules (examples 4-39 and 4-40) and cases of anadiplosis (examples 4-41 to 4-43), i.e. when a word that is used at the end of a discourse unit is then used again at the beginning of the next one, are <u>not</u> annotated as repetitions.

4-39: Counter-example of repetititon: co-occurrence (1) - FR002-S

I	thought	**of**	two	kind	**of**	(0.510)	two	kinds	**of**	experiences
						<UPL>	<RS	RS	RS>	
								<SM>		

4-40: Counter-example of repetititon: co-occurrence (2) - FR001-S

a	reflection	was	carried	out	**on**	death	**on**	[…]	sickness

on	marriage	as	an	institution	or	**on**	marriage	as	love

4-41: Counter-example of repetititon: anadiplosis (1) - FR002-F

it	’s	a	tradition	who	goes	back	to	one	great	**saint**	**saint**	er
												<FP>

it	’s	a	woman	er	Sainte	er						
				<FP>	<RS>+<W>	<FP>						

4-42: Counter-example of repetititon: anadiplosis (2) - FR015-F

it [the punching ball] makes erm a (0.300) a loud (0.230) a loud noise it's

you	can	't	sleep	with	**it**	**it**	's	every	fifteen	seconds	you	hear	bang
							<FS>						

4-43: Counter-example of repetititon: anadiplosis (3) - FR010-F

if	it	if	**it**	**hits**	**me**	**it**	**hits**	**me**	if	**it**	**doesn**	**'t**	**it**	**doesn**	**'t**
<R0	R0	R1	R1>												

The phenomenon of repetition is restricted in this study to complete and fully propositional lexical words: successive truncations, filled pauses or unfilled pauses are thus not eligible for repetition (Example 4-44).

4-44: Counter-example of repetititon: successive filled pauses - FR042-F

corpus	linguistics	[…]	and	**eh**	**eh**	**eh**	**eh**	try	to
			<C>	<FP>	<FP>	<FP>	<FP>		

H. Restarts

Definition: Unlike repetitions where the repeatable and the repeated are formally identical, restarts are repetition where the propositional content of the repeated is modified in some way, for example by a substitution, a deletion, or an insertion.

Annotation: The restart is marked by RS and can involve one or several item(s). It is always further specified in [anno-2] (see Section 4.2.3.2 for more details). When the restart is used conjointly with a truncation, it is not specified in [anno-2] (as illustrated in Example 4-48), unless it also involves a substitution, a deletion or an insertion (as in 4-49).

4-45: Annotation of restarts (with morphosyntactic substitution) - FR002-S

I	had	to	took	eh	**to**	**take**	er	the	train
				<FP>	**<RS**	**RS>**	<FP>		
						<SM>			

4-46: Annotation of restarts (with insertion) - FR033-F

he	just	(0.160)	showed	it	to	friends	**personal**	**friends**
		<UPL>					**<RS**	**RS>**
							<I>	

4-47: Annotation of restarts (with deletion) - FR005-S

I	went	to	the	**to**	movies	with	er
			<L>	**<RS>**			<FP>
				<Del>			

4-48: Annotation of restarts with truncation (1) - FR038-F

I	've	al	(0.230)	**I**	**'ve**	**already**	seen	him
		<T	<UPA>	**<RS**	**RS**	**RS>**+T>		
			<N>					

4-49: Annotation of restarts with truncation (2) - FR021-S

for	some	holiday	(0.240)	in	En	**to**	**England**
			<UPA>		<T	**<RS**	**RS>**+T>
						<SP>	

I. Discourse markers

Definition: Discourse markers are linguistic elements that are independent of the sentential structure (i.e. they are syntactically optional). They may occur at the beginning, middle, or end of a discourse unit or form a unit of their own. They have little or no semantic meaning in themselves, but are multifunctional in marking the interactional aspect between the participants (Müller 2005; Schiffrin 1987; Schourup 1999).

Annotation: Discourse markers are marked by the tag DM. They can involve one word (<DM>) or two or more word units (<DM DM>). They can be repeated or truncated (see 4-52 and 4-53).

4-50: Annotation of one-word discourse markers - FR002-F

you	are	(0.790)	**well**	you	hear	music
		<UPL>	**<DM>**	<RS	RS>	
					<SP>	

4-51: Annotation of two-word discourse markers - FR016-F

you	can	meet	a	lot	of	people	**I**	**mean**
							<DM	**DM>**

not	only	(0.390)	English	speaking	people	but	also	Spanish
		<UPL>						

4-52: Annotation of repeated discourse markers – FR008-F

well	but	(0.270)	**well**	I	said	**well**	(0.140)
<DM>	<C>	<UPL>	**<DM>**			**<DM>**+<R0	<UPL>
		<S>					<N>+<S>
						reported speech	

well	of	course	I	I	want
<DM>+R1>			<R0	R1>	

4-53: Annotation of truncated discourse markers - FR046-F

it	’s	okay	<laughing/>	so	y	**you**	**know**	there	are
				<C>	<T	**<DM**+T>	**DM>**		

J. Conjunctions

Definition: This category includes three conjunctions of coordination, namely *and, so* and *but.*

Annotation: Conjunctions are marked by the letter C. Conjunctions always involve one word: they will thus always be tagged <C>, but they may be truncated, repeated, or lengthened.

4-54: Annotation of conjunctions (1) - FR017-F

nine	hours	for	the	agregation	(1.020)	**but**	n	next	year
			<L>		<UPL>	**<C>**	<T	T>	

4-55: Annotation of conjunctions (2) - FR017-F

it	’s	my	weakness	**and**	erm	(1.330)	that	’s	the	problem
				<C>	<FP>	<UPL>				

4.2.3.2 Second level of (dis)fluency annotation ([anno-2])

The second tier of (dis)fluency annotation contains tags that specify more precisely some of the features annotated in [anno-1]. More specifically, nesting of (dis)fluency features as well as five characteristics of restarts are annotated in [anno-2]. The same technical aspects apply for the tags in [anno-2], i.e. bracketing, + sign in case of multiple annotation, word-level annotation etc.

A. Lexical insertion

Restarts can first involve the insertion of new lexical elements, e.g. *young* in Example 4-56. Inserted items are marked by the tag <I> in [anno-2], or <I […] I> if the insertion involves more than one word, as shown in 4-56 and 4-57, respectively.

4-56: Annotation of restarts with one-word insertion - FR002-S

who	search	for	er	(0.490)	for	girls	for	**young**	girls
		<R0	<FP>	<UPL>	R1>		<RS	RS	RS>
			<N>	<N>				**<I>**	

4-57: Annotation of restarts with two-word insertion - FR002-F

eh	during	the	war	the	**second**	**world**	war
				<RS	RS	RS	RS>
					<I	**I>**	

B. Morpho-syntactic substitution

Restarts can also include a morpho-syntactic substitution (i.e. a change in number, gender, or tense). For example, in 4-58, *two kind* (not annotated) is substituted by *two kinds*. The substituted elements are then marked by <SM> or <SM SM>.

4-58: Annotation of restarts with morpho-syntactic substitution (1) - FR002-S

I	thought	of	two	kind	of	(0.510)	two	**kinds**	of	experiences
						<UPL>	<RS	RS	RS>	
								<SM>		

4-59: Annotation of restarts with morpho-syntactic substitution (2) - FR005-F

we	are	(0.340)	organising	(0.280)	twice	a	week	er
		<UPA>		<UPA>				<FP>

conversation	er	(0.320)	**conversations**					
	<FP>	<UPL>	<RS>					
			<SM>					

C. Propositional substitution

Propositional substitutions, which involve the replacement of one word by another from the same grammatical category, may also be included in the restart. For instance, a verb may be substituted for another (as in 4-60 and 4-61), an indefinite determiner may be replaced by a definite determiner, or a preposition may be replaced by another. The substituted words are marked by <SP> or <SP SP>.

4-60: Annotation of restarts with propositional substitution (1) - FR005-F

we	hadn	't	eh	we	**weren**	't	very	successful
			<FP>	<RS	RS	RS>		
					<SP>			

4-61: Annotation of restarts with propositional substitution (2) - FR002-S

we	were	by	(0.440)	we	**went**	by	car
			<UPL>	<RS	RS	RS>	
					<SP>		

D. Deletion

Some words might be deleted in the restart (as compared to the repeatable). The tag <Del> is placed just before the place where the word(s) should have occurred, had it (they) not been omitted. Due to technical constraints[121], the tag marks the place where a deletion occurs and not the actual deleted word(s).

4-62: Annotation of restarts with deletion (1) - FR008-S

I	went	to	the	(0.270)	to	Mexico
			<L>	<UPL>	<RS>	
					<Del>	

[121] EXMARaLDA does not make it possible to annotate in [anno-2] if there is no annotation in [anno-1]. Compare Examples 4-62 and 4-63: in the former case, the deleted word (*the*) could have been annotated in [anno-2] because the word is annotated for some other (dis)fluency feature in [anno-1]. In 4-63, however, the deleted words (*a bit*) are not annotated in [anno-1]: EXMARaLDA would not have allowed for an annotation in [anno-2] under those words. Although marking the *place* where a deletion occurs could be seen as a drawback in the annotation, it makes it easier to see the relationship between the restart and the deletion.

4-63: Annotation of restarts with deletion (2) - FR002-S

I	was	a	bit	er	(1.050)	I	was	it	was	a	bit	odd
				<FP>	<UPL>	<RS	RS>	<RS	RS	RS	RS>	
					<P>		**<Del>**	<SP>				

E. Word order

The last characteristic of restarts that is annotated in [anno-2] is a change in word order: the same words are used, but in a different order (as compared to the repeatable). They are marked by the tag <Or Or>.

4-64: Annotation of restarts with word ordering - FR022-F

tennis	table	ta	**table**	**tennis**	sorry
		<T	<RS+T>	RS>	
			<Or	**Or>**	
					<ET>

F. Nesting

(Dis)fluency features are sometimes embedded, or "nested", within another feature. For example, in 4-65, the first filled pause er is uttered between the two parts of the repetition (kings), and in 4-66, both a filled and an unfilled pause are embedded within the repetition of to.

Nested (dis)fluency features are marked in [anno-2] with the tag <N> or <N N>. All features that are annotated in [anno-1] can potentially be nested within other features.

4-65: Annotation of nesting (1) - FR002-F

kings	**er**	kings	(0.220)	er	Charles	the	fifth
<R0	<FP>	R1>	<UPA>	<FP>			
	<N>						

4-66: Annotation of nesting (2) - FR002-S

we	wanted	to	go	to	**er**	**(0.550)**	to	a	beach
				<R0	<FP>	<UPL>	R1>		
					<N>	**<N>**			

4.2.3.3 Third level of (dis)fluency annotation ([anno-3])

Annotations in [anno-3] are freer than those in [anno-1] and [anno-2]: the third level of annotation mainly contains occasional remarks or notes about the discourse in the transcription, such as:

- the presence of editing terms, such as *sorry* (marked by `<ET>` or `<ET ET>`, see Example 4-64);
- grammatical mistakes (marked by `error`);
- an erroneous pronunciation (marked by `mispronunciation`; see Example 4-67);
- emphasis (for repetitions; marked by `emphasis` – see 4-37);
- reported speech (marked by `reported speech`, as in 4-52);
- …

Annotations in this tier do not have to strictly follow the technical guidelines (especially the bracketing system) and are more specifically intended for later interpretation and analysis.

4-67: Annotation of mispronunciation - FR032-S

where	people	were	ver	were	very	grateful
			<T	<RS	RS>+T>	
		mispronunciation				

4.2.3.4 Search syntax for the concordancing and extraction in EXAKT

EXAKT, the search tool in EXMARaLDA, makes it possible to look for linguistic information for each tier independently. The tool functions with regular expressions[122]. Note that extraction may also be performed by other means, such as using scripts.

The search syntax presented in Table 4-10 can be used in EXAKT to look for specific words or character strings in the **transcription tiers**.

[122] Some documentation on the use of regular expressions is provided on the official website of EXMARaLDA (http://www.exmaralda.org/pdf/Quickstart_Regular_Expressions_EN.pdf; last accessed 11/03/2018); see in particular the use and meaning of symbols and metasymbols.

Search	Search syntax
Paraverbal information	\<\b\p{L}+/\>
word	word
word + words	word.
be + being	be(|ing)
Length of unfilled pauses (UPA and UPL)	\(\d{1,2}\.\d{1,3}\)

Table 4-10: Search syntax (transcription tiers)

Likewise, to compute the **concordances of (dis)fluency features** marked in [anno-1] and [anno-2], the search syntax presented in Table 4-11 can be used in EXAKT.

Search	Search syntax
Filled pauses (FP)	<FP>
Unfilled pauses (UPL + UPA)	<UP
Unfilled pauses from LINDSEI-FR and LOCNEC transcriptions (UPL, *some may be shorter than 200 ms)*	<UPL>
Unfilled pauses added to LINDSEI-FR and LOCNEC transcriptions (UPA, *equal to or longer than 200 ms)*	<UPA>
Vowel lengthening	<L>
Vowel lengthening used conjointly with another (dis)fluency feature	((|\+)<L>(|\+))
Truncated words	<T
Abandoned truncated words	<T>
Completed truncated words	<T(\+|$)
Completed truncated words with restart	RS>\+T>
Truncated words used conjointly with another (dis)fluency feature	((|\+)(|<)T(|>)(|\+))
Foreign words	<W
Foreign words used conjointly with another (dis)fluency feature	\+(|<)W
False starts	<FS>
False starts used conjointly with another fluenceme	((|\+)(|<)FS(|>)(|\+))
Repetitions	<R0

Double repetitions (R0 R1)	R1>
Triple repetitions (R0 R1 R2)	R2>
Repetitions used conjointly with another (dis)fluency feature	((|\+)(|<)R\d(|>)(|\+))
Restarts	<RS
Restarts used conjointly with a truncation	RS(|>)\+T>
Restarts used conjointly with another (dis)fluency feature	((|\+)(|<)RS(|>)(|\+))
Insertions	<I
Morpho-syntactic substitutions	<SM
Propositional substitutions	<SP
Substitutions	(<SM|<SP)
Deletions	<Del
Change of order	<Or
Discourse markers	<DM
One-word discourse markers	<DM>
Two-word discourse markers	DM>
Discourse markers used conjointly with another (dis)fluency feature	((|\+)(|<)DM(|>)(|\+))
Conjunctions	<C>
Nested elements	<N

Table 4-11: Search syntax (annotation tiers)

4.2.3.5 The ARC annotation scheme

One of the aims of the ARC project this thesis is part of was the comparison of (dis)fluency features across languages (French vs. English) and modalities (spoken and sign languages). To this end, a generic annotation scheme was also collaboratively designed by the four doctoral researchers of the project to ensure a minimum level of comparability between our respective results. However, given that each sub-project within the ARC had its own research objectives and agenda, the ARC scheme was either further adapted to answer specific research questions, or designed to be compatible with a previously developped annotation scheme. In the present case, I designed the annotation scheme presented above so that, with minor adaptations, the annotations can be converted into 'ARC annotations'.

In the interest of space, I summarise the main points of divergence in Table 4-12 below. For further details on the ARC annotation scheme, see Crible *et al*. (2015a) and for a first attempt at cross-modal comparison of (dis)fluency features, see Crible *et al*. (2017).

	Dumont (2015)	Crible *et al*. (2015a)
Global architecture	• 3 annotation tiers • 2 tiers with part-of-speech tags	• 1 main annotation tier • 1 tier for diacritics (i.e. misarticulation, lengthening, embedding, word order, completion) • no part-of-speech tags
Terminology	• "(Dis)fluency features"	• "(Dis)fluency markers"; "fluencemes" • Distinction between "simple" and "compound" fluencemes
	• "Repetition"	• "Identical repetition"
	• "Restart"	• "Modified repetition"
Scope of (dis)fluency features	• Distinction between discourse markers (DM) and coordinating conjunctions (C)	• Discourse markers include coordinating and subordinating conjunctions, as well as interjections
	• No distinction between lexical and parenthetical insertions	• Distinction between lexical and parenthetical insertions
	• Syntactic completion: not annotated	• Syntactic completion: annotated
Annotation format & tags	• Angle brackets	• Angle brackets
	• '+' used for multiple annotations of the same item • The tags are ordered alphabetically	• The tags are juxtaposed (without blank space), first the simple, then the compound fluencemes
	• Restarts: only the restart is annotated (i.e. not the original utterance)	• Modified repetitions: have a 2-part structure (like repetitions), i.e. the original utterance and the modification. Their annotation also uses a numbering system

Table 4-12: Comparison between Dumont (2015) and Crible et al. (2015)

4.2.4 Implementation and evaluation of the annotations

4.2.4.1 Implementation

Following Spooren and Degand's (2010) guidelines, the development of the annotation scheme was coupled with a "**warming-up phase**" and a "**calibration phase**" during which I tested the annotation system on a small sample from LINDSEI-FR+. This enabled me not only to get acquainted with the categories to be annotated, but also to adapt and improve the annotation scheme to account for unforeseen phenomena or issues. Only after these two phases were LINDSEI-FR+ and LOCNEC+ interviews fully annotated within EXMARaLDA.

The annotation of LINDSEI-FR+ and LOCNEC+ was performed in the Partitur-Editor tool of EXMARaLDA and combined the two main annotation methods: while some (dis)fluency features were **automatically** marked (though with manual disambiguation), others were fully **manually** annotated. An overview of the annotation method of the various (dis)fluency features can be found in Table 4-13 below. As explained previously (see Section 2.3.3, and more particularly Section 2.3.3.2), automatic annotations have the advantage of being quick and of alleviating the cognitive load during the annotation process, but they may also cause 'noise': complementary manual checks are thus required. Manual annotations are cognitively demanding, especially when numerous features are annotated at the same time. They are also prone to formal errors such as the misspelling of a tag or the oversight of a closing bracket, but they are also more accurate when it comes to the annotation of complex patterns.

The mark-up indicating filled and unfilled pauses, vowel lengthenings, truncations and foreign words in the original LINDSEI-FR and LOCNEC transcriptions was **automatically** converted into the corresponding stand-off (dis)fluency tag and integrated in [anno-1]. Nevertheless, each of the automatically generated tags was manually checked to weed out errors or to correct the potential erroneous annotation of complex patterns. Conjunctions and repetitions, which were not annotated in LINDSEI-FR and LOCNEC, were also partially automatically detected and annotated thanks to a handmade script[123]. It should nonetheless be noted that the complexity of the patterns involving repetitions limited the efficiency and accuracy of this automatic annotation.

False starts, restarts and their sub-categories, discourse markers as well as nesting were annotated strictly **manually**. For this type of annotation, an annotation panel in the Partitur-Editor can be used to support the consistent application of the tagset (the so-called 'annotation specification' that I created for this purpose can be found in Appendix 9.5). Although the annotation panel aims at facilitating manual annotations, it proved not to be ideally adapted to the (dis)fluency annotation system as designed in this study, with its

[123] The script was written by Sophie Roekhaut.

brackets, plus signs, different levels etc. As a result, except for the first few interviews, I simply manually typed the tags on the keyboard for each annotation.

(Dis)fluency features	Automatic annotation	Manual annotation / disambiguation
Conjunction	Yes	Yes
Discourse marker	No	Yes
Filled pause	Yes	Yes
False start	No	Yes
Foreign word	Yes	Yes
Lengthening	Yes	Yes
Repetition	Yes	Yes
Restart	No	Yes
Deletion	No	Yes
Insertion	No	Yes
Lexico-grammatical substitution	No	Yes
Propositional substitution	No	Yes
Word order	No	Yes
Truncation	Yes	Yes
Unfilled pause	Yes	Yes
Nesting	No	Yes

Table 4-13: Automatic and manual annotation of (dis)fluency features

I worked **speaker after speaker** (and not, for example, task after task) in order to get used to each speaker's specific speaking style and to be able to detect his or her potential idiosyncratic (dis)fluency patterns more easily. I also took some **notes** on the annotation process of each interview (e.g. how I annotated a complex pattern) and briefly summarised my impression of each speaker's (dis)fluency for future reference.

Because the annotation phase and time alignment corrections are intimately connected (especially in the case of unfilled pauses), I proceeded in two stages. I first **annotated** each raw aligned file, checking the automatically generated tags and annotating the other (dis)fluency features. I also listened to the audio file to disambiguate problematic cases if necessary (if the time alignment was not accurate, I came back to the tag after correcting the time alignment). Then, I went over the annotated file a second time and mainly focused on the accuracy of the **time alignment**: I paid particular attention to the alignment (and measurement) of unfilled pauses, as well as the transcription of overlaps and unclear passages (the <unknown /> tags in the transcriptions). While doing so, I also **verified the annotations** and corrected potential slips, or added tags that I had missed.

On average, **two to four hours** were required to annotate and correct the alignment of **5 minutes of speech** in LINDSEI-FR+ and LOCNEC+ – most of the time was devoted to the correction of the time alignment – but this average estimation hides wide discrepancies: while some interviews could be annotated in half a day, others (fortunately not the majority) required up to three days.

With respect to the number of annotators, while some brief passages of LINDSEI-FR were annotated by the other ARC PhD students too, and diverging annotations discussed, the overwhelming majority of the interviews were only annotated by myself, i.e. a "one-coder-does-all solution" (Spooren & Degand 2010:254). It is legitimate to question whether **one single annotator** can remain coherent in his/her annotations (e.g. can refrain from being more sensitive to particular aspects of (dis)fluency), especially considering the fact that mental fatigue also undoubtedly influences the quality of the annotations. However, as Spooren and Degand (2010:254) write, "[o]f course the coding will be subject to individual strategies developed by the coder, but these strategies will presumably be systematic and there is no reason to assume that such strategies will be conflated with the phenomena of interest". In other words, while it is not impossible that I am more sensitive to particular features over others, the annotation of these features will presumably be systematic. Besides, the fact that many elements could be annotated at least partially automatically, and that I developed the annotation system in a data-driven fashion offer additional safety belts. More generally, I felt that annotating in itself also helped in raising my awareness of all (dis)fluency features and that my overall perception of disfluencies became much more acute over time (and not only when I was annotating!). With regard to the issue of mental tiredness, it was hopefully circumvented by the careful planning of the annotation phase, which expanded over several months.

The interviews were only **annotated once**. Because annotating is very time-consuming, it was unfortunately not possible to ask a second annotator to annotate the corpora, or to re-annotate the two corpora myself. However, as explained above, I went over the annotations twice and I also checked intra-annotator reliability, the results of which are set out below. Before moving on to this analysis, however, some reflections on the annotation process are presented.

4.2.4.2 Retrospective thoughts on the annotation process

Annotating LINDSEI-FR+ and LOCNEC+ was a long enterprise, and despite the partial automation of the process, the cognitive strain – due to the manifold simultaneous annotations – did feel quite heavy at times. Careful planning and scheduling, I believe, was key in maintaining a high level of attentiveness through the files. As for the tool I used, although the EXMARaLDA interface is user-friendly, some important functions are not very intuitive, and it was often necessary to go back several times to the user manuals available online in order to prevent unwelcome surprises at later stages. I only tested other tools (such

as Praat and ELAN) occasionally, but I am convinced they could also be used for similar annotation tasks. In any case, the annotation protocol does not have to be used within EXMARaLDA.

Lastly, as regards the (dis)fluency features, in some cases, I noticed that the borderline between the categories of restarts and of false starts was a bit blurry. Other researchers have also reported this difficulty in distinguishing false starts from restarts (e.g. Gráf 2015:38). Likewise, the distinction between conjunctions and discourse markers was at times confusing. In this regard, it might be important to underline that during the early stages of the annotation, I considered *and, so* and *but* eligible to be tagged as discourse markers as long as they matched the definition. However, as there were many cases where I found it extremely difficult to discriminate between the two categories with a fair degree of certainty, I chose to annotate all occurrences of *and, so* and *but* as conjunctions, thereby leaving further considerations as to their exact status for a later stage in the analysis.

The next section aims to examine the degree of intra-rater reliability of the annotations in LINDSEI-FR+ and LOCNEC+.

4.2.4.3 Intra-annotator reliability

The assessment of inter-rater reliability (also called inter-rater agreement) provides "a way of quantifying the degree of agreement between two or more coders who make independent ratings about the features of a set of subjects" (Hallgren 2012:23). When the data are coded twice by the same coder, the term "intra-rater reliability" is generally adopted. As underlined by Artstein and Poesio (2008:557), reliability is "a prerequisite for demonstrating the validity of the coding scheme". For an overview of the methodological issues related to the assessment of inter-rater reliability, see e.g. Hallgren (2012); for a survey of methods for measuring agreement among corpus annotators, see e.g. Artstein and Poesio (2008).

To measure the reliability of the (dis)fluency annotations, **I re-annotated 18 speaking tasks** ten months after finishing the annotation. The sample consisted in 3 set topics, 3 free discussions and 3 picture descriptions in each corpus[124]. The tasks were **picked randomly** in each corpus, and represent 6% of the corpus files, and **over 5,000 annotations** in [anno-1].

When using categorical data, **Cohen's kappa** (κ) can be used to measure the extent of the agreement between judges (Howell 2013:166). The κ statistic may range from 0 to 1. Landis and Koch (1977:165) suggested useful benchmarks for the **interpretation** of this statistic, with scores ranging from 0 to 0.20 representing a slight level of agreement, from 0.21 to 0.40 a fair

[124] The set topic tasks are the following: FR012-S, FR023-S, FR040-S, EN002-S, EN018-S and EN031-S. The free discussion tasks are: FR011-F, FR016-F, FR034-F, EN010-F, EN025-F, and EN040-F. The picture description tasks are: FR007-P, FR020-P, FR048-P, EN020-P, EN047-P and EN050-P.

level of agreement, from 0.41 to 0.60 a moderate level of agreement, from 0.61 to 0.80 a substantial level of agreement and over 0.81 an almost perfect level of agreement.

Cohen's kappa was run on the data to determine the level of agreement between the first and the second round of annotations in [anno-1]. The κ statistic reaches **.945** ($p < .000$), which, according to Landis and Koch's (1977) guidelines, represents an **almost perfect agreement**. Ensuring a high level of reliability was a major consideration while annotating, and much effort was put into manually checking the annotations: it appears that the methodology adopted admirably succeeded in this respect.

Looking further at the **individual (dis)fluency features**, it appears that not only was the overwhelming majority of the annotated features re-annotated in the second round, but they were also re-annotated in the same way (i.e. with the same tag).

(Dis)fluency annotation	Number of annotations in round 1	Number of annotations in round 2	% of identical annotations
Conjunction	650	640	98.5%
Discourse marker	590	542	91.9%
False start	78	55	70.5%
Filled pause	639	633	99.1%
Foreign word	23	21	91.3%
Lengthening	261	261	100%
Repetition	902	808	89.6%
Restart	524	473	90.3%
Truncation	187	186	99.5%
Unfilled pause	1,301	1,290	99.2%
No annotation	*9,353*	*9,151*	*97.8%*
Totals	**14,508**	**14,060**	**96.9%**

Table 4-14: Annotation of individual (dis)fluency features

As displayed in Table 4-14, the accuracy rate for the identification and annotation of five (dis)fluency features – namely **conjunctions, filled pauses, vowel lengthenings, truncations, unfilled pauses** – is **nearly perfect** (> 98%). Likewise, the identification of words that did not need any annotation is also very high: 97.8% of the words that had not been annotated in the first round remained unannotated in the second. For **discourse markers, repetitions, restarts and foreign words**, the percentage of identical re-annotations proves to be slightly lower, but is still **very high** (c. 90%). Looking more closely at the crosstabulated data, it appears that some discourse markers were actually either not annotated at all or re-

annotated as conjunctions[125] in the second round. As for repetitions, restarts and foreign words that were annotated in the first round but not in the second, the examination of the data revealed that I either missed them during the re-annotation, or interpreted them differently (for example, by re-annotating a repetition as a restart).

False starts have the lowest accuracy rate: **70.5%** of the false starts from the first round of annotations were re-annotated as such in the second. Two explanatory factors may be at play here. As previously underlined, a number of disagreements are due to difficulties in drawing the line between the category of false starts and that of restarts. More specifically, it seems that the presence of pauses or prosodic aspects also influenced my analysis. For example, in 4-68, I first annotated *I thought* as a false start because this segment is left grammatically incomplete. Moreover, it is followed by quite a long pause and the intonational pattern also shows a clear separation between *I thought* and *I'd expected*, which follows the pause. In the re-annotated data, however, due to the repeat of the personal pronoun *I, I'd expected* was instead interpreted as the restart of *I thought*. Incidentally, this type of coding disagreement suggests that further analyses into the prosodic context of false starts could greatly deepen our understanding of this phenomenon, which, for now, may still remain "intrinsically ambiguous" (Spooren & Degand 2010:253). The second factor that may explain the lower accuracy rate of false starts has to do with interruptions by the interviewer: in a few exceptional cases, I seemed to use the tag for false starts to indicate that the speaker was interrupted by the interviewer and could not complete his or her utterance (*cf*. Example 4-69). These cases should, however, be considered within the frame of interactive (dis)fluency, which is out of the scope of the present thesis. Despite the slightly lower level of agreement for false starts, there is no doubt that they are still an interesting and worthy category to investigate "as long as we recognise the limitations of a scheme which delivers less than ideal levels of reliability, and use the resulting annotated [data] accordingly" (Craggs & Wood 2005:293).

4-68: FS/RS disagreement - EN010-F

	I	thought	(0.430)	I	'd	expected	it	to	be	fantastic
First annotation		<FS>	<UP>							
Re-annotation			<UP>	<RS	RS	RS>				

4-69: FS disagreement - FR011-F

A	well	a	bit	far	to	do	every	evening	for	example
FR011		in	the	mountain	it	's				

[125] *Cf. supra*. Although I did correct my early annotations of *and, so* and *but* as discourse markers, some seem to have escaped my attention.

First annotation						<FS>				
Re-annotation										

To sum up, the (dis)fluency annotations offer a more than satisfactory level of reliability for further analyses. All annotation disagreement cases were re-examined one by one in the sample of 18 speaking tasks, and coding errors corrected.

The last section of this chapter provides a quantitative overview of the (dis)fluency features that will be explored in the following chapters.

4.2.5 A glimpse into the annotated (dis)fluency variables

The (dis)fluency annotations were extracted using the search syntax shown in Table 4-10 and in Table 4-11. I screened the annotations of each (dis)fluency feature once again, with particular attention to the (dis)fluency features that had a slightly lower level of reliability. Table 4-15 below reveals the final breakdown of each (dis)fluency feature in LINDSEI-FR+ and in LOCNEC+.

Annotated (dis)fluency features	**LINDSEI-FR+**		**LOCNEC+**		**Total**
Conjunctions	4,849	12.95%	6,522	22.99%	11,371
Discourse markers	2,016	5.39%	3,213	11.27%	5,229
False starts	656	1.75%	628	2.20%	1,284
Filled pauses	7,576	20.23%	2,997	10.52%	10,573
Foreign words	436	1.17 %	56	0.20%	492
Lengthenings	2,909	7.77%	1,026	3.60%	3,935
Repetitions	3,748	10.01%	2,750	9.65%	6,498
Restarts	1,775	4.74%	1,651	5.79%	3,426
Truncations	1,614	4.31%	837	2.94%	2,451
Unfilled pauses	11,863	31.68%	8,818	30.94%	20,681
Totals	**37,442**	**100%**	**28,498**	**100%**	**65,940**

Table 4-15: Raw number of occurrences of (dis)fluency features in LINDSEI-FR+ and LOCNEC+
Note: the (dis)fluency features are ordered in alphabetical order

Table 4-15 shows that the first level of annotation in **LINDSEI-FR+**, which contains the 10 generic tags, includes precisely 37,442 (dis)fluency features. Great differences can be observed between the different categories, ranging from 11,863 occurrences (31.68%) for unfilled pauses to 436 (1.17%) for foreign words. Likewise, in **LOCNEC+**, among the 28,498 features annotated, 30.94% are unfilled pauses, and a small percentage accounts for truncations, false starts or foreign words. Figure 4-13 and Figure 4-14 provide a visual representation of the cumulative frequencies of (dis)fluency features in LINDSEI-FR+ and LOCNEC+, respectively. It is striking that in the learner corpus, filled and unfilled pauses account for slightly more than 50% of the annotations. In the native corpus, the cumulative percentage is slightly lower, but still impressive (c. 40%). This means that as much as one out of two annotations in LINDSEI-FR+ and that two out of five annotations in LOCNEC+ is a pause (filled or unfilled).

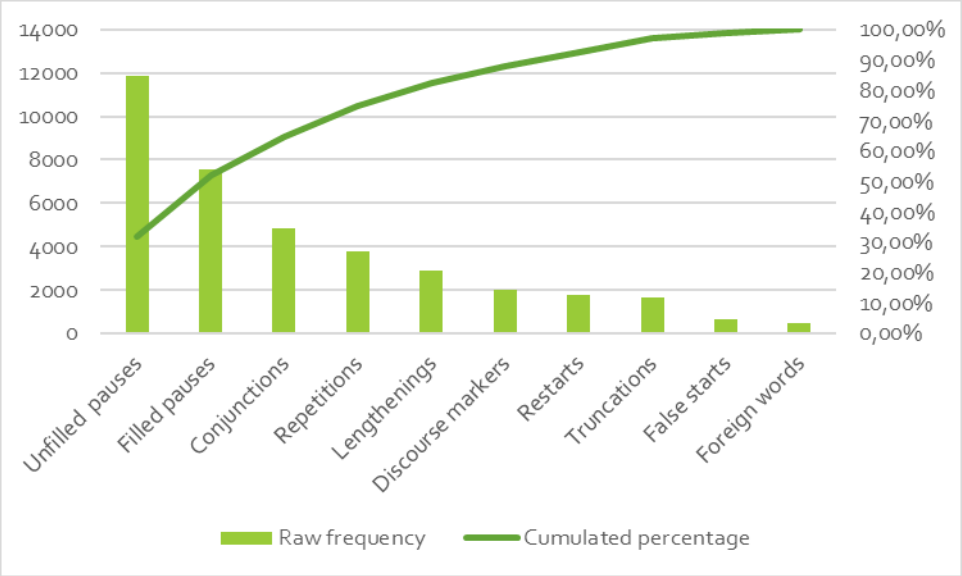


Figure 4-13: Frequencies and cumulated percentages of (dis)fluency features in LINDSEI-FR+
Note: the (dis)fluency features are ordered in decreasing order of frequency

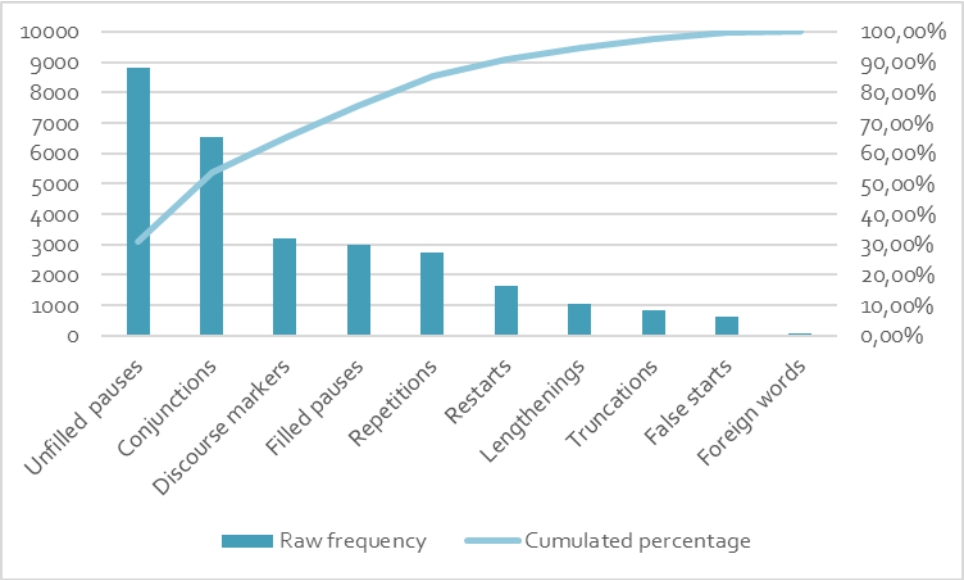


Figure 4-14: Frequencies and cumulated percentages of (dis)fluency features in LOCNEC+
Note: the (dis)fluency features are ordered in decreasing order of frequency

The figures presented above, as well as the temporal data and word counts set out in Section 4.1.8, are used to measure the normalised frequency of each (dis)fluency feature per corpus. The analysis of these features is presented in the following chapter.

4.3 CONCLUSION

This chapter reported on the two major transformations of LINDSEI-FR and LOCNEC, namely time alignment at the word-level (Section 4.1) and (dis)fluency annotation (Section 4.2).

The benefits of time alignment are not to be under-estimated: it not only provides **precise and reliable temporal data** which were previously inaccessible for research, but it also greatly **facilitates the annotation** – and increases its quality – by enabling quick access to the original audio data.

The main drawback of the time alignment procedure, however, is that it is extremely **time-consuming.** Although the time required to generate and manually correct time alignment depends on many elements, the quality of the audio recoding is a key factor. In this respect, I would advise researchers willing to engage in the collection of new spoken corpora to pay particular attention to the range of recording devices available and to the recording set up (the use of one microphone per speaker, the recording place etc.). Besides, for LINDSEI-FR and LOCNEC, time alignement was carried out after the transcription phase. Time and manpower could certainly be saved if both were carried out at the same time.

One one the main difficulties in aligning the corpora pertained to **overlapping speech** and the way overlaps were marked in the transcriptions, and more specifically, the fact that only the beginning of overlapping speech was indicated. Although I acknowledge that transcription, and transcription of overlapping speech in particular, is a very complex task, marking the end of the overlap could be beneficial for the time alignment procedure.

In Section 4.2, I presented the **(dis)fluency annotation protocol**, its main principles, and illustrated the annotation of each (dis)fluency feature. This annotation protocol, which allows for the annotation of simple and more complex features (i.e. involving one or several words, such as repetitions) as well as embedded patterns (e.g. a pause within a repetition), was applied to the time aligned version of the LINDSEI-FR+ and LOCNEC+ interviews.

The annotation scheme was based on an array of studies, iteratively applied to real corpus data, and amended accordingly to reach the final version. In spite of these amendments, the annotation of some (dis)fluency features still proved to be difficult at times, and the **borderline between two categories** was not always straightforward, especially between restarts and false starts. Annotation by **multiple annotators**, including discussion of problematic cases, should be recommended for similar cases (it was unfortunately not possible in this case for practical reasons).

Annotating a dozen (dis)fluency features is a highly **cognitively demanding task**, and manual annotation (like automatic annotation) has its limits. Prior to using the annotated data, it is insightful to carry out an inter-rater reliability analysis: it shows the extent to which

annotations are reliable and it also pinpoints the areas where, perhaps, the researcher should consider revision.

Lastly, despite the intrinsic difficulties and limitations involved in such enterprises, collaborative initiatives attempting to design **cross-linguistic and/or cross-modal annotation** schemes are extremely valuable and promising. Deeper insights could definitely be gained from analyses using such cross-linguistic annotation of (dis)fluency features, for example, by gauging the extent of transfer from the mother tongue to the foreign language, as compared to native speaker behaviour.

PART III

Chapter 5
A QUANTITATIVE SKETCH OF LEARNER AND NATIVE SPEAKER (DIS)FLUENCY

Everybody's tongue slips now and again,
most often when the tongue's owner is tired,
a bit drunk, or rather nervous. So errors of this
type are normal enough to be called normal.

The Articulate Mammal

(Aitchison 1989:244)

This chapter is the first chapter of the third part of this thesis, which presents the corpus findings on learner and native (dis)fluency. It is primarily concerned with the **description and illustration of the 14 (dis)fluency measures in the speech of French-speaking learners and of native speakers**. While the main focus is on the univariate analysis of each (dis)fluency variable separately in the two corpora, the interrelationships between (dis)fluency features are examined in the next chapter (Chapter 6). Chapter 7 then zooms in on the learner data and analyses the relationship between the learners' empirical (dis)fluency measures and their assessed CEFR fluency level.

After a short introduction providing a first overview of the data in the two corpora, Section 5.2 focuses on the temporal (dis)fluency measures and Section 5.3 then scrutinises the annotated (dis)fluency features.

Before embarking on the first section, it is important to underline that, to avoid making this chapter (too) number crunching, only the most important figures are mentioned in each section. A full account of the figures and statistical tests is provided in a summary table (Table 5-17) in Section 5.4.

5.1 INTRODUCTION

The introductory section of this chapter aims to present a cursory overview of the (dis)fluency of French-speaking learners as compared to that of native-speakers. Fourteen (dis)fluency measures are put under scrutiny in this chapter, as summarised in Table 5-1.

Temporal (dis)fluency measures		
1.	Speech rate	SR
2.	Mean length of runs	MLR
3.	Phonation-time ratio	PTR
4.	Mean length of unfilled pauses	MLUP

Annotated (dis)fluency features[126]		
1.	Unfilled pauses	UP
2.	Filled pauses	FP
3.	Conjunctions	C
4.	Repetitions	Rep
5.	Lengthenings	L
6.	Discourse markers	DM
7.	Restarts	RS
8.	Truncations	T
9.	False starts	FS
10.	Foreign words	W

Table 5-1: Overview of the 14 (dis)fluency variables analysed in Chapter 5

One of the major benefits of having access to time aligned data in the frame of a study on (dis)fluency is that it is possible to measure **temporal aspects of (dis)fluency** both accurately and reliably. The temporal data in LINDSEI-FR+ can be summarised by four figures:

- 160 words per minute: the learners' mean speech rate;
- 5.6 words: the mean number of words between two unfilled pauses (i.e. the mean length of runs);
- 83%: the proportion of speech vs. the proportion of pausing time (17%) in a typical LINDSEI-FR+ interview;
- 0.5 second: the average length of an unfilled pause.

In the native speaker corpus, the corresponding figures are the following: 222 words per minute, 8 words per speech run, 87% of speech per interview (13% of pausing), and 0.5 second on average per unfilled pause. Unsurprisingly, the figures for speech rate, length of run and phonation time ratio are slightly higher for the native speakers than for the learners. More surprising, however, is the similarity between mean length of unfilled pauses, but, as will become clearer below, this similarity hides more subtle differences.

With respect to the **annotated (dis)fluency features**, the corpus data reveal that all the learners produce at least one occurrence of each of the 10 annotated features listed in Table

[126] That is, the 10 (dis)fluency features annotated in [anno-1], see Section 4.2.

5-1, except for foreign words: four LINDSEI-FR+ speakers (8% of the learners) do not produce any foreign word in their interview. The native speakers from LOCNEC+ also produce these (dis)fluency features, but 70% of them do not produce foreign words in their interviews, and one speaker (2%) does not produce any false start either.

Estimates of the **overall frequency of (dis)fluency features** in speech greatly vary in the literature and caution is advised when encountering such figures. They are indeed dependent on, among others, the range of features considered, whether or not unfilled pauses are included in the count (as well as the chosen threshold(s)), and whether or not the researcher has adopted a disfluency bias (for example, by counting only "disfluent" repetitions, or only "disfluent" pauses). The type of speech (e.g. monologues or dialogues) and speakers considered (e.g. L1 or L2) are undoubtedly also of prime importance. Reported frequency counts of (dis)fluency features range from 2 to 26 "disfluencies" per 100 words, that is, a difference by a factor of 13! Based on a review of previous literature, Fox Tree (1995:710) reports an average estimate of **6%, exclusive of pauses,** in native speech – it is unfortunately not entirely clear whether this average is exclusive of unfilled pauses, filled pauses, or both. In learner language, Kormos and Dénes (2004:154) found c. 5 disfluencies (operationalised as repetitions, restarts and repairs) *per minute* in the speech of learners of English, plus an additional c. 30 unfilled pauses and 16 or 8 filled pauses (for intermediate and advanced learners, respectively), which, in total, amounts to **between 40 and 50 disfluencies per minute**.

On average in LINDSEI-FR+, each learner produces as many as **39.36 (dis)fluencies every hundred words**. This means that about two fifths of each learner's discourse has been annotated for one of the 10 (dis)fluency features considered in this study. In the native corpus, this proportion is much lower, though still quite high as compared to previous frequency counts: each native interview averages **22.02 (dis)fluency features** per hundred words. Besides, as can be seen from Figure 5-1, although there is some variability in each corpus, there is no overlap between the mean total frequencies in LINDSEI-FR+ and LOCNEC+: the learner who produces the least (dis)fluency features per hundred words still produces slightly more of them than the native speaker who produces the most (29.51 vs. 28.89 phw).

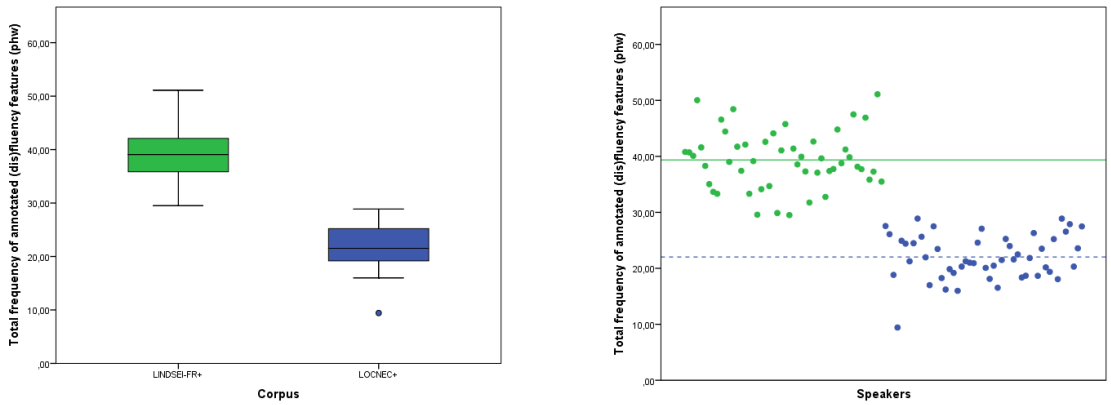


Figure 5-1: Boxplots and stripcharts of the total frequency of annotated (dis)fluency features (phw) in LINDSEI-FR+ and LOCNEC+

An independent samples *t*-test revealed that, all (dis)fluency features taken together, the mean difference between learners in LINDSEI-FR+ and native speakers is **significant** (t = 18.80, p = .000). Moreover, this difference represents a **large-sized effect** (d = 3.76). This suggests that, even at a high level of proficiency, French-speaking learners still produce highly **significantly more (dis)fluency features** per hundred words on average than native speakers in the same type of communicative task.

Zooming in on the frequency of each annotated (dis)fluency feature taken independently, it is obvious that these do not come up in speech with the same frequency, some being far more prevalent than others. Moreover, when native and non-native speaker (dis)fluency features are ranked in decreasing order of frequency (see Figure 5-2 and Figure 5-3), it is interesting to see that there are similarities, but also differences, between learners and native speakers.

In both LINDSEI-FR+ and LOCNEC+, **unfilled pauses** have the **highest mean frequency**, and **truncations, false starts and foreign words** are found at the far right side of the graph (i.e. they have the **lowest mean** frequencies). The six features that have an intermediate frequency, however, come up in a different order in the learner and in the native corpus. For example, filled pauses, which come second in terms of mean frequency in LINDSEI-FR+, appear in fourth position in LOCNEC+, after conjunctions and discourse makers (NNS conjunctions appear in third position, and discourse markers in sixth). These differences seem to suggest that, contrarily to **native speech where coherence and pragmatic features** (conjunctions and discourse markers) seem to be more salient (at least in terms of frequency compared to the other features), in **learner speech**, it is the **two types of pauses** as well as delaying strategies (**repetitions and lengthenings**) that tend to be more pervasive. Besides, statistical analyses reveal significant differences in mean frequency between the learner and the native speaker data for all annotated (dis)fluency features, except for conjunctions and discourse markers (see Figure 5-4 for a visual representation, as well as Table 5-17 for more details on the results of *t*-tests on separate variables).

Bearing in mind that differences may also lie at subtler levels, such as the functional use of the features, all in all, the results presented in this introductory section might be first indicators of a potential different underlying structure of (dis)fluency features between learner and native speaker speech (see Chapter 6).

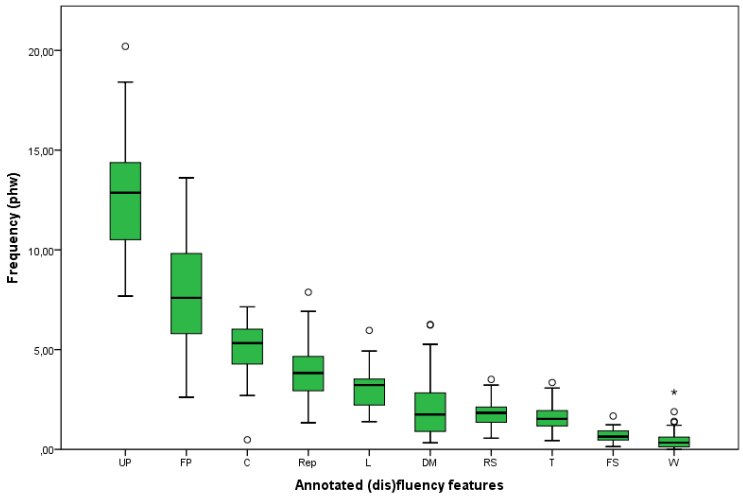


Figure 5-2: Boxplots of the 10 annotated (dis)fluency features in LINDSEI-FR+ (ranked in decreasing order of frequency)

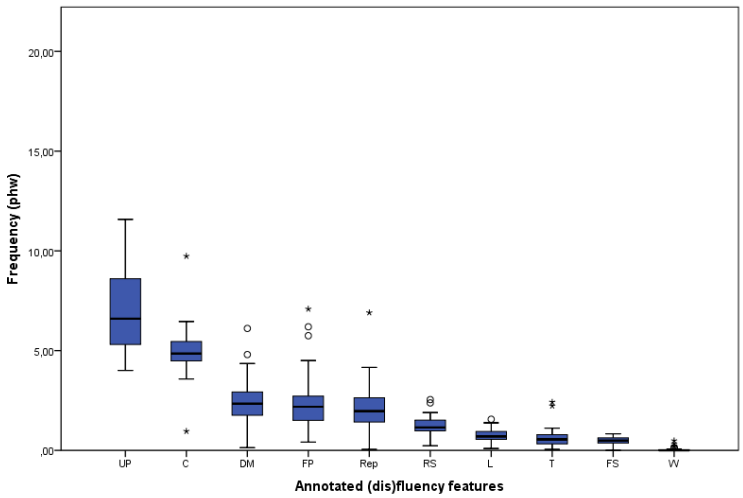


Figure 5-3: Boxplots of the 10 annotated (dis)fluency features in LOCNEC+ (ranked in decreasing order of frequency)

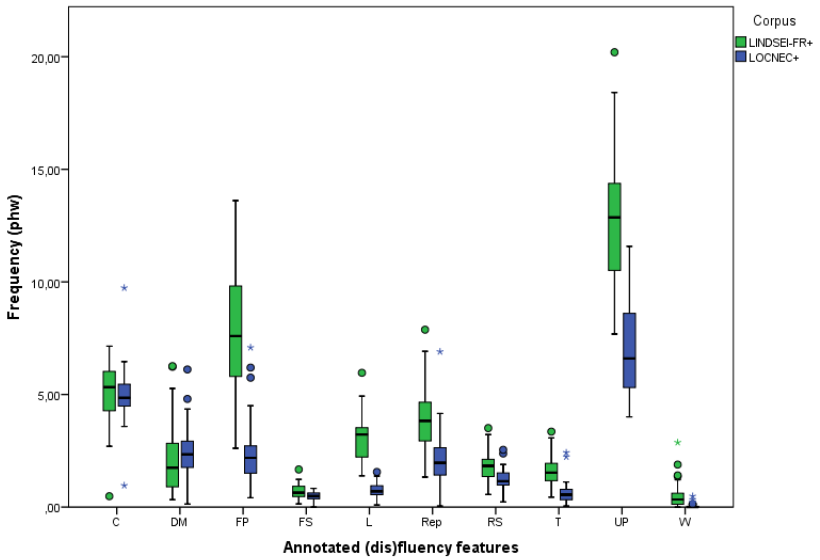


Figure 5-4: Boxplots of the 10 annotated (dis)fluency features in LINDSEI-FR+ and LOCNEC+ (in alphabetical order)

5.2 TEMPORAL (DIS)FLUENCY MEASURES

Four temporal (dis)fluency measures were calculated based on the time aligned data from LINDSEI-FR+ and LOCNEC+, namely speech rate, mean length of runs, phonation-time ratio, and mean length of unfilled pauses. The first three come under close scrutiny in the following sections, but, for reasons of clarity of exposition, the mean length of unfilled pauses is analysed together with the frequency of unfilled pauses in Section 5.3.1.

5.2.1 Speech rate

Following, *inter alia*, Lennon (1990), Götz (2013a) and Gráf (2015), speech rate (SR) is measured in **words per minute** (wpm). It is calculated by dividing the total number of (unpruned) words by the total speaking time (including UP time)*60 (*cf.* previous chapter for details on the calculation of words and time).

The **learners**' mean speech rate in LINDSEI-FR+ amounts to **162.6 words per minute**. As can be seen in Figure 5-5, there is considerable variability in the learner data with SRs ranging from 131.62 to 194.04 wpm (FR028 has a mean SR of 216.48 but is an outlier). The **native speakers'** mean speech rate in LOCNEC+ interviews is substantially higher: **222.13 wpm** on average. The data further reveal that there is some overlap between the learner and native speaker distributions: four native speakers (EN043, EN004, EN006, and EN005) have a lower mean speech rate than the highest non-native speech rate (the outlier excluded). This suggests that some learners in the corpus perform as "well" as their native speaker counterparts.

The French-speaking learners from LINDSEI-FR+, although they are highly proficient, speak **statistically slower** (t = -15.48; p = .000) on average than the native speakers from LOCNEC+. This difference represents a **large-sized effect** (d = 3.10), and implies that *all* the learners have a slower speech rate than the NS mean, although there is some overlap between the two groups.

The results are in line with previous results from the literature. In a corpus of learners of English (LINDSEI-GE), Brand and Götz (2011) found a very similar mean speech rate for German-speaking learners: about 160 words per minute. Osborne (2010) also reported close, though slightly lower, values for B2/C1 learners of English, and significant differences between high-intermediate learners and native speakers (also Ginther, Dimova & Yang 2010; Munro & Derwing 2001; Rohr 2017).

As explained in the first theoretical chapter of this thesis, **speech rate variability** can be attributed to two main factors. First, as shown by e.g. Guz (2015), part of the variability might simply be due to each speaker's own **speaking style**. Although the present data

unfortunately do not allow for an analysis of the relationship between the learners' L1 and L2, her results indicated that the two are positively correlated: fast (slow) speakers tend to speak faster (slower) in an L2 as well (see also Towell, Hawkins & Bazergui 1996). Likewise, native speakers are also characterised by differing speaking styles and speech rates (e.g. Raupach 1980).

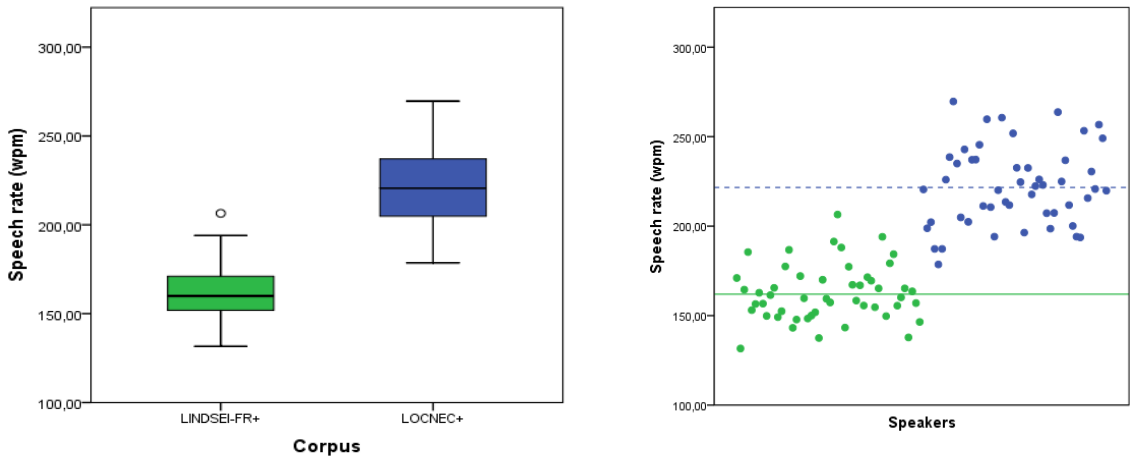


Figure 5-5: Boxplots and stripcharts of speech rate (in words per minute) in LINDSEI-FR+ and LOCNEC+

The second factor that may account for the large variability in the data is **proficiency level**: previous research consistently reported a strong correlation between proficiency level and speech rate (e.g. Baker-Smemoe *et al.* 2014; Cucchiarini, Strik & Boves 2000; Derwing *et al.* 2004; 2009; Kormos & Dénes 2004; Rossiter 2009; Tonkyn 2012). This aspect will be considered in Chapter 7, where learners' performances are related to their assessed CEFR fluency level.

5.2.2 Mean length of runs

Mean length of run (MLR) is calculated by averaging the number of words per run. As explained in Chapter 3, a run is defined as a word or a sequence of words that occurs between two unfilled pauses (Götz 2013a; Grosjean 1972; Tavakoli 2016) – see also Figure 3-3 and Figure 3-4 for illustrations of speech runs.

As visually represented in Figure 5-6, in LINDSEI-FR+, the learners utter **5.64 words per run** on average. The distribution of invividual means ranges from 3.85 (FR041) to 6.05 (FR010) (or to 6.32, the two outliers included). The native speaker mean in LOCNEC+ is much higher, reaching **8.01 words per run**, with a considerable dispersion too (min: 5.22; max: 11.41/15.15, outliers ex-/included). As can be observed, the two distributions overlap a lot, and some learners outperform the native speaker mean.

The mean difference between the two speaker groups is statistically **significant** (t = -7.12; p < .000) and represents a **large-sized effect** (d = 1.42). A Cohen's d of 1.42 means that c. 48 % of

the two groups overlap and that 92 % of the NS group is above the mean of the learner group (i.e. four native speakers have a lower MLR than the NNS mean).

The results corroborate previous findings indicating that learners, on average, utter shorter speech runs than native speakers (Ginther, Dimova & Yang 2010; Guz 2015; Hincks 2010; Rohr 2017). The length of runs is generally seen as a reflection of the **degree of automation of speech processes** (the longer the runs, the greater the automatisation) (*cf.* e.g. Cucchiarini, Strik & Boves 2000; Derwing *et al.* 2004; Kormos & Dénes 2004; Préfontaine, Kormos & Johnson 2015). The present results thus suggest that even high-intermediate to advanced learners have not automated linguistic processes to the same extent as native speakers.

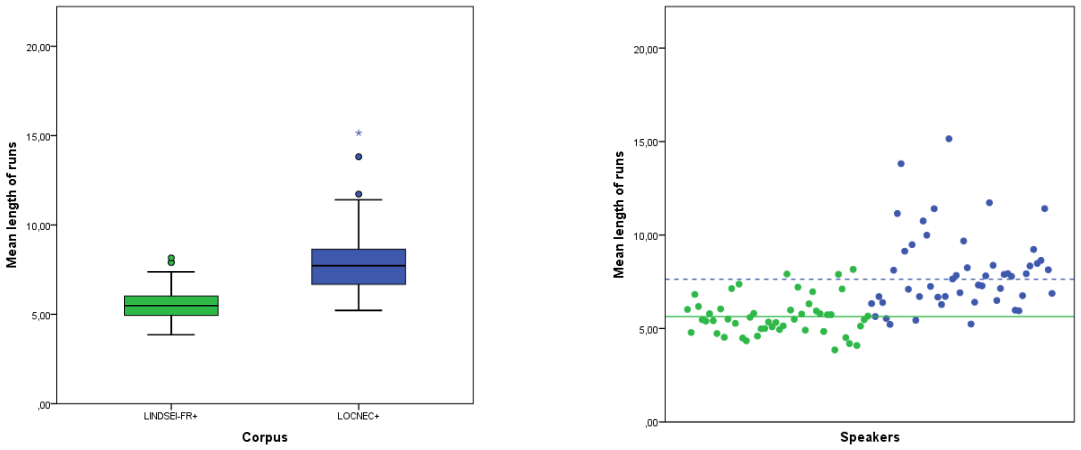


Figure 5-6: Boxplots and stripcharts of mean length of runs in LINDSEI-FR+ and LOCNEC+

More interestingly perhaps, what the present results bring to light is the **considerable dispersion of individual mean lengths of runs in both speaker groups**. In learner speech, MLR has been shown to be related to proficiency level (e.g. Baker-Smemoe *et al.* 2014; Kormos & Dénes 2004), but, while it may account for the variability in LINDSEI-FR+, this factor cannot be extended to explain the tremendous dispersion in the native speaker data. A tentative explanation is that native (and learner) runs are very **idiosyncratic** and that speakers differ in their preference for shorter, or longer, runs.

5.2.3 Phonation time ratio

The phonation time ratio (PTR) is the time spent talking as a percentage proportion of the total time needed to produce the speech sample. A higher phonation time ratio can be associated with a higher fluency (i.e. less pausing).

The mean phonation time ratio in LINDSEI-FR+ amounts to **82.75%** and the average in the native corpus to **86.78%**. The boxplots and stripcharts shown in Figure 5-7 further reveal that there is a great overlap between the two distributions, with NNS individual PTRs ranging

from 69.87% (FR041) to 91.62% (FR042), and NS ratios ranging from 78.68% (EN013) to 92.76% (EN015).

A t-test confirmed that the **learners** in LINDSEI-FR+ have a **significantly lower mean phonation time ratio** than the native speakers in LOCNEC+ ($t = 4.76$; $p < .000$), and the effect size of this difference is large ($d = 0.951$): c. 65% of the two groups overlap and c. 83% of the native speakers have a higher PTR than the learner mean.

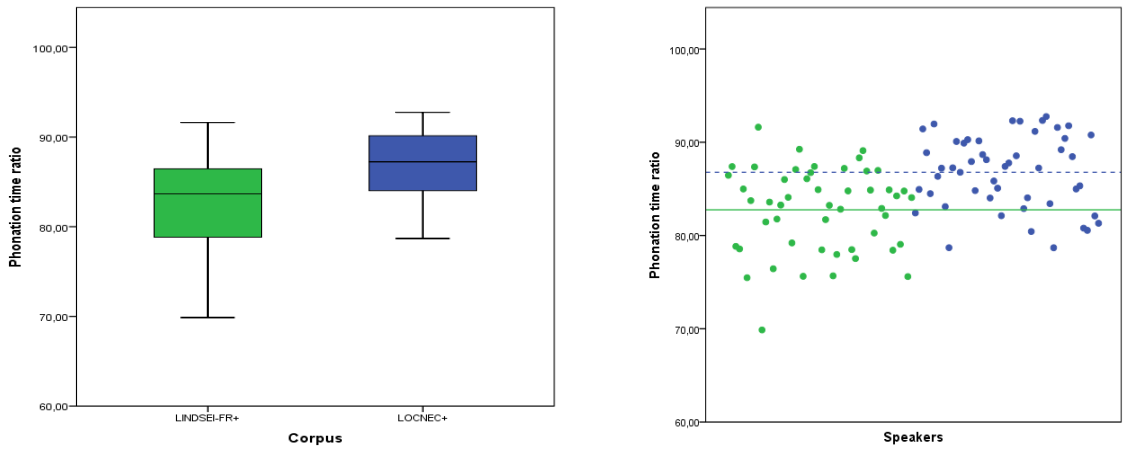


Figure 5-7: Boxplots and stripcharts of phonation time ratio in LINDSEI-FR+ and LOCNEC+

Although previous studies have also reported lower PTRs in learner speech as compared to native speech (Kahng 2014; e.g. Kormos & Dénes 2004), the mean PTR reported for LINDSEI-FR+ is far higher than those reported elsewhere. This might partly be due to the high **proficiency** level of the learners or to speaking task properties (Iwashita *et al.* 2008; Tavakoli 2016). However, as PTR heavily relies on the accurate measurement of UP time, another likely explanation might be the way unfilled pauses were detected and measured in this study, i.e. automatically, through time alignement, and with a threshold slightly lower than that used in some other studies (0.2 vs. 0.25 sec.).

5.3 ANNOTATED (DIS)FLUENCY FEATURES

The ten annotated (dis)fluency features under investigation are analysed in the following sections. The focus lies on the mean frequencies and dispersion in learner and native speech, rather than on the detailed analysis of the use or position of the features. The order of the sections follows the mean frequency of the features in LINDSEI-FR+, from the most to the least frequent.

5.3.1 Unfilled pauses

It is widely accepted that the **identification of unfilled pauses** (UPs) is "a very difficult exercise when done entirely manually, and we have noticed that most linguists, even highly competent ones, tend to miss many silent [unfilled] pauses, especially when they are coupled with other phenomena (such as hesitation or syllable lengthening)" (Campione & Véronis 2005:44). Bearing this warning in mind, the detection of unfilled pauses in LINDSEI-FR+ and LOCNEC+ combined (1) a manual transcription of perceived unfilled pauses ("UPL" annotation tags[127]) with (2) an automatic detection of unfilled pauses of at least 0.200 second ("UPA" tags).

Table 5-2 below displays the proportion of manually transcribed (UPLs) and automatically detected intra-turn unfilled pauses (UPAs) in LINDSEI-FR+ and LOCNEC+. As can be observed, the proportion of automatically detected UPs is non-negligible: 34% of the learner UPs and 47% of the native UPs were, in fact, not marked down in the transcriptions of LINDSEI-FR and LOCNEC.

	LINDSEI-FR+	**LOCNEC+**
Manually transcribed unfilled pauses ("UPL")	65.97% (7,826)	52.47% (4,627)
Automatically detected unfilled pauses ("UPA")	34.03% (4,037)	47.53% (4,191)
Totals	**100% (11,863)**	**100% (8,818)**

Table 5-2: Proportion (absolute frequency) of UPL and UPA in LINDSEI-FR+ and LOCNEC+

[127] As a reminder, the manually transcribed unfilled pauses (UPLs) correspond to the ".", ".." and "..." in LINDSEI-FR and LOCNEC transcriptions. No lower threshold was applied for those unfilled pauses because it was thought that the fact that they were perceived mattered more than their empirical length. This, however, implies that some of the UPLs might be lower than 200 ms.

The fact that about half of the native pauses were not perceived and transcribed (as compared to a third of the learner pauses) suggests that native pauses are, overall, less noticeable than learner pauses. L1 and L2 unfilled pauses might not be perceived because of several factors, including the length of the pause, its position in the utterance, or its combination with other (dis)fluency features. Consider the following examples (**UPAs** are in bold font and UPLs are underlined):

5-1: FR044-F - UPL and UPA (1)

> I think that it's less common now **(0.560)** people are sticked to their roots and they: (0.230) rather (0.100) the= they try to

5-2: EN021-F - UPL and UPA (1)

> she had a small boy (0.930) erm **(0.320)** and **(0.310)** and then **(0.370)** he started to fall in love with her

5-3: EN021-F - UPL and UPA (2)

> we did a quiz the other night **(0.640)** er **(0.310)** and the: **(0.510)** executive the committee **(0.420)** got completely thrashed like everybody else **(0.400)** which is a bit embarrassing

In Example 5-1, which comes from the learner corpus, the first unfilled pause was not perceived, despite being quite long. This might simply be due to the fact that the pause lies at a clause boundary and is thus not disruptive (*cf.* also Candea 2000; Pawley & Syder 2000 on the distinction between structuring and non-structuring UPs). By contrast, the other two pauses (*(0.230)* and *(0.100)*) are comparably much shorter, but they were perceived and transcribed, probably because they are preceded and followed by other (dis)fluency features (a lengthening, a restart, and a truncation). In Example 5-2, which is from a native speaker, the first unfilled pause was perceived and transcribed (although it is at a juncture, it is quite long and precedes a filled pause). The next three pauses, however, were not perceived, although the native speaker clearly hesitates about what to say. Finally, Example 5-3 shows five unfilled pauses that were not perceived in native speech. Some lie at clause boundaries, but others seem to be more closely associated with difficulties in the precise wording (*cf*. the lengthening and the restart). It is also worth mentioning that the perception of an unfilled pause cannot always be associated with longer actual duration, as can be seen by the first UPL in Example 5-4 (0.050 second).

5-4: FR044-F - UPL and UPA (2)

> we will have to go to schools talk to people and (0.050) after this we will have to: (0.660) to compile er the results

It unfortunately falls out of the scope of this thesis to analyse more precisely the occurrences where unfilled pauses were perceived or not, but such an analysis would undoubtedly prove particularly insightful. The findings and examples shown before, however, clearly highlight the need to complement a manual identification of unfilled pauses with more automatic methods.

In what follows, the term "unfilled pause" will be used as a cover term for both UPLs and UPAs, unless otherwise specified.

5.3.1.1 Frequency of unfilled pauses

In total, 11,863 unfilled pauses were detected in LINDSEI-FR+. They account for 1h 39 min. In the native corpus, the 8,818 unfilled pauses total 1h 15 min.

The average frequencies of unfilled pauses in LINDSEI-FR+ and LOCNEC+ are the highest among the annotated (dis)fluency features investigated in this thesis. The corpus data reveal that, on average, the **French learners produce more than one unfilled pause every 10 words**: they occur with a frequency of 12.69 UPs per hundred words in LINDSEI-FR+. This very high mean confirms the widespread claim that unfilled pauses are a particularly pervasive phenomenon in spontaneous learner speech. The learner mean is comparable with earlier findings: Cucchiarini *et al.* (2002) found a mean of 31 UPs per minute in learner Dutch, De Jong *et al.* (2012a) found a mean of 27 UPs phw for learners of Dutch, and Götz (2013a) found a mean of 15 UPs phw in learner English. In the **native corpus**, an average of **one unfilled pause every 14 words** was measured, i.e. 6.95 UPs phw on average. Thus, it appears quite clearly that LINDSEI-FR+ learners pause much more frequently than native speakers. This observation is further substantiated by a *t*-test comparing the mean frequency of UPs in learner and native English and which indicates that there is a **significant difference between the two groups** ($t = 12.052$; $p < .000$).

Yet, in both corpora, **variation is quite large**, and larger in the learner data than in the native data, as is also illustrated in Figure 5-8 below. The means of individual learners range between 7.68 (which is higher than the native speaker mean) and 20.20 UPs phw; in LOCNEC+, the minimum frequency is 4.00, and the highest frequency is 11.58 UPs phw (which is lower than the learners' mean). Therefore, although there is quite a lot of overlap between the two distributions, all the learners produce more UPs per hundred words than the average native speaker from LOCNEC+.

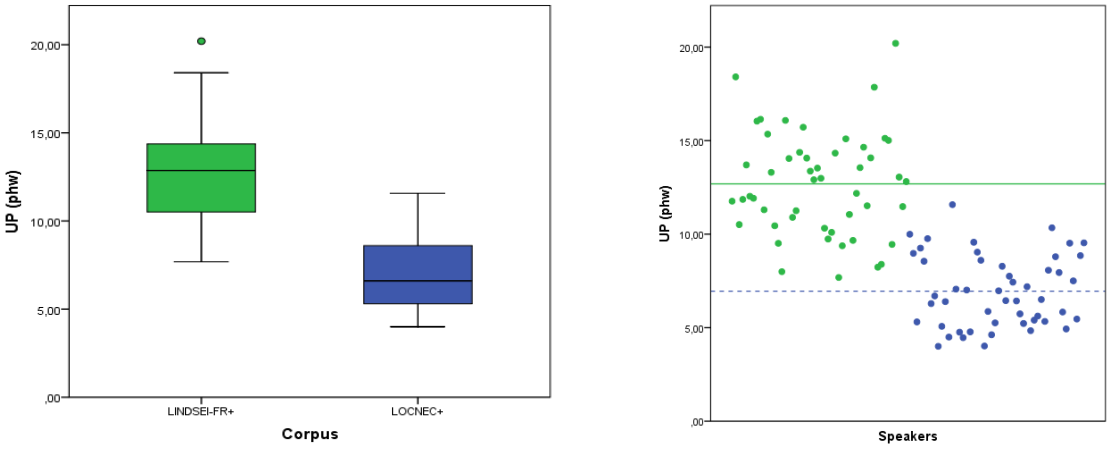


Figure 5-8: Boxplots and stripchart of UPs (phw) in LINDSEI-FR+ and LOCNEC+

It is nearly impossible to determine the exact reason why speakers pause in their speech. Some insights into the functions of unfilled pauses might nonetheless be gained from the contextual use of unfilled pauses, i.e. whether they are used together with other (dis)fluency features and, if so, which ones. Three such "clusters" of unfilled pauses emerge from the examination of the corpus data. Unfilled pauses are, for example, regularly used together with the **conjunctions** *and*, *so*, and *but*, as in Example 5-5. Due to the presence of the conjunction, the cluster is typically found at clause boundaries. In this association, UPs seem to emphasise the segmentation of speech, and are arguably a positive use of pauses.

5-5: FR007-S - UP+C

> he was married to erm (0.550) a German woman **(0.480) and** they had lived in Zaire for quite a long time

Unfilled pauses are also often used before or after a **filled pause**, or they may also surround the filled pause, as illustrated in 5-6 (and 5-5). These uses of unfilled pauses could perhaps be associated with planning or formulation problems. However, some can be found at clause boundaries, and may thus function similarly as UP+C (i.e. as segmentators; *cf*. also Examples 5-2 and 5-3).

5-6: EN026-F - UP+FP+UP

> I will do the er language courses **(0.550) erm (1.010)** initially as a sort of to to augment my

A third common association of unfilled pauses is with **repetitions**: in such associations, the pause typically lies between the original utterance (the *repeatable*) and the *repeated*, as in Example 5-7. Such uses of unfilled pauses also seem to emphasise planning and formulation issues.

5-7: FR002-S - Rep+UP

> we decided **to (0.150) to** take **a (0.110) a** taxi

Although these three clusters clearly stand out in both the learner and the native corpus, many more exist. Future analyses could review such combinations more systematically and, possibly, uncover different pausing patterns in learner vs. native speech.

5.3.1.2 Mean length of unfilled pauses

The precise analysis of the mean length of unfilled pauses requires time aligned corpus data. In LINDSEI-FR+ and in LOCNEC+ interviews, all unfilled pauses have been measured automatically and precisely in milliseconds.

The corpus data show that the mean length of UPs is very close in the two corpora, though (surprisingly) slightly lower in the learner corpus: the **average learner unfilled pause is 0.506 second long**, and the **average native pause is 0.520 second long**. A visual representation is shown in Figure 5-9. In the spontaneous language of French native speakers, Campione and

Véronis (2005) found a very close geometric mean of the length of unfilled pauses: in their data, UPs are 0.496 second on average. A *t*-test showed that, although there is a slight tendency for learners to produce slightly shorter unfilled pauses, this difference is, in fact, **not significant** ($p > .005$).

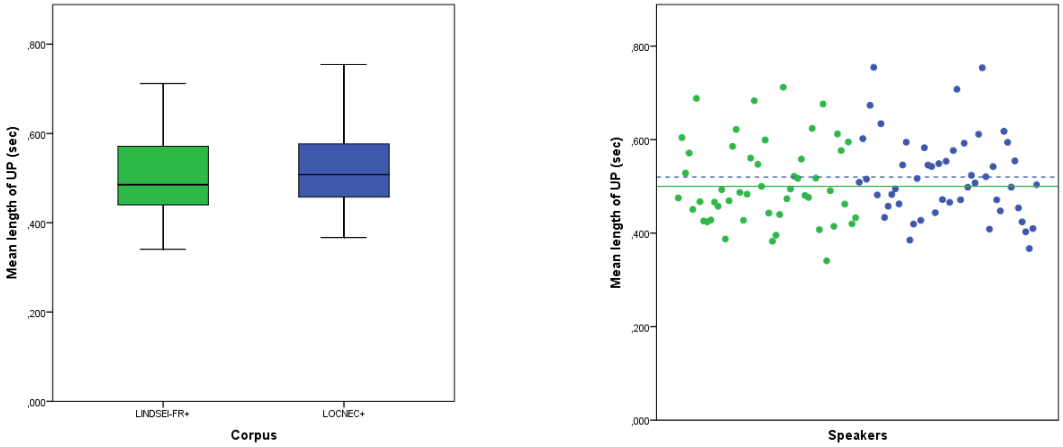


Figure 5-9: Boxplots and stripcharts for mean UP length (in sec) in LINDSEI-FR+ and LOCNEC+

These results are slightly surprising given that the literature indicates that learners tend to produce longer pauses on average than native speakers (Guz 2015; Rohr 2017). Moreover, previous research indicates that the mean length of pauses is affected by the mother tongue background of the learners, and that French native speakers pause on average longer than English native speakers (Grosjean & Deschamps 1975; Préfontaine, Kormos & Johnson 2015). In our data, a possible influence from the learners' L1 does not seem to be at play. However, there is also some evidence in the literature that learners at higher proficiency levels are able to produce more native-like pausing patterns (Riazantseva 2001). The learners in LINDSEI-FR have a high proficiency level, so it does not seem impossible that they have internalised the native length of pauses (while still producing more UPs on average than native speakers).

Another likely hypothesis for the absence of significant difference between the mean length of L1 and L2 unfilled pauses might be found in the distinction between UPLs and UPAs. While the mean lengths of learner and native UPs mentioned above are based on all unfilled pauses (i.e. UPLs and UPAs without distinction), many studies are, in fact, based on the duration of perceptively transcribed pauses only (i.e. UPLs only; e.g. De Jong & Bosker (2013)). To check whether there might be a different tendency when perceived and automatically-detected unfilled pauses are distinguished, the mean lengths of UPLs and UPAs were compared in LINDSEI-FR+ and LOCNEC+. As displayed in Table 5-3, the mean length of perceived UPs in LINDSEI-FR+ is 0.581 second, and 0.663 second in LOCNEC+; the mean length of automatically detected UPs in LINDSEI-FR+ is 0.370 second and 0.391 second in the native corpus. The figures seem to confirm that **perceived unfilled pauses are generally longer** than automatically detected UPs. However, as was the case for the mean length of UPs, the mean length of learner UPLs and UPAs is slightly *lower* than native UPLs and UPAs. *T*-tests indicate that **the mean length of both perceived and transcribed unfilled pauses is actually**

significantly different in the two groups (UPLs: t = -2.75; p < .05; UPAs: t = -2.68; p < .05): **learner UPLs *and* UPAs are significantly *shorter*** than the native counterparts. This finding is very perplexing, to say the least.

	LINDSEI-FR+	**LOCNEC+**
Mean length of UPLs	0.581 ms	0.663ms
Mean length of UPAs	0.370 ms	0.391 ms

Table 5-3: Mean length of UPLs and UPAs in LINDSEI-FR+ and LOCNEC+

To get further insights into the results, different lengths of unfilled pauses were examined. Following Campione and Véronis (2002), who highlighted the fact that the distribution of the length of unfilled pauses is trimodal in sponatenous speech, the unfilled pauses in LINDSEI-FR+ and LOCNEC+ pauses were classified into **short (< .200 sec), medium (< 1.000 sec) and long (> 1.000 sec) unfilled pauses**. The results are visually represented in Figure 5-10.

In both the learner and the native corpora, the overwhelming majority of unfilled pauses are of medium length: c. 80% (9,480 UPs) of the L2 unfilled pauses and c. 86% (7,560 UPs) of the L1 unfilled pauses last between 0.200 and 1 second. Learners, however, have a higher proportion of short (11.32%, 1,343 UPs vs. 6.62%, 584 UPs) and long unfilled pauses (8.77%, 1,040 UPs vs. 7.64%, 674 UPs) than native speakers.

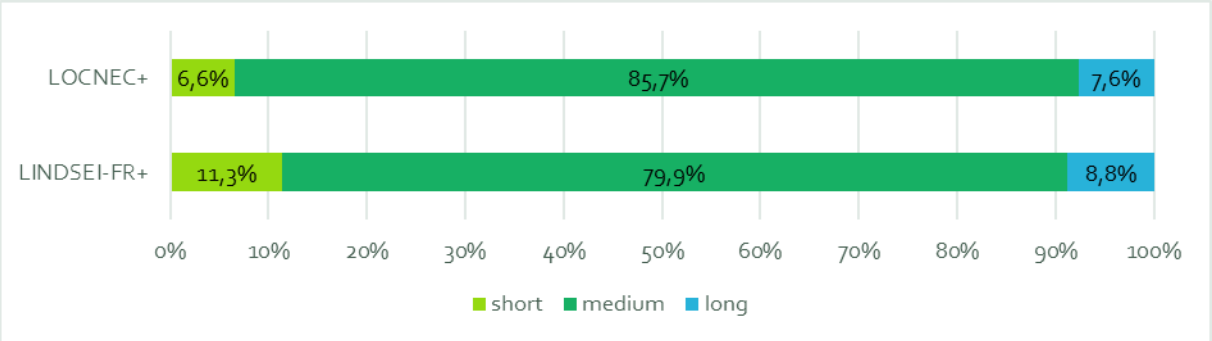


Figure 5-10: Proportion of short, medium and long UPs in LINDSEI-FR+ and LOCNEC+

Although there is a slightly higher proportion of long unfilled pauses in LINDSEI-FR+, the difference with native speakers is still very small. What might contribute to the impression of longer UPs in learner speech, however, is the maximum length of long unfilled pauses, the (dis)fluent context around the pause and the overall speech rate of the speaker (Duez 1985; Goldman *et al.* 2010). Examples 5-8 and 5-9 illustrate the two longest unfilled pauses in LINDSEI-FR+ and LOCNEC+, respectively. Note that the long learner unfilled pause, which is more than 6 seconds long, is situated in a larger sequence of disfluencies comprising an unfilled pause, a lengthening and a repetition. By way of comparison, the longest native UP (nearly 5 seconds) is preceded by a lexical editing expression indicating that the speaker is searching for a specific word (*what's the word*), as well as by another long pause and an approximator (*some sort of*).

5-8: FR030-P - long UP

> a woman has asked a painter to (0.300) <laughing/> (0.360) to: **(6.560)** to paint yes to represent her **(1.720)** her portrait

5-9: EN035-F - long UP

> I think there's like some sort of **(2.350)** erm what's the word **(4.780)** this barrier like between students and the (0.670) the people who live here

Another possible explanation for the perplexing results might be that, due to higher cognitive demands, the learners prefer shorter, but more regular, lulls to plan the next words, while native speakers might prefer longer and less frequent pauses to plan longer speech runs. A last possible explanation could be the subtly different context of the interviews. While learners may have unconsciously thought that long periods of silence were "inappropriate" in the frame of an interview (which specifically aimed to analyse their L2), native speakers may not have felt the same pressure to "save face".

5.3.2 Filled pauses

Using the tag <FP>, 7,576 instances of filled pauses were retrieved from the learner corpus, and 2,997 were found in the native corpus.

The corpus data reveal that, in the speech of **French learners of English**, FPs have a mean frequency of 7.80 occurrences per hundred words, as visually represented in Figure 5-11. In other words, learners produce a filled pause every 15 words. **Native speakers** produce fewer filled pauses: 3.38 FPs phw on average. The learners' mean is thus nearly three times as high as the native speakers', and it is also slightly higher than the average reported for German learners of English in LINDSEI-GE (5.12 FPs phw) (Götz 2013a:110). An independent-samples *t*-test on mean frequency of filled pauses per hundred words was statistically significant, revealing that **on average, learners in LINDSEI-FR+ produce more filled pauses than the native speakers in LOCNEC+**. A follow-up test further indicates that Cohen's *d* is 1.99, indicating that the distributions of L1 and L2 filled pauses barely overlap. These results thus lend credence to the hypothesis of an overall overuse of filled pauses by learners.

It is, however, also obvious from Figure 5-11 that the data are distributed differently in the learner and native corpora, with the NSs apparently behaving much more homogeneously than LINDSEI-FR+ learners, despite the fact that there are three NS outliers. The examination of Figure 5-11 reveals that the range between the learner who produces the least and the most FPs phw is quite important, from 2.61 to 13.61 (i.e. 5 times as many) and individual native speakers' means range between 0.41 and 7.09 (i.e. 17 times as many). Levene's Test for Homogeneity of Variances proves that the variances in the two speaker groups are significantly different ($F = 27.26$, $p < .000$).

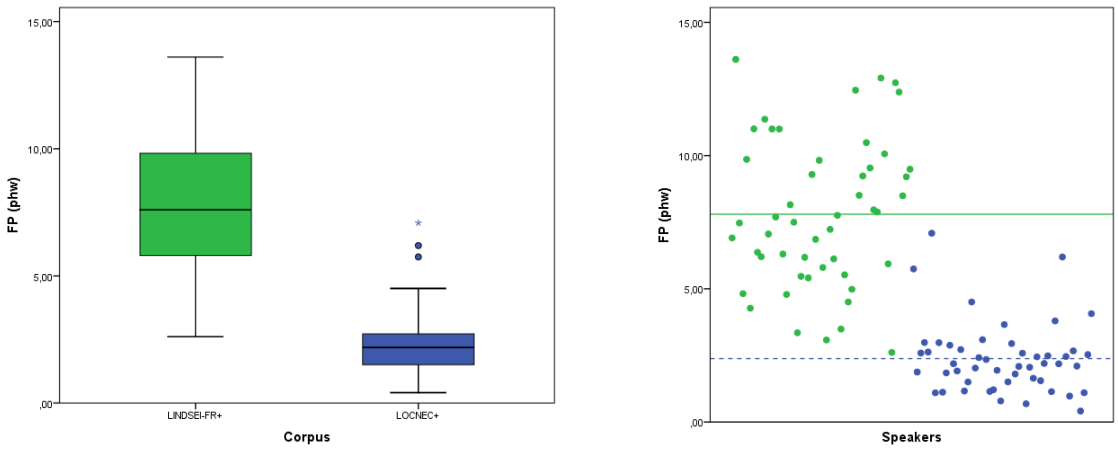


Figure 5-11: Boxplots and stripchart of FPs (phw) in LINDSEI-FR+ and LOCNEC+

Figure 5-12 reveals the breakdown of the different filled pauses in LINDSEI-FR+ and LOCNEC+. It is strikingto see that, while **learners have a clear preference for the FP *er*** (64%; 4,822 occurrences), **native speakers** rather prefer the form ***erm*** (58%; 1,741 occurrences), and, to a lesser extent, *er* (30%; 909 occurrences). *Eh*, *em*, and *mm* are rarely used in either learner or native speech. Illustrations of the different filled pauses are shown in 5-10 through 5-13.

As underlined by Gilquin (2008), *er* is very close to the common French filled pause *euh*, and it might well be that the learners **transfer**, to some extent, their pausing behaviour to their L2 (*cf.* also Clark & Fox Tree 2002). However, this hypothesis does not explain why there is such a huge difference between the learners' use of *er* and *eh*, which are also very close to the French *euh*. Besides, previous literature has underlined that **nasal and non-nasal filled pauses** may not have the same function in speech: whereas *uh* (the Americal spelling for *eh/er*) is indicative of a short delay used lexical identification, *um* (i.e. *em/erm* in British English) is a signal of a long upcoming suspension (Fox Tree 2001). While the association of *er* and *eh* with lexical retrieval seems to find some support in the results for learner speech, it seems perplexing that *erm*, which allegedly signals long delay, is the preferred native filled pause.

5-10: FR002-F - er and eh

they have to (0.630) <noise/> (0.450) to put **er** different **eh** kinds of **er** dresses

5-11: FR002-P - em

it seems that the painter is very (0.630) **em** (0.860) near (0.410) the reality

5-12: FR042-F – mm and eh

they are puzzled just because of that **mm** particular verb not because they **eh**

5-13: EN026-F - erm and eh

it was rather bizarre but **erm** but it it **eh** actually if you'd looked into it

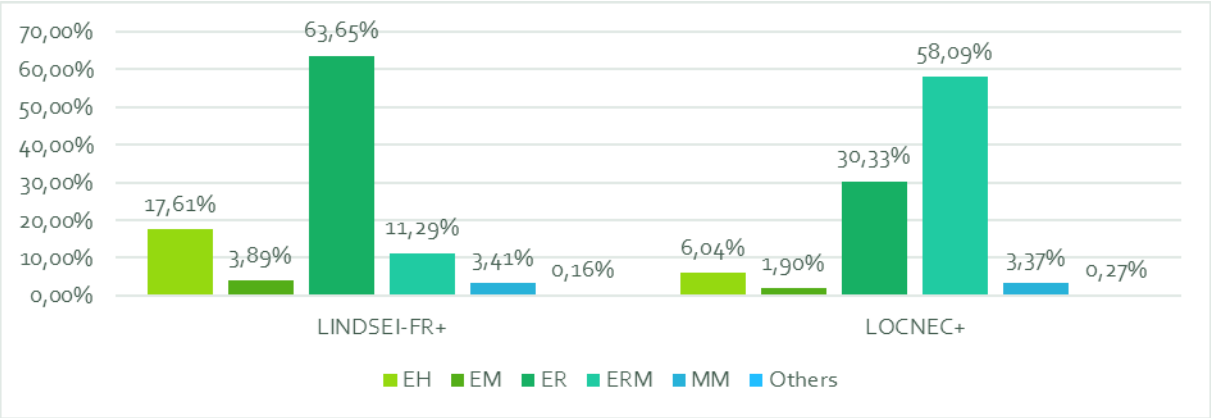


Figure 5-12: The different FPs in LINDSEI-FR+ and LOCNEC+

Although the difference of preferred forms of filled pauses between LINDSEI-FR+ and LOCNEC+ is striking, there also seems to be **considerable individual differences** within each corpus. Consider Figure 5-13 and Figure 5-14, which show the use of the different forms of filled pauses for three learners and three native speakers. While about 60% of the filled pauses of FR012 are *er*s, these only amount to 43% for FR022 and a small 19% for FR029. Similarly, whilst EN020 nearly exlusively uses *erm*s, EN025 and EN037 seem to have a much more varied use of filled pauses. Note also that the various forms of filled pauses may follow one another at a quick pace, as illustrated in 5-14.

5-14: FR005-S - many forms of FPs

er and I helped **eh** some Irish people (0.350) to **mm** to care **erm** (0.090) with **eh** handicapped people

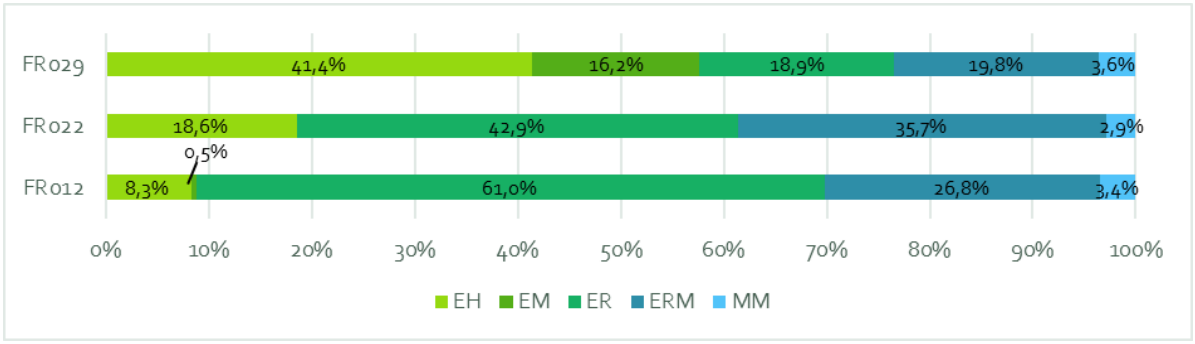


Figure 5-13: The use of FPs by three learners from LINDSEI-FR+

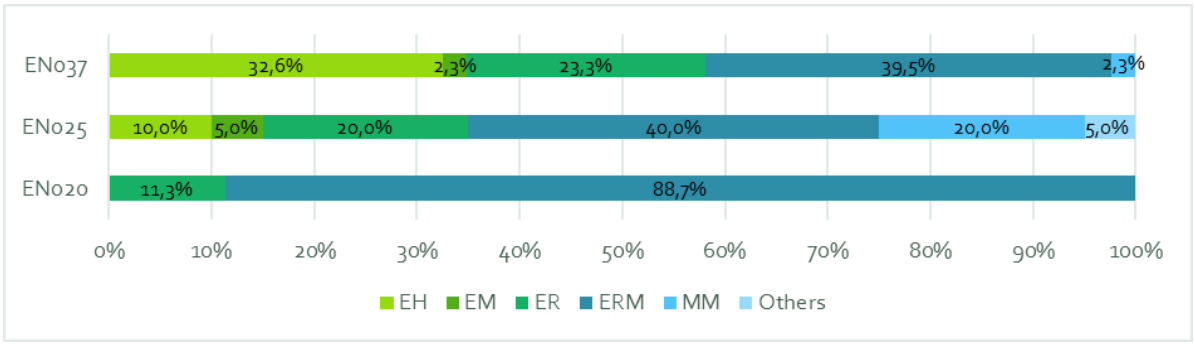


Figure 5-14: The use of FPs by three native speakers from LOCNEC+

To get further insights into the use of filled pauses in L1 and L2 speech, their contextual use was analysed (Table 5-4). As was the case for unfilled pauses, filled pauses are **mainly used in clusters,** i.e. in combination with other (dis)fluency features in adjacent position. Only 25% of the L1 and L2 filled pauses are used on their own (Examples 5-15 and 5-16). This sheds a different light on Riggenbach's (1991) claim that the presence of clusters of filled (and unfilled) pauses is indicative of non-fluent speakers.

5-15: FR030-P - FP used alone

she says **er** is it really me

5-16: EN001-F - FP used alone

they're quite accepting of you know whether I wanted to **er** go to university or not go to university

	LINDSEI-FR+	**LOCNEC+**
Stand-alone use of FPs	25.37% (1,922)	24.12% (723)
FP used in clusters	74.63% (5,654)	75.88% (2,274)
Totals	**100% (7,576)**	**100% (2,997)**

Table 5-4: Proportion (absolute frequency) of FPs used alone and in clusters in LINDSEI-FR+ and LOCNEC+

When they are used in clusters, both L1 and L2 filled pauses are generally combined with one or two **unfilled pauses** (*cf*. previous section) or with a **conjunction**. This latter association was already pointed out by Clark and Fox Tree (2002), who noted that the filled pauses *uh* and *um* "are often cliticised onto prior words and never onto following words" (*ibid.*:101), especially what they call introductory conjunctions such as *and*, *but* or *so* to form "an.duh" (*and eh; and er*), "bu.tuh" (*but eh; but er*), and "so.wuh" (*so eh; so er*). Examples of combinations of conjunctions with filled pauses are shown in 5-17 to 5-19. Some of these clusters might be associated with planning issues, especially when they are found at the end of speech runs, as is the case in 5-18 and 5-19. Lastly, contrarily to Levelt's claim, L1 and L2 filled pauses are only rarely used jointly with restarts (see Example 5-20).

5-17: FR040-P - C+FP

she realised she was smiling **and er** he painted her (0.280) you know like a

5-18: FR038-F - C+FP

B: it's (0.340) still a bit vague **but eh**

A: and was it you who chose the top= the (0.270) the general background

5-19: EN026-F - C+FP

B: you're expected to pay back

A: oh yes

B: **so er**

A: it's quite interesting

B: so it's quite interesting so I mean I

5-20: FR046-F - FP with a restart

but I didn't saw that **eh** see that

5.3.3 Conjunctions

The conjunctions *and, so* and *but* are the third most frequent (dis)fluency feature in LINDSEI-FR+. In the learner corpus, 4,849 conjunctions were retrieved using the tag <C>, and 6,522 occurrences were retrieved in LOCNEC+.

Figure 5-15 offers a visualisation of the L1 and L2 conjunctions (Cs). As can be observed, conjunctions have a mean frequency of **5.15 per hundred words in learner speech**: they occur once every 20 words on average. In the **native corpus**, conjunctions occur with a slightly lower frequency: there is one conjunction every 20 words on average (4.96 phw). A *t*-test revealed that the difference between learner and native means is **not significant**: the learners do not produce significantly more conjunctions than the native speakers. The examination of Figure 5-15 reveals that there there is some **variation** both in the learner and the native data. Individual learners range between 2.79 (outlier excluded) and 7.20 conjunctions per hundred words, while native speakers range between 3.58 and 6.49 (ouliers excluded). However, Levene's test for equality of variances indicates that the variances are not significantly different ($F = 2.78$; $p > .05$).

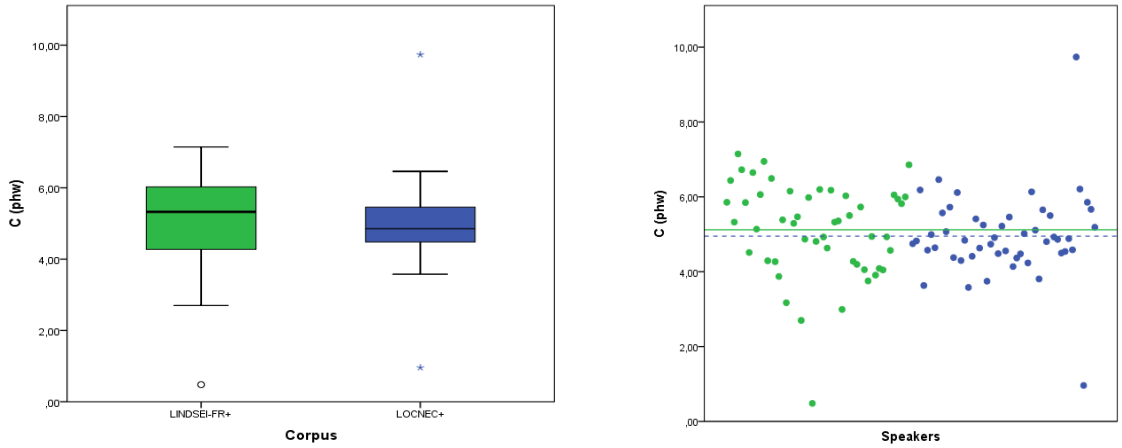


Figure 5-15: Boxplots and stripchats of Cs (phw) in LINDSEI-FR+ and LOCNEC+

Zomming in more closely on the three types of conjunctions under investigation, it appears that learners and native speakers behave, overall, very similarly in their use of *and*, *so*, and *but.* As can be seen in Figure 5-16, ***and* is the most frequently used conjunction in both groups**: in each corpus, c. 58% of the conjunctions are occurrences of *and*. *But* and *so* represent nearly the same proportion, but while there are slightly more *but*s in learner speech (23.8% vs. 17.9%), there are slightly more *so*s in LOCNEC+ (21.2% vs. 20.3%). Figure 5-16 **does**

not reveal any clear pattern of over- or underuse of these three individual conjunctions in learner speech.

It falls out of the scope of this thesis to provide an in-depth analysis of the multifunctionality of *and*, *so*, and *but*, even more so because several functions can be performed simultaneously by the same item and as it is often difficult to disentangle them. Therefore, no attempt was made to identify the function(s) of each item. Nonetheless, some of the main functions of *and*, *so*, and *but* are illustrated in the following sub-sections.

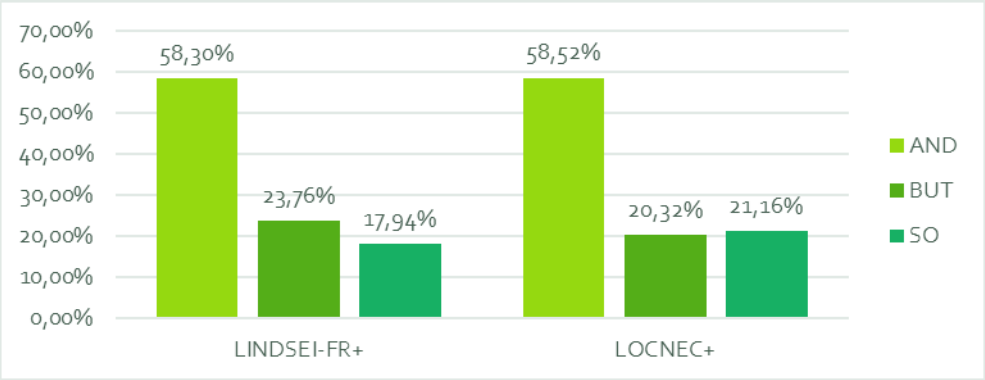


Figure 5-16: Proportion of 'and', 'but' and 'so' in LINDSEI-FR+ and LOCNEC+

5.3.3.1 And

Previous research has also noted the high frequency of *and* in speech. In Chafe (1982) and Fung and Carter (2007), for example, it was the most common among the range of conjunctions studied. *And* has been shown to be particularly multifunctional. First, it can convey the idea of **additivity**, as illustrated in 5-21.

5-21: EN038-F - additive use of 'and'

B: I got myself sacked (0.560) er <unknown/>

A: because you ate too many peppers

B: ate too many peppers **and** I was late too many times **and** I didn't like him particularly **and** he **and** he hated me

However, Chafe (1982) pointed out that, in many cases, *and* simply joins what he calls "idea units" of talk. Similarly, Fung and Carter (2007) noted that *and* is often used as a "**continuer**" (or "staller", in Stenström's (1994) words) to provide the speaker with a conversational space to expand upon (i.e. it signposts the speaker's desire to hold the floor). Such uses of the conjunction seem to correspond to what Schiffrin described as a "structural coordinator of ideas which has pragmatic effect as a **marker of speaker continuation**" (Schiffrin 1987:152 my emphasis). This structural use of *and* is apparent from Examples 5-22 to 5-24. In 5-22, the speaker recalls a journey in the Andes. In the excerpt, the rapid succession of *and*s links and structures the numerous steps in the journey. In 5-23, the conjunction *and* links together a first semantic unit (*there's a fountain in Delph*) with a second one (*you have to drink from it*).

The use of *and* in Example 5-24 is particularly interesting. The learner uses the first *and* to mark the transition between her first utterance (*she watches the picture*) and the following one, but she then decides to replace the word *picture* by *portrait*. This restart is signalled by the discourse marker *well*. After having corrected the word, the learner seems to reaffirm the transition between her previous unit and the following one (*she thinks it's really awful*) by reiterating the conjunction *and.*

5-22: EN041-F – 'and' as marker of continuation

there's this one you come across (0.590) **and** you just come out over the top of a mountain pass (0.540) **and** you've got these staired (1.260) tracks they're all staired (0.680) **and** you just walk around the corner **and** (0.540) there's this (0.260) three point valley (0.390) all meeting up **and** there's a sort of (0.340) stick of rock which goes out (0.540) **and** then there's a settlement on the top of it

5-23: FR003-S - 'and' as marker of continuation

there's a fountain in Delph **and** you have to: to drink from it

5-24: FR001-P - 'and' as marker of continuation

then she she watch= she watches the: the picture **and** er well the portrait **and** she thinks it's really awful

It is noteworthy that, from a contextual point of view, the most frequent use of *and* is the stand-alone use, but while native *and*s are used without other (dis)fluency features in adjacent position in 50% of the cases, this proportion only amounts to 25% in learner speech. The second most frequent use of *and* is in a cluster with at least one unfilled pause, as illustrated before. More specifically, *and* is used in the cluster **UP+*and*** in c. 13% of the cases in both learner and native speech. By way of comparison, there are only about 4% of *and*+UP clusters in both LINDSEI-FR+ and LOCNEC+.

5.3.3.2 But

Compared with *and*, the conjunction *but* does not seem to have a structural function. In many cases, it is used to mark "an upcoming unit as contrasting action" (Schiffrin 1987:152), as in 5-25 and 5-26.

Some uses of *but* do not seem to carry this contrastive meaning and are used to tone down the previous utterance, to **indicate hesitation about what to say next**, or even to express the desire to yield the floor. In Example 5-27, the first *but* appears to indicate that the speaker hesitates about what to say next, which seems to be confirmed by the filled and unfilled pauses preceding and following the conjunction. Alternatively, the first *but* in Example 5-27 may express the interviewee's desire to hand the floor (the interviewer does, indeed, take the floor shortly afterwards). Such a use of *but* is not restricted to learners: it also features in native speech, as in Example 5-28. Incidentally, it is noteworthy that in both 5-27 and 5-28, the conjunction is cliticised with the following filled pause (*cf.* Clark & Fox Tree 2002). Lastly, in 5-29, the learner talks about her disappointment about a play she saw (*it was the first*

impression I had). In the excerpt, she first indicates a transition with *and* to the next unit, which is a contrasting utterance starting with *but* (*but it was funny*). Then, she signals her uncertainty about what to say next by another *but* (which is followed by a filled pause as well as a long unfilled pause). She finally decides to round up the discussion by using the conjunction *so*, followed by a clear statement of the end of the discussion (*that's all I can say*).

5-25: FR015-F – contrastive use of 'but'

> a a big amphi of five hundred seats (0.200) **but** there were not er <laughing/> not enough eh not enough

5-26: EN001-F - contrastive use of 'but'

> she doesn't live with my parents **but** she's there quite a lot of the time

5-27: FR015-F – 'but' as hesitation marker

> B: so I I don't remember anything about it (0.330) **but** er (0.380) <unknown/>
>
> A: what was your impression
>
> B: oh it was nice (0.500) **but** I heard **but** I was (0.480) near to the the

5-28: EN008-S – 'but' as hesitation marker

> B: next summer or bef= hopefully before if I can **but** er
>
> A: perhaps
>
> B: yeah I'd like to it'd be really

5-29: FR019-S – two uses of 'but'

> it was the first (0.230) impression I had (0.170) and er (0.310) **but** it was funny (0.360) **but** er (0.990) so (0.270) that's all I can say <laughing/>

As was the case for *and,* L1 and L2 *but*s are primarily used on their own, though less frequently in LINDSEI-FR+ (27% of the cases) than in LOCNEC+ (43%). When they are used in clusters, *but* is most frequently combined with an unfilled pause, either in frontal or back position (i.e. UP+*but* or *but*+UP). Both clusters occur with nearly the same frequency in both learner and native speech (c. 9%).

5.3.3.3 So

While Buysse (2012) found that Dutch-speaking learners of English overuse *so* compared to English native speakers, Müller (2005) found that German-speaking learners of English underuse *so.* In LINDSEI-FR+, the French-speaking learners tend to use slightly fewer *so*s than the native speakers from LOCNEC+, but the difference is very small (*cf.* Figure 5-16).

Most prior research has focused on *so* for marking "result" (Schiffrin 1987) or "inference" (Blakemore 1988; 2002; Fraser 1999) between utterances – such uses of *so* are illustrated in 5-30 and 5-31.

5-30: FR013-S - inferential use of 'so'

I've only been through Koeln (0.730) on the train (0.210) **so** (0.300) I've not really seen it

5-31: EN001-F - inferential use of 'so'

I did (0.990) English literature (0.350) and language (0.550) and French **so** there was reading involved in (0.510) most of my courses

Yet, the use of *so* is not restricted to inference. *So* may also function as a "marker of cohesion" (Howe 1991; Bolden 2009), as a "topic sequencer" (Johnson 2002), as a "marker of elaboration" (Buysse 2009), or it may mark "a speaker's readiness to relinquish a turn" (Schiffrin 1987:218). In learner speech, *so* has been shown to have as many as ten functions (Buysse 2012). Some examples are discussed below.

In Example 5-32, the learner uses *so* (together with a filled and an unfilled pause) to introduce a restart (*people who succeeded in Spanish*). In 5-33 and 5-34, *so* is used conjointly with a filled or unfilled pause at the end of a turn, possibly to indicate that the speaker is ready to relinquish the floor. While the interviewer takes the floor after *so* in 5-33, this is not the case in 5-34. An example of the use of *so* in native speech is shown in 5-35. The speaker uses six *so*s in the passage, as if to sequence his discourse. The first *so* (*(0.850) so erm (0.550)*) seems to be used to come back to the main topic after a digression, i.e. the retelling of an experience that taught him a lesson. The second and third *so*s (*(0.640) so we went out and so we had a good time*), indicate a result or an inference. The next *so* appears to be used to make a temporary summary (*(0.900) erm (1.200) so everything was going very well*). The last two *so*s (*(0.410) so erm she wasn't over that and so she shouldn't have gone out*) are, again, used to make inferences.

5-32: FR046-F - 'so' used with a restart

she said only eh people who have (0.560) a certain er percent= eh **so** (0.460) people who succeeded er in Spanish can go with me

5-33: FR046-F - 'so' used to mark readiness to relinquish the floor

B: there was er no er large difference between us **so** er

A: yeah

5-34: FR002-S - 'so'

something like er five minutes away from the station **so** (0.370) we had to eh to take eh the car

5-35: EN007-S - 'so'

I thought I might tell you about something that happened in the first year (0.250) or someone I went out with in the first year and that that taught me a lesson (1.010) actually erm (0.540) the girl's name was <name/> and she was very nice (0.850) **so** erm (0.550) I asked her out and she agreed (0.640) **so** we went out for a while […] it was near the end of term (0.510) it was near Christmas coming up to Christmas **so** we had a good time […] that was that was great that was very good (0.900) erm (1.200) **so**

everything was going very well […] she was still carrying along a lot of er (1.070) what you would call you know emotional baggage or whatever (0.410) **so** erm she wasn't over that properly **so** she really she shouldn't have gone out with someone else

What seems to emerge from the above examples of *and*, *so*, and *but* is that, contrarily to some other (dis)fluency features, they are generally used between utterances and are rarely found in intrusive places (i.e. in the middle of closely-knit chunks of words). This is fully consistent with previous research (Crible 2017a; Cuenca 2013; Gilquin & Granger 2015; Gilquin 2016; Valdmets 2013). Furthermore, what *and*, *so*, and *but* also have in common is that they are often combined with filled and/or unfilled pauses to form (dis)fluency clusters (see also Crible 2017a; 2017b; Crible & Cuenca 2018). Future research could investigate whether different clusters might be associated with different functions.

5.3.4 Repetitions

Repetitions, which are included in Skehan's (2003) repair fluency category, are the fourth most frequent (dis)fluency feature in learner speech. To obtain the data, I looked for the tags <R0 to obtain the number of repetitions, and R1>, R2> etc. to get the number of simple (R1>), double (R2>) etc. repetitions. A script was created to automatically count the number of words in the *repeatable* (i.e. the initial utterance).

LINDSEI-FR+ contains a total of 3,748 repetitions, against 2,750 in LOCNEC+. On average, the French **learners utter 3.94 repetitions** per hundred words, which means that repetitions occur once every 25 words, whereas **native speaker** repeat themselves **2.15 times every hundred words** (one repetition every 46 words). The figure for the native speakers is very similar to those reported elsewhere (e.g. MacGregor, Corley & Donaldson 2009) but the mean for the learners is higher than those reported previously for other learner groups. De Jong *et al.* (2012a) found an average of 2.1 repetitions per hundred words in the speech of intermediate to advanced learners of Dutch from varied L1s, Gráf (2017) reported a mean of 1.91 "repeat" per hundred word in Czech learners of English, and Götz (2013a) found an average of 0.69 repetitions per hundred words in German-speaking learners of English (but she only considered a closed list of repeats).

As was the case for the other (dis)fluency features so far, the range between the lowest and the highest mean is quite large (see Figure 5-17). In the learner corpus, repetitions range between 1.33 phw and 6.92 phw (outlier excluded); in LOCNEC+, they range between 0.05 and 4.16 phw (outlier excluded). This indicates that there is a **large inter-speaker variability in each corpus**. Levene's test for equality of variances further indicates that the variability is larger in the learner group than in the native group ($p < .000$).

A statistically significant **comparison of means** shows that **learners produce more repetitions**, on average, than native speakers ($p < .000$; $t = 7.07$); a post-hoc Cohen's *d* test

further indicates that the **effect size is very large** (d = 1.16). This result is in line with previous studies (Götz 2013a; Rohr 2017).

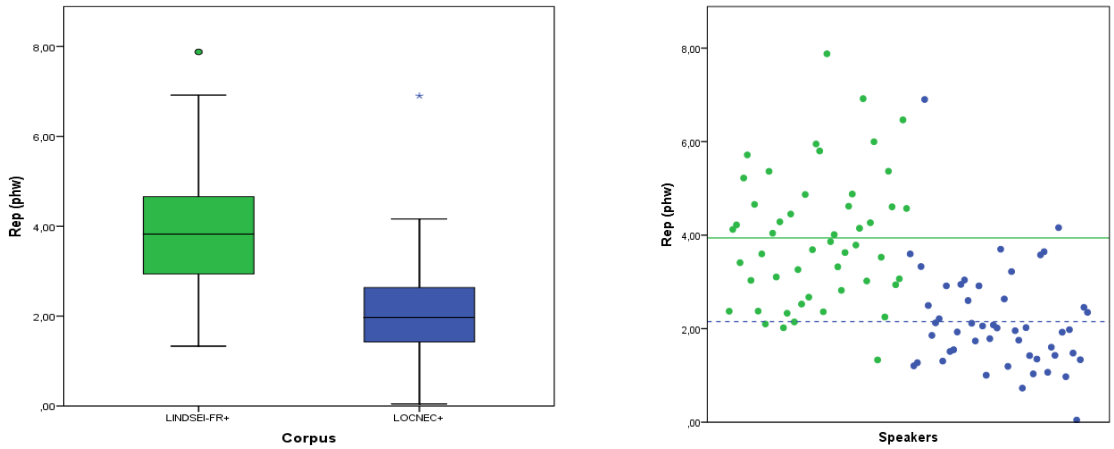


Figure 5-17: Boxplots and stripchart of repetitions in LINDSEI-FR+ and LOCNEC+

The two corpora contain repetitions of varying spans: a speaker can repeat him- or herself once, twice, or more. Previous research indicates that single repetitions (i.e. with a single *repeated*) are more frequent than double, triple or even longer repetitions (i.e. with two, three, or more *repeated*) (Biber *et al.* 1999; Gráf 2017). The figures from Table 5-5 strongly support these results **Single repetitions** constitute the **overwhelming majority of repetitions in the two corpora**: in LINDSEI-FR+, 92.2% of the repetitions are single repetitions, which is nearly the same proportion as in LOCNEC+ (93.4%). **Multiple repetitions account for less than 10% of the repetitions in each corpus** – 7.8% in LINDSEI-FR+ and 6.6% in LOCNEC+. Some typical simple and multiple repetitions from the learner and the native corpus are shown in 5-36 to 5-38, respectively. In the two corpora, most multiple repetitions are **double repetitions** (i.e. where there are two *repeated*, as in 5-38 for LOCNEC+), and only a very small proportion are longer repetitions. In LINDSEI-FR+, there is one instance of a quintuple repetition (see 5-39), while there are three in the native corpus.

	LINDSEI-FR+	**LOCNEC+**
Single repetitions	92.24% (3,457)	93.42% (2,569)
Multiple repetitions	7.76% (291)	6.58% (181)
Double repetitions	*6.86% (257)*	*5.93% (163)*
Triple repetitions	*0.75% (28)*	*0.40% (11)*
Quadruple repetitions	*0.13% (5)*	*0.14% (4)*
Quintuple repetitions	*0.02% (1)*	*0.11% (3)*
Totals	100% (3,748)	100% (2,750)

Table 5-5: Proportion (absolute frequency) of simple and multiple repetitions in LINDSEI-FR+ and LOCNEC+

5-36: FR042-F - simple repetitions

```
time to: (0.240) to really do (0.230) fieldwork for example this is this is rather
impossible
```

5-37: EN001-F - simple repetition

```
er I I've never really had that problem
```

5-38: EN046-S - (simple and) double repetition

```
well the problem was (0.280) was that (0.410) the the the contract that I'd signed
```

5-39: FR025-S - quintuple repetition

```
B: eh human rights are (0.650) also erm (0.380) how do you say that (0.250) eh
A: violated
B: violated in in in in in in Europe as well
```

As could already be seen from the previous examples, repetitions may invole the re-iteration of one or more words. Table 5-6 reveals the breakdown of the number of words in the initial utterance in the repetition. It appears clearly that the **repetition of single words is predominant both in LINDSEI-FR+ and in LOCNEC+**. In the learner corpus, 82.5% of the repetitions are one-word repetitions and 14.09% are two-word repetitions. Repetitions of three words or more account for less than 3.5% of the learner repetitions. Similarly, in the native corpus, the proportion of repetitions sharply decreases as the number of words in the *repeatable* increases. The results so far thus strongly support Biber *et al.*'s (1999:1055) findings that "the likelihood of the repetition decreases sharply with the number of words repeated", and are also comparable with previous findings (Gráf 2017; Kapatsinski 2004).

No. of repeated words	**LINDSEI-FR+**	**LOCNEC+**
1 word	82.50% (3,092)	67.85% (1,866)
2 words	14.09% (528)	24.40% (671)
3 words	2.59% (97)	5.71% (157)
4 words	0.72% (27)	1.46% (40)
5 words	0.08% (3)	0.47% (13)
6 words	0.00% (0)	0.07% (2)
7 words	0.02% (1)	0.00% (0)
8 words	0.00% (0)	0.04% (1)
Totals	**100% (3,748)**	**100% (2,750)**

Table 5-6: The number of words in Ro in LINDSEI-FR+ and in LOCNEC+

Quite surprisingly, however, **compared to French learners, native speakers from LOCNEC+ use a lower proportion of one-word repetitions** (67.85%), **and a higher proportion of longer repeatables**. For example, native speakers use about 10% more two-word repetitions than learners (24.40% vs. 14.09%). From the data, it appears that these two-word repetitions

generally correspond to the sequence subject + verb (such as *it took, I did, she lives, there was* etc.), and a small number of two-word repetitions correspond to a sequence of two conjunctions (especially *and then*) or more eclectic sequences such as *not much, in the,* or *all the*. The fact that native speakers use proportionally more two-word repetitions than learners is rather unexpected, but could perhaps be explained by the native speakers' renowned preference for building discourse on the basis of "routinised building blocks" or "automatised chunks of words", which are argued to be processed as one single unit (see e.g. Chambers 1997; De Cock 2004; Towell, Hawkins & Bazergui 1996). Learners, who may not have integrated such sequences to the same extent, would thus be more likely to repeat single words over sequences of words.

No. of repeated words	**No. of repetitions**	**LINDSEI-FR+**	**LOCNEC+**
1	Single	75.21% (2,819)	62.36% (1,715)
	Double	6.43% (241)	4.91% (135)
	Triple	0.69% (26)	0.40% (11)
	Quadruple	0.13% (5)	0.07% (2)
	Quintuple	0.03% (1)	0.11% (3)
2	Single	13.61% (510)	23.49% (646)
	Double	0.43% (16)	0.84% (23)
	Triple	0.05% (2)	0.00% (0)
	Quadruple	0.00% (0)	0.07% (2)
3	Single	2.59% (97)	5.56% (153)
	Double	0.00% (0)	0.15% (4)
4	Single	0.72% (27)	1.42% (39)
	Double	0.00% (0)	0.04% (1)
5	Single	0.08% (3)	0.47% (13)
6	Single	0.00% (0)	0.07% (2)
7	Single	0.03% (1)	0.00% (0)
8	Single	0.00% (0)	0.04% (1)
Totals		**100.00% (3,748)**	**100.00% (2,750)**

Table 5-7: The relationship between number of words in Ro and number of repeated in LINDSEI-FR+ and LOCNEC+

To examine the relationship between the number of repeated words and the empan of repetitions (i.e. the number of repetitions), I crossed the data from Table 5-5 and Table 5-6. The result is displayed in Table 5-7 below. It is striking to see that when single words are repeated (number of repeated words = 1), they can be repeated once, twice, and up to five times in a row in the two corpora. **As the number of repeated words increases,** however, **the number of repetitions clearly decreases**, and double, triple or quadruple repetitions become the exception. Overall, the results thus highlight a negative relationship between the number of repeated words and the number of repetitions: as one increases, the other decreases. This result was to be expected given the fact that, from a cognitive point of view,

single words, or short sequences of words, are more easily kept activated in the short-term memory than longer sequences: the longer the *repeatable*, the higher the cognitive load, and the less likely it becomes that this sequence of words can be repeated more than once.

Examples 5-40 to 5-42 below illustrate the typical link between length of the repeatable and length of the repeated.

5-40: EN038-F - quintuple repetition of one word

I went into this pub and **the the the the the the** Dutch owner who owned it

5-41: EN005-F - quadruple repetition of two words

it was it was it was it was it was a really good trip

5-42: EN005-S - double repetition of three words

erm **I don't I don't I don't** really get the chance to see her

Henry and Pallaud (2004) rightly underlined in their article that the ***repeatable* and the *repeated* are not always in contiguous position**: one or more elements may actually be inserted between the original utterance and the repetition. This potential material between the initial utterance and the repeated is also sometimes referred to as the "hiatus" (Clark & Wasow 1998).

I examined, in LINDSEI-FR+ and in LOCNEC+, the proportion of repetitions where the *repeatable* and the *repeated* are in adjacent position vs. when they are split up. To obtain the figures, I examined whether there was an <N> tag in the annotation tier [anno-2], which corresponds to hiatuses.

	LINDSEI-FR+	**LOCNEC+**
Direct repetitions (<R0 R*n*>)	58.96% (2,210)	67.38% (1,853)
Indirect repetitions (<R0 <*tag*> R*n*>)	39.01% (1,462)	31.64% (870)
Others	2.03% (76)	0.98% (27)
Totals	**100% (3,748)**	**100% (2,750)**

Table 5-8: The proportion (absolute frequency) of direct and indirect repetitions in LINDSEI-FR+ and LOCNEC+

Table 5-8 displays the proportion of direct and indirect repetitions in the two corpora. It appears that **direct repetitions** (i.e. repetitions where the repeatable and the repeated are immediately adjacent) **form the majority of native and learner repetitions**. In LINDSEI-FR+, c. 59% of the repetitions have an empty hiatus. In LOCNEC+, this proportion is even higher as c. 67% of the repetitions are classified as direct repetitions. Illustrations of typical direct repetitions can be seen in 5-43 and 5-44. Indirect repetitions, where the repeatable and the

repeated are separated by a hiatus, account for about 40% of the occurrences in LINDSEI-FR+ and a third of the occurrences in LOCNEC+. **The hiatus typically consists in a pause, filled or unfilled, in a discourse marker, or in a combination of those**, as illustrated in 5-45 through 5-47.

5-43: FR019-S - direct repetition

there was only one actor who **was was** not that good

5-44: EN001-F - two direct repetitions

I I dropped it **rather than rather than** do it as a minor

5-45: FR036-P - indirect repetition with an unfilled pause as hiatus

and er **she** (0.260) **she** shows erm the painting

5-46: EN046-F - indirect repetition with an unfilled pause and a discourse marker as hiatus

New England is just supposed to be glorious with **all the** (0.430) you know **all the** trees

5-47: FR019-S - indirect repetition with a filled and an unfilled pause as hiatus

I'll erm (0.490) **I'll** try to interview two different cases

An interesting finding which has not been reported in previous literature is that, while the hiatus typically occurs between the original utterance and the *repeated*, it may also be found either **within a constituent, or between the iterations of the *repeated***. These cases account for 1 to 2 percent of all L1 and L2 repetitions. In Example 5-48, where the hiatus is placed just before the final iteration of the repetition, it seems as if the learner tries to make a fresh, fluent start after a single repetition. The same phenomenon also occurs in LOCNEC+ (Example 5-49). In example 5-50, however, the unfilled pause in hiatus is placed in the middle of the *repeated*. This type of hiatus is only attested in LINDSEI-FR+ (although the native speakers tend to use more multiple-word repetitions than learners).

5-48: FR004-S - the hiatus

they they (1.120) **they** picked me up

5-49: EN029-F - the hiatus

I I (1.130) **I** had to watch videos of television

5-50: FR023-S - the hiatus

this was the this (0.380) **was the** first time I went by plane

5.3.5 Vowel lengthenings

Vowel lengthenings (Ls) were already marked in the original transcriptions of LINDSEI-FR and LOCNEC. In order to obtain all instances of lengthened words, I extracted the tag <L> in the aligned and annotated corpora, which amounts to 2,909 occurrences of lengthenings in LINDSEI-FR+ and 1,026 in LOCNEC+.

The corpus data reveal that, on average, **the learners** in LINDSEI-FR+ produce **3.11 vowel lengthenings per hundred words**, i.e. 1 lengthening every c. 30 words. The mean for the **native speakers** in LOCNEC+ is quite lower, with a mean of **0.77 phw**, or barely one lengthening every c. 130 words. The average for the NSs is thus more than 3 times lower than that of the learners. Figure 5-18, which provides a visual representation of the descriptive statistics in the two corpora, shows that it is not only the L1 mean that is lower than the learners' mean: in fact, nearly all the native speakers produce fewer vowel lengthenings per hundred words than the learners who produces the least. It is also striking to see that the dispersion of the learner data is far bigger than the dispersion of the native speaker data, as confirmed by Levene's test ($p < .000$).

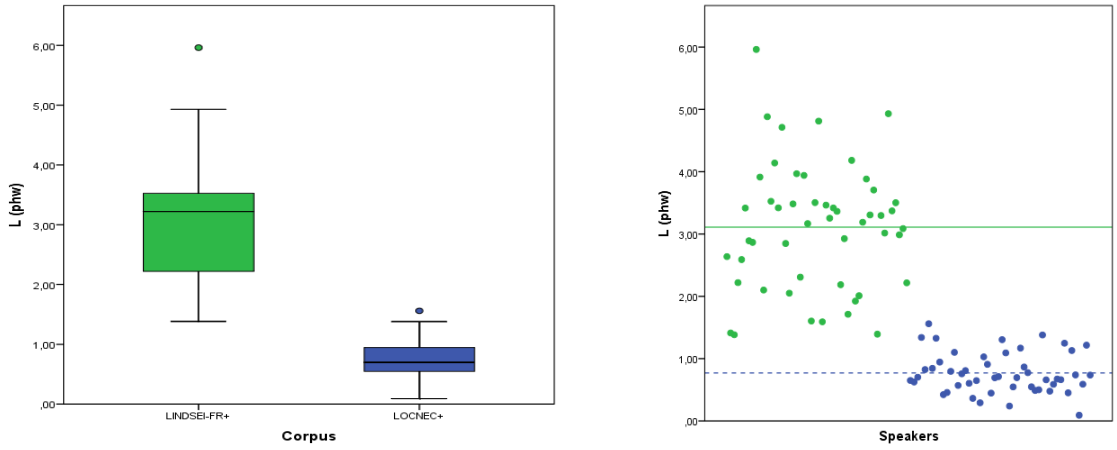


Figure 5-18: Boxplots and stripchart of vowel lengthenings phw in LINDSEI-FR+ and LOCNEC+

A *t*-test on the mean frequency of lengthenings was statistically significant ($t = 15.58$; $p < .000$), revealing that, **on average, learners in LINDSEI-FR+ lengthen more vowels than the native speakers from LOCNEC+**. The significance of the *t*-test is further strengthened by a very large effect size for the mean difference (Cohen's $d = 3.117$).

These results can be interpreted as a clear case of **overuse on the part of the learners**, and they support the claim that non-native speakers, who are more often confronted to planning problems, regularly resort to lengthenings when they need to gain some additional planning time (Clark & Fox Tree 2002; Rohr 2017). However, it ought to be underlined that there might also be an element of **L1 transfer** that could partly account for the large difference between French learners and native speakers. Duez (2001) explains that French and English differ (among others) with respect to the proportion of lengthenings because "English, which is a

closed syllable language, uses longer syllables less often than French, which is a CV-type language" (*ibid.*:118; my translation).

In addition to investigating the extent to which lengthenings are used in native and learner speech, I also looked at the **nature of the words (POS) that are most frequently lengthened**. The five most frequent POS (automatically generated from TreeTagger) are displayed in Table 5-9. It appears that **determiners are particularly prone to lengthening**, especially in LOCNEC+: as they account for c. 38% of the lengthenings in LINDSEI-FR+ and c. 74% in LOCNEC+. Most of these determiners actually correspond to either *the:* or *a:* (*cf.* also Table 5-10). The POS "**TO**" (which corresponds to the word form *to*) is the second most frequently lengthened POS in LINDSEI-FR+, followed by **personal pronouns, adverbs and prepositions**. In the native corpus, **conjunctions** take the fourth place (instead of adverbs in the NNS corpus), but the other categories are the same. Interestingly, whereas in LOCNEC+, there is a sharp divide in terms of frequency between the top 1 POS (determiners) and the other POS, it is much less so in LINDSEI-FR+.

POS (TreeTagger)	**LINDSEI-FR+**
DT – determiner	37.81% (1100)
TO – *to*	19.90% (579)
PP – personal pronoun	15.57% (453)
RB – adverb	6.18% (180)
IN – preposition	5.08% (148)
Others	*15.46% (449)*

POS (TreeTagger)	**LOCNEC+**
DT – determiner	73.97% (759)
TO – *to*	8.38% (86)
PP – personal pronoun	5.36% (55)
CC - connector	4.09% (42)
IN – preposition	2.73% (28)
Others	*5.47% (56)*

Table 5-9: The top 5 POS of the lengthened words in LINDSEI-FR+ and LOCNEC+ (proportion and absolute frequency)

Going a step further, Table 5-10 reveals the breakdown of the top 10 most frequently lengthened words in LINDSEI-FR+ and LOCNEC+. The data reveal that **the 3 most frequently lengthened words** – *the:*, *to:* and *a:* – **are the same in the native and the learner corpora**. However, whereas the form *the:* accounts for c. 27% of the lengthened words in LINDSEI-FR+, it totals more than two thirds of the lengthenings in LOCNEC+. This difference can partly explain the different frequencies of the POS DT (determiner) in LINDSEI-FR+ and LOCNEC+ from Table 5-9. The other forms in LOCNEC+ only account for a small proportion of the occurrences of lengthenings (< 10% each). By contrast, in the learner corpus, the lengthened word *to:* is nearly as frequent as *the:*, totalling 617 instances (c. 21%). It is also interesting to see that in both corpora, five personal pronouns (*we:*, *she:*, *I:*, *they:*, *you:*) stand among the top 10 both in NS and NNS speech (*cf.* also Table 5-9, where personal pronouns come third). Two illustrations of typical lengthened words are provided in 5-51 for LINDSEI-FR+ and in 5-52 for LOCNEC+.

Lengthened word	LINDSEI-FR+
the:[128]	27.63% (804)
to:	21.21% (617)
a:	9.35% (272)
we:	3.71% (108)
she:	3.16% (92)
I:	3.06% (89)
very:	2.44% (71)
they:	2.16% (63)
and:	2.09% (61)
you:	2.02% (59)

Lengthened word	LOCNEC+
the:	67.64% (694)
to:	8.38% (86)
a:	6.14% (63)
and:	3.99% (41)
so:	2.82% (29)
she:	1.85% (19)
I:	1.07% (11)
we:	0.88% (9)
they:	0.88% (9)
you:	0.68% (7)

Table 5-10: The top 10 lengthened words in LINDSEI-FR+ and LOCNEC+ (proportion and absolute frequency)

5-51: FR026-F - lengthening of a determiner (L used alone)

I (0.210) erm I go to **a:** dance course

5-52: EN006-P - lengthening of a personal pronoun (L used alone)

the contrast to (0.430) the second picture where **she:** looks like she hates the picture

It appears from Table 5-11 that, in addition to using on average more lengthenings than native speakers, the **learners from LINDSEI-FR+ also use lengthenings in a different context than native speakers**. Native speakers of English use about half of the lengthenings alone (i.e. without combining it with other adjacent (dis)fluency features), and half of the lengthenings in clusters of adjacent (dis)fluency features. By contrast, only c. 18% of the learner lengthenings are used alone and the **majority (c. 81%) are used in clusters**. This may be a further indication that learners need to combine more than one device to gain sufficient time to plan what to say next or how to cope with a planning problem. Note, however, that the results that Duez (2001) obtained for **native speakers of French** are situated in between the figures of our NNS and NS: in her data, she found that 65% of the lengthened syllables are combined with other (dis)fluency features. Illustrations of lengthenings used alone or in clusters are provided in 5-51 and 5-52, and 5-53 and 5-54, respectively.

[128] Note that in LINDSEI-FR+ and LOCNEC+, no difference is made between the weak and the strong form of the determiners *the* (*the:* vs *the[i:]*) and *a* (*a:* vs *a[ei]*) and I am aware that this might constitute a caveat in the data, though this does not impact the comparability of the data from Table 5-10.

	LINDSEI-FR+	LOCNEC+
Stand-alone use of Ls	18.56% (540)	43.66% (448)
Ls used in clusters	81.44% (2,369)	56.34% (578)
Totals	100% (2,909)	100% (1,026)

Table 5-11: Proportion (absolute frequency) of Ls used alone and in combination with other (dis)fluency features in LINDSEI-FR+ and LOCNEC+

5-53: FR025-S - lengthening of a personal pronoun (L used conjointly with a repetition and an UP)

we can be glad about what **we:** (0.860) what we have achieved

5-54: EN042-P - lengthening of a determiner (L used conjointly with an UP, a FP and a repetition)

he er (0.200) he shows (0.660) **the:** er the lady the portrait and she obviously isn't very happy

5-55: EN004-S - lengthening of 'to' (L used conjointly with an UP and a FP)

when I was (0.320) say about fourteen I I used to want **to:** (1.190) er open a computer shop

Examples 5-53 and 5-54 are very typical of the use of lengthenings in clusters: the **combination of lengthenings with repetitions and/or filled and unfilled pauses is very frequent in the corpora**. For instance, 20% of the lengthenings in LINDSEI-FR+ are either used in the cluster L+Rep, or L+Rep+UP. Incidentally, these types of clusters do not entirely support Clark and Fox Tree's (2002:86) claim that lengthenings do not signal the initiation of delays, but rather the *continuation of ongoing delays*. Although it is true that lengthenings often appear in the middle of a (dis)fluency cluster, they also often appear at their beginning, as in 5-55.

All in all, the investigation of learner and native vowel lengthenings revealed that learners use about three times as many lengthenings as NSs. Although the nature of the words that are lengthened is similar in the two corpora, LOCNEC+ speakers favour the lengthening of the determiner *the* over all other word forms. This is not the case for LINDSEI-FR+ speakers. Lastly, learners and native speakers also differ with respect to their propensity of associating lengthenings with other (dis)fluency features: whereas the proportion of lengthenings used alone and in clusters is nearly equal for the NSs, only a minority of the lengthenings are used alone in LINDSEI-FR+.

5.3.6 Discourse markers

Discourse markers were extracted by querying the tag <DM. In LINDSEI-FR+, 2,016 occurrences of discourse markers were retrieved, against 3,213 in LOCNEC+.

Discourse markers have a mean frequency of **2.04 phw in LINDSEI-FR+**, which means that learners utter one discourse marker every 50 words on average. In **LOCNEC+**, the native speaker mean is slightly higher, with **2.42 DMs phw**. As can be observed in Figure 5-19, there is also **considerable variability both in the learner and in the native data**. In learner speech, individual means range from 0.33 to 6.26 DMs phw, that is, nearly twenty times as many. Likewise, in the native data, some speakers use very few discourse markers (min. 0.14 DM phw) and some use many of them (max. 6.11 DMs phw). The variability in the frequency of discourse markers is rarely emphasised in previous research, but these initial results suggest that it might be important for future studies of DM use in both L1 and L2 to combine pooled and individual data.

An independent samples *t*-test reveals that, although there is a tendency for native speakers to use more discourse markers than learners, the difference between the learner and native means is **not significant**. This result diverges from previous studies, but might simply be due to the different range of discourse markers considered (*cf*. below).

A number of discourse markers can be used to signal hesitation, as exemplified in 5-56 with *well*. In this example, the learner seems to hesitate about the use of *postpone*, as suggested by the filled pause preceding the word. She then uses the DM *well* to introduce a reformulation (*to take another one*). Another example includes the use of *you know* or *I mean* (Examples 5-57 and 5-58). *You know* may function as "a speaker appeal for hearer cooperation in a discourse task" (Schiffrin 1987:63) and *I mean* as a forewarning of upcoming adjustments (*ibidem*).

5-56: FR008-F - DM 'well' with restart

> someone help= helped me (0.300) to er postpose my plane (0.090) **well** to take another one

5-57: FR025-P - DM 'you know'

> she asks him erm (0.300) to do it again (0.500) and to (1.040) **you know** erm improve eh the reality <laughing/>

5-58: EN044-F - DM 'you know' and 'I mean'

> the university's on a main street (0.330) **you know I mean** this (0.200) it has (0.380) **you know** (0.300) it's surrounded by the fields

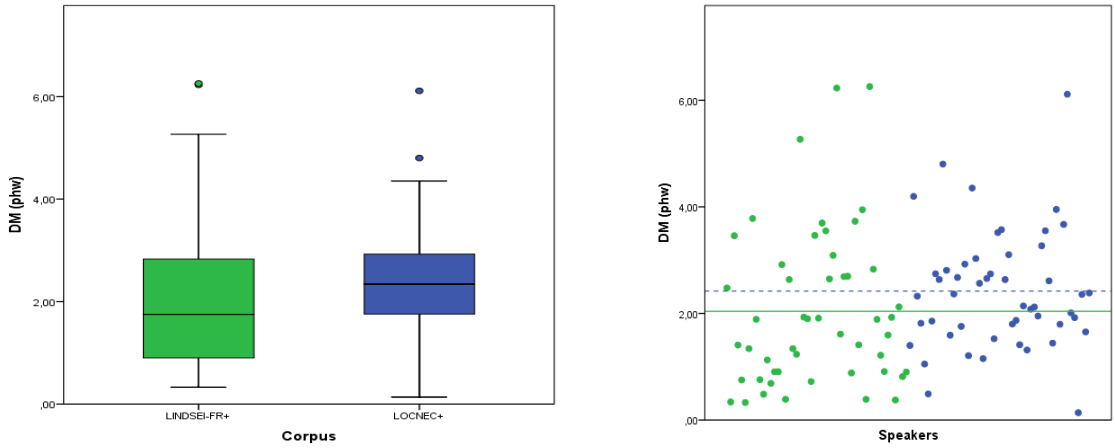


Figure 5-19: Boxplots and stripchart of discourse markers phw in LINDSEI-FR+ and in LOCNEC+

Discourse markers in LINDSEI-FR and LOCNEC have already been thoroughly analysed in several studies[129]. Gilquin (2008:120), for example, investigated "a representative selection of smallwords", including *kind of*, *well*, *you know*, and *I mean*. Gilquin and Granger (2015) examined the following two-word discourse markers in several components of LINDSEI (including LINDSEI-FR) and in LOCNEC: *and so*, *and then*, *I mean*, *in fact*, *sort of*, and *you know*. In a later study, Gilquin (2016) further investigated seven discourse markers in LINDSEI and LOCNEC, namely *and so*, *and then*, *I mean*, *like*, *sort of*, *well*, and *you know*. In those studies, both quantitative and qualitative aspects have been discussed. The remainder of this section is consequently restricted to the contextual use (1) of discourse markers in general and (2) of the five most frequent discourse markers as identified in the previously cited studies, i.e. *I mean*, *in fact*, *you know*, *well*, and *like*.

Table 5-12 displays the proportion of discourse markers that occur alone and the proportion of discourse markers that occur in clusters of adjacent (dis)fluency features. In both LINDSEI-FR+ and LOCNEC+, the **majority of discourse markers are used in clusters**. However, while learners use a relatively modest proportion of DMs alone (21%), stand-alone DMs seem far more common in native speech: 41% of discourse markers are used alone. This suggests that learners, who tend to use fewer DMs on average, might use them in more (dis)fluent contexts than native speakers.

	LINDSEI-FR+	**LOCNEC+**
Stand-alone DMs	21.14% (430)	41.37% (1,332)
DMs in clusters	78.86% (1,604)	59.63% (1,888)
Totals	**100% (2,034)**	**100% (3,220)**

Table 5-12: Proportion (absolute frequency) of DMs used alone and in clusters in LINDSEI-FR+ and LOCNEC+

[129] See also Crible (2017a; 2018; forthcoming) for a large-scale and in-depth study of discourse markers in native French and native English.

5-59: ENo02-F - stand-alone use of 'like'

I'd really like to be able to do it cos **like** I think it's a really food advantage

Zooming in on *you know*, *I mean*, *in fact*, *like*, and *well*, the same tendency can also be observed (Figure 5-20). For example, learners use a greater proportion of *you know*s in clusters than alone (77% vs. 23%), and the proportion of L2 *you know*s in clusters is higher than that of native speakers (77% in LINDSEI-FR+ vs. 61% in LOCNEC+). Interestingly, contrary to the other native DMs which occur preferably in clusters, ***like*** is used slightly more often alone than in clusters, as illustrated in 5-59.

When they are used in clusters, the discourse markers *you know*, *I mean*, *in fact*, *like*, and *well* are most frequently **combined with unfilled pauses, either in front or back position** (UP+DM and DM+UP). Some interesting differences can, however, be noted between L1 and L2 speakers. Whereas learner *you know*s occur equally frequently in the pattern UP+DM and DM+UP, native speakers have a clear preference for the pattern UP+*you know*. The opposite is true for *I mean*: learners use this DM preferably with an unfilled pause preceding it (UP+*I mean*), but NSs use the two patterns UP+DM and DM+UP equally frequently. As for *well*, learners tend to use it more frequently in the pattern UP+*well*, but NSs have a marked tendency to use the other pattern (*well*+UP). Two examples of the combination of discourse markers with unfilled pauses are provided below (5-60 and 5-61). Other recurrent clusters involving discourse markers include the patterns **C+DM** (Examples 5-62 and 5-63) and **FP+DM** (see Example 5-64).

5-60: FR008-F - UP+DM

it was (0.780) incredible (0.260) **I mean** people (0.380) are always thinking about other people

5-61: FR040-S - DM+UP

but it's (0.050) on the street **you know** (0.300) people throwing everything or just going to the river

5-62: FR033-F - C+DM

some of them were homosexual (0.670) erm others not but **well** the book was (0.290) more or less well received

5-63: ENo02-F - C+DM

we tried to learn French and **like** it's just (0.540) I didn't like it

5-64: ENo12-F – FP+DM

it's all right it's erm **well** different from la= last year

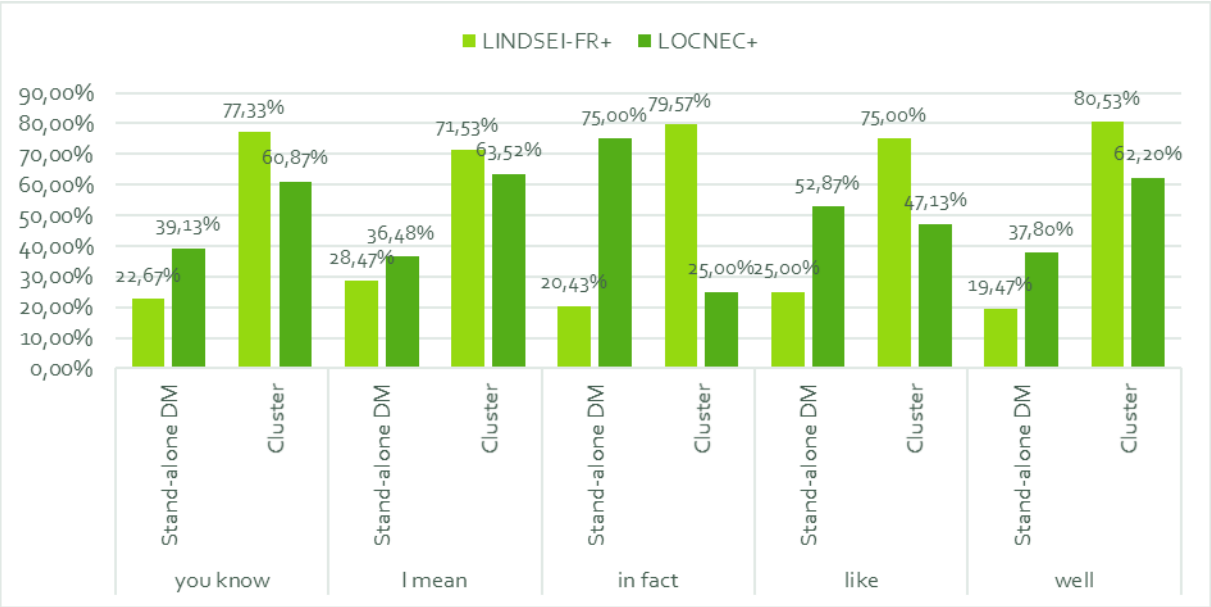


Figure 5-20: Proportion of 'you know', 'I mean', 'in fact', 'like', and 'well' used alone or in clusters in LINDSEI-FR+ and in LOCNEC+

5.3.7 Restarts

In LINDSEI-FR+ and LOCNEC+, restarts (RSs) were extracted by querying the tag <RS. In the learner corpus, 1,775 restarts were extracted, against 1,651 in LOCNEC+. Several characteristics of restarts (substitution, insertion etc.) were also annotated in the annotation tier ANNO-2, and will be analysed below.

The corpus data show that, on average, LINDSEI-FR+ **learners** produce **1.84 restart per hundred words**, that is, one restart every c. 54 words. This mean is comparable to that reported by De Jong *et al.* (2012a) for intermediate learners of Dutch, i.e. 1.6 phw. In LOCNEC+, the mean number of restarts is slightly lower and amounts to **1.25 RS phw** (1 restart every c. 80 words). These means suggest that restarts are rather infrequent in speech. However, as can be observed in Figure 5-21, the dispersion is quite large in the learner data, with individual means ranging from 0.56 to 3 .22 RSs phw (the outlier, FR028, excluded), as well as in LOCNEC+, where the lowest frequency is 0.23 and the highest is 1.89 RSs (EN038 and EN004, the 2 outliers, excluded).

An independent samples *t*-test reveals that **learners produce significantly more restarts per hundred words than native speakers** (t = 5.44; p < .000). Moreover, the effect size of this difference is large (d = 1.12). This finding is in line with previous findings (Kahng 2014; Rohr 2017).

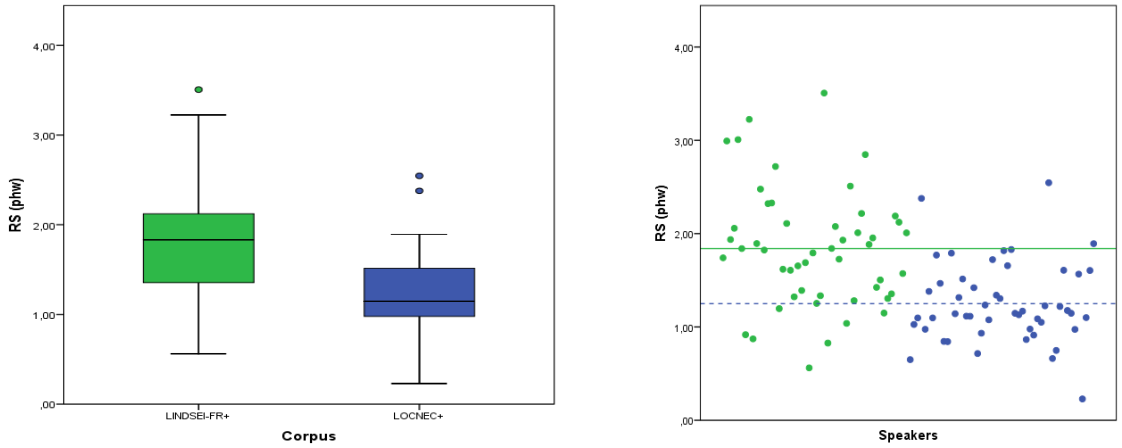


Figure 5-21: Boxplots and stripchart ofr restarts phw in LINDSEI-FR+ and LOCNEC+

Several characteristics of restarts were annotated in the annotation tier [anno-2]. For example, it was specified whether the restart includes an insertion or a substitution. The proportion of restarts that contain a substitution, an insertion, a deletion, or a word re-ordering is displayed in Table 5-13.

	LINDSEI-FR+	**LOCNEC+**
Propositional substitution	40.00% (710)	40.64% (671)
Insertion	19.77% (351)	30.04% (496)
Morpho-syntactic substitution	18.08% (321)	11.08% (183)
Restart after truncation	15.66% (278)	13.20% (218)
Deletion	9.75% (173)	13.26% (219)
Ordering	0.68% (12)	0.79% (13)
Other	1.75% (31)	3.15% (52)

Table 5-13: Proportion (absolute frequency) of sub-categories of restarts in LINDSEI-FR+ and LOCNEC+

Forty percent of the restarts in both LINDSEI-FR+ and LOCNEC+ contain a **propositional substitution**. In 5-65, for example, the learner substitutes the word *reply*, transferred from French, by *answer*, which is more appropriate in this context. It is also interesting to note that the learner pauses very briefly before the word to be substituted, and introduces the substitution with a filled pause (*cf*. Levelt's (1983) main interruption rule). In Example 5-66, the native speaker likewise pauses (*erm*) before the prepositional phrase that is later replaced (i.e. the reparandum *on the site*). The substitution immediately follows the reparandum and starts with the re-utterance of the beginning of the constituent (*we stayed*).

5-65: FR010-F - restart with propositional substitution[130]

if people come and (0.040) reply things er ***answer*** **things** in in

5-66: EN003-F - restart with propositional substitution

we stayed erm on the site **we stayed *in the camp*** of David Crockett

Insertions are included in 20% and 30% of the learner and native restarts, respectively. As can be observed in Example 5-67, an insertion can **add propositional content to the utterance** (although, in this example, the insertion (*whites*) is grammatically incorrect). Likewise, native speakers may also restart a constituent to add some detail: in 5-68, the speaker EN049 restarts the noun phrase a river by inserting an adjective (*a chocolate river*). Note that, in those two examples (and as was the case in 5-65 and 5-66), the reparandums (*marks* and *a river*) are immediately preceded by a (dis)fluency feature: an unfilled pause and the repetition of the determiner *a*. These features arguably already signal the forthcoming restarts. Alternatively, speakers (especially learners) may restart and use an insertion to **correct an ungrammatical utterance** (*find job*/*find a job* in 5-69). Lastly, restarts may include a **parenthesis**. In 5-70, for example, the speaker inserts a long parenthesis before restarting her utterance. The end of the insertion and the beginning of the actual restart are marked (or linked) by an unfilled pause and the conjunction *and*.

5-67: FR013-S - restart with insertion

there are no (0.410) marks on the roads ***whites*** **marks**

5-68: EN049-F - restart with insertion

there's a a a river **a *chocolate*** **river** in this massive big room

5-69: FR014-F - restart with insertion

it depends on er (0.340) where I can find job ***a*** **job**

5-70: EN046-S - restart with insertion

luckily I ma= I made good friends with one of the ***cos there was five of us living in this house*** **(0.350)** **and** **I made good friends with one of the** girls

Morpho-syntactic substitutions can be found in c. 18% of the learner restarts, and in 11% of the native restarts. Such substitutions generally pertain to verbal forms (5-71 to 5-73), but not exlusively (*cf*. Examples 5-74 and 5-75).

5-71: FR020-S - restart with morpho-syntactic substitution

there are (0.510) less people who (0.250) who believes in it **who *believe*** **in it**

[130] In the following examples, the restart is shown in bold font. The insertion, substitution etc. is in italics. Other (dis)fluency features that are relevant for the interpretation of the example are underlined.

5-72: FR035-S - restart with morpho-syntactic substitution

we find it **we *found* it** em

5-73: EN049-F - restart with morpho-syntactic substitution

he said ***was saying*** (0.410) they're not gonna give me my son back

5-74: FR026-P - restart with morpho-syntactic substitution

to draw (0.480) erm (0.380) a a women **a *woman*** (0.580) portrait

5-75: EN050-S - restart with morpho-syntactic substitution

there were certain objective ***objectives*** that we wanted to (0.180) complete

Additionally, speakers also regularly restart their utterance **after a truncation**. In the learner corpus, about 16% of the restarts follow a truncation, and in native speech, 13% of the restarts can be found after a truncated word. More details on the joint use of restarts and truncations can be found in Section 5.3.8.

Some restarts include the **deletion** of some material. Interestingly, there are more such cases in native than in learner speech (13.26% vs. 9.75%). In Example 5-76, the learner restarts the prepositional phrase, but without the determiner *the*. Similarly, in 5-77, the native speaker restarts his utterance after two pauses and omits the adjective *little*.

5-76: FR005-S - restart with deletion

I went to *the:* **to** movies

5-77: EN050-S - restart with deletion

it was like *little* er (1.310) **it was like** dried stuff

In exceptional cases, restarts involve a **change in word order**, as illustrated in the following two examples. It is worth underlining that the learner in 5-78 produces three pauses before his initial utterance (*eh erm (1.550)*), which suggests that he might already be aware of the issue with word order. The restart with the new word order is introduced by an unfilled pause and a truncation. Finally, the learner apologises for the mistake (*sorry*). An example from LOCNEC+ is shown in 5-79.

5-78: FR022-F - restart with word re-ordering

I play badminton and eh erm (1.550) tennis table (0.330) ta= ***table tennis*** <laughing/> sorry

5-79: EN022-S - restart with word re-ordering

but er it was a really **it was *really a*** good film

Lastly, the "other" category includes restarts due to articulation issues, such as in the following example (5-80):

5-80: FR038-F - restart

er will probably be er something (0.250) er on Austrainan **Australian** literature

Before concluding this section, it is important to underline that one restart can involve more than one of the characteristics outlined above. Consider example 5-81, where the insertion of the adjective *Irish* causes the substitution of *a* by *an*. Restarts may also follow one another, as in 5-82, where *somebody* is first replaced by *we*, and *all* is then inserted in the second restart.

5-81: FR005-S - restart

eh a friend of mine ***an Irish*** **friend of mine** (0.340) and me we decided to to dance rock 'n 'roll

5-82: EN050-S - restart

somebody saw ***we*** **saw** <unknown/> **we** ***all*** **saw** some <name/> muskox

To conclude this section, restarts are significantly more frequent in learner than in native speech; there is, however, considerable overlap between the L1 ad L2 distributions. Six main subtypes of restarts have been illustrated. While restarts with propositional substitutions are nearly as frequent in learner than in native speech, the figures indicate that learners tend to use fewer restarts with insertions, and more restarts with morpho-syntactic substitutions than L1 speakers. The categories identified above could be related to Levelt's (1983) appropriateness-adjustment and error-correction repairs. At first sight, however, it seems that there is no perfect correspondence between Levelt's two categories and the sub-categories of restarts identified here: insertions, for example, might be considered to correspond to either adjustment repairs (as in 5-67) or error repairs (as in 5-69).

5.3.8 Truncated words

Truncated words (Ts) were extracted from the corpora by querying the tag <T, which resulted in 2,451 hits. There are 1,614 truncations in LINDSEI-FR+, and 837 in LOCNEC+.

While the French-speaking learners utter, on average, 1.64 truncations per hundred words, the native speakers produce only 0.63 truncation phw. The native mean is very close to the mean reported for French native speakers (Pallaud 2002; Henry & Pallaud 2004). The **learners thus use over twice as many truncations per hundred words than native speakers**.

As can be observed in Figure 5-22, the dispersion of the data is quite large, especially in LINDSEI-FR+, where the learner who produces the most truncations (3.07 phw, outlier excluded) utters nearly 5 times as many truncations as the learner who utters the least (0.44 phw). In LOCNEC+, the dispersion of the data is less important, with a maximum of 1.11 truncations per hundred words (outliers excluded). Levene's test of equality of variances indicates that there is **more variability in the learner group** than in native speech ($F = 9.91$; $p < .000$).

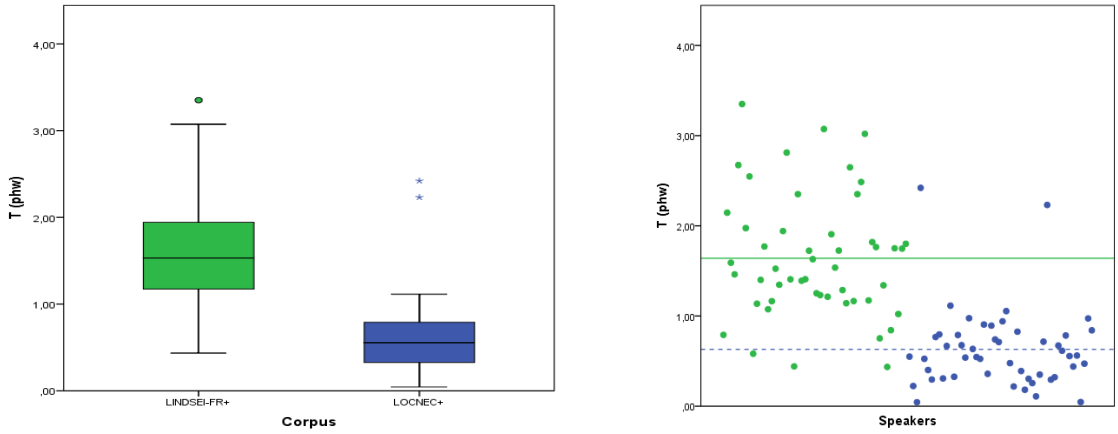


Figure 5-22: Boxplots and stripchart of T phw in LINDSEI-FR+ and LOCNEC+

A two samples *t*-test reveals that the difference between the means in LINDSEI-FR+ and LOCNEC+ is very **significant** (t = 8.86; p < .000) and that the effect size of this difference is very large (d = 1.772). In other words, the **learners produce significantly more truncations** per hundred words than the native speakers.

In their analyses, Henry and Pallaud (Pallaud 2002; Pallaud & Henry 1995; Henry & Pallaud 2004) and Gilquin (2008) made a distinction between different types of truncations, e.g. depending on whether the truncated word is uttered in full later on in the utterance or not. Following their work, **the proportion of completed and abandoned truncations in LINDSEI-FR+ and LOCNEC+** was also examined. The results are displayed in Figure 5-23.

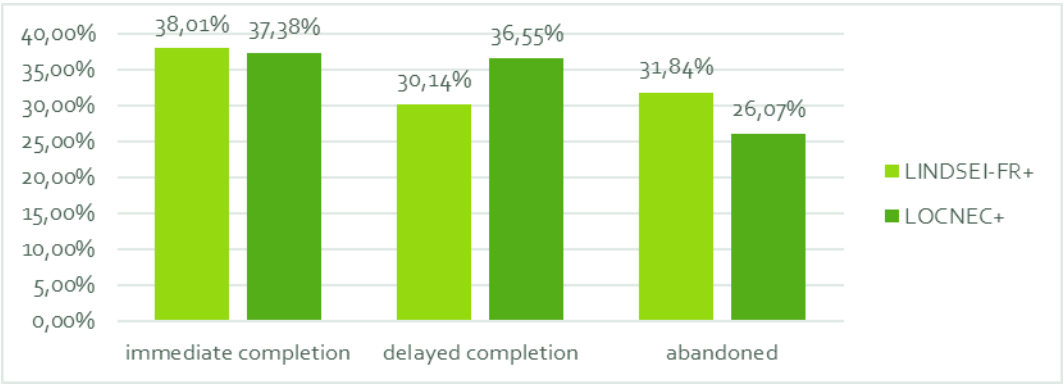


Figure 5-23: Proportion of completed and abandoned Ts in LINDSEI-FR+ and LOCNEC+

It appears very clearly from Figure 5-23 that **the overwhelming majority of truncated words both in LINDSEI-FR+ and LOCNEC+** are uttered in full after the truncation. In the learner corpus, 68% of the truncations are **completed**, and 74% are completed in LOCNEC+. A substantial proportion of the native and learner truncations **are abandoned**, and the speaker never utters the word in full: 32% in LINDSEI-FR+ and 26% in LOCNEC+. The results obtained for native (and learner) English strongly support Henry and Pallaud's (2004) results. They found that completed truncations are the most frequent in native French, before what they call "unfinished interruptions" (i.e. abandoned truncations).

When truncations are completed, the **completion** may follow three different patterns. The completion can occur **immediately after the truncation** (Gilquin's (2008) "stutters", *cf*. the second truncation in 5-87). It can also occur **after some delay** (Gilquin's (*ibid*.) "delays"). Truncations in LINDSEI-FR+ and LOCNEC+ are typically delayed by a restart (the re-utterance of one or several words prior to the interruption, which may be slightly modified, as illustrated in 5-83), or by a filled or unfilled pause (5-84). Other (dis)fluency features are but rarely used (5-85). Lastly, the completion of the truncation can occur **after one, or more, other truncation(s)** of the same target word (Example 5-86) (*cf.* De Gaulmyn 1987). This pattern is thus a specific type of delayed truncation.

5-83: FR002-S - delayed completed truncation (with modified retracing)

so er her **fa=** his **father** decided to take me by car

5-84: FR002-F - delayed completed truncation (after FP)

the fight is very **su=** er **successful**

5-85: EN012-F - delayed completed truncation (after DM)

English or whatever **pri=** well **primary** education cos that's what I'm doing at the moment

5-86: FR027-S - multiple adjacent truncations

a film about the life of **Beet= Beet= Beethoven**

When a speaker uses a truncation but never utters the full word (i.e. an **abandoned truncation**), he/she can either **use another word** (or sequence of words) instead of the target word, or even **change the structure of his or her utterance** altogether. In example 5-87, the speaker started uttering a word (*d*=) but never completed it. Instead, after an unfilled pause, she restarts her utterance, repeats the personal pronoun *they*, and replaces the target word by *just live*. Another example is provided in 5-88 for LOCNEC+.

5-87: FR001-F - abandoned truncation (and completed truncation)

they **d=** (0.520) they just live er next to **m= my** er

5-88: EN004-F - abandoned truncation

no you **c=** I think there is a: an an option which does

5.3.9 False starts

False starts (FSs) were extracted with the tag <FS>. The learner corpus totals 656 false starts and LOCNEC+ 628.

In the learner corpus, false starts occur with a **mean frequency of 0.70 FSs phw**: the French learners from LINDSEI-FR+ produce about **one false start every c. 140 words** on average. All

the learners produce false starts in their interviews (min. 2 FSs for FR050), ranging from 0.14 to 1.67 FS phw. As can be observed in Figure 5-24, the mean frequency of false starts in **LOCNEC+** is slightly lower than in LINDSEI-FR+, reaching **0.48 FSs phw**, which amounts to one false start every c. 200 words. The dispersion of individual NS means ranges between 0.00 and 0.83 FSs phw: there is thus a great degree of overlap between the L1 and L2 distributions, and some native speakers produce more false starts than the learner mean frequency.

The very low frequencies of false starts in the two corpora confirm the claim that this type of radical interruption, where the speaker gives up his/her utterance and starts anew with fresh material, is a **hardly observable phenomenon in the spontaneous speech of both native and non-native speakers** (Wisniewski 2015).

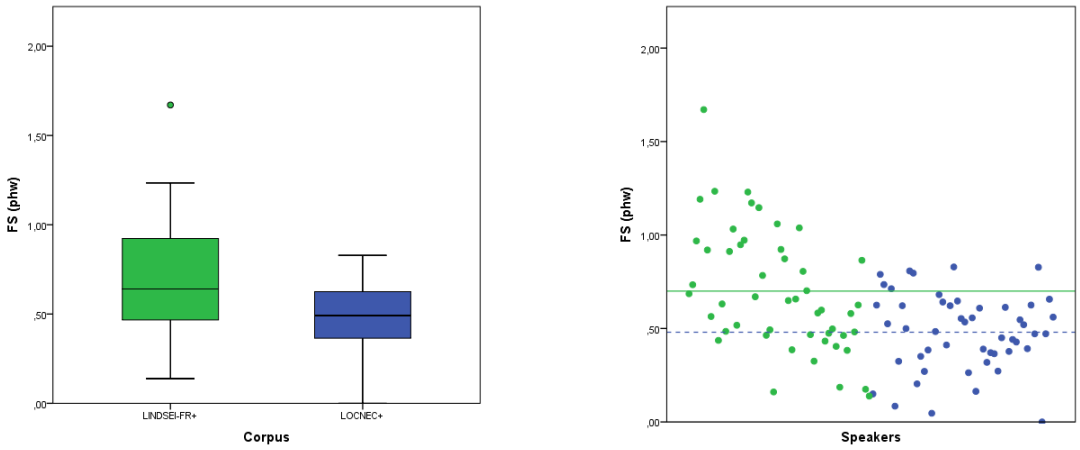


Figure 5-24: Boxplots and stripchart of FSs phw in LINDSEI-FR+ and LOCNEC+

An independent samples *t*-test indicates that the difference between the mean frequency of FSs in LINDSEI-FR+ and LOCNEC+ is significant ($t = 3.96$; $p < .001$), with a medium-sized effect for this difference ($d = 0.79$) (Cohen 1977). In other words, the French-speaking **learners produce more false starts, on average, that the native speakers in LOCNEC+**.

Given the fact that it is cognitively easier for a speaker to go on with his/her utterance, even after some delay, than to fully re-plan utterances, false starts are likely to be employed after other strategies have been exploited. To investigate this aspect of false starts, the **context of occurrence of FSs** was examined (see Table 5-14). In LINDSEI-FR+, c. 21% of the false starts are used alone, with no immediately adjacent annotated (dis)fluency feature. In LOCNEC+, this proportion is slightly higher, accounting for c. 31% of the false starts. Both in the learner and in the native corpus, the **majority of the false starts are used in clusters** with at least one other adjacent (dis)fluency feature (79% and 69%, respectively).

	LINDSEI-FR+	LOCNEC+
Stand-alone FS	21.04% (138)	31.05% (195)
FS used in clusters	78.96 % (518)	68.95% (433)
Totals	100% (656)	100% (628)

Table 5-14: Proportion (absolute frequency) of FSs used alone and in clusters in LINDSEI-FR+ and in LOCNEC+

Examples of FSs used alone are presented in 5-89 and 5-90. In these two examples, the false start is immediately followed by a fresh start with new linguistic material. By contrast, when FSs are used conjointly with other (dis)fluency features to form **clusters of adjacent features**, false starts are typically accompanied by a **filled or unfilled pause and/or a discourse marker**. Whilst for the speaker, such elements provide additional planning time, for the listener, the use of pauses or discourse markers emphasises the grammatical incompletion and the fresh start. Consider, for example, typical clusters in 5-91 and 5-92.

5-89: FR015-F - FS used alone

it's you can't sleep

5-90: EN037-P - FS used alone

you like me to sort of speech bubble it **if you** I'd sort of put words into their mouths

5-91: FR010-F - FS with FP and UP

Louvain la Neuve **is known for** er (0.240) everything is (0.140) close and near

5-92: EN030-S - FS with UP

it's in the shape of a big glass pyramid (0.290) and inside **you have** (0.640) the roof comes down in the day to let the sun in

All in all, FSs are not a very frequent phenomenon in speech, but they are more regularly used in learner speech. Such interruptions are typically accompanied by other (dis)fluency features in adjacent position. On the one hand, this could be interpreted as a sign that, for speakers, in-depth re-planning is particularly cognitively demanding and that additional delays (i.e. pauses or discourse markers for example) are required. On the other hand, the joint use of a false start with other (dis)fluency features might signal to the listener that the utterance will be left unfinished and that a new utterance is beginning: such clusters could thus also contribute to the listener's cognitive fluency.

5.3.10 Foreign words

Foreign words (Ws) feature **in both the learners' and the native speakers' interviews**. In our corpus of French learners of English, 436 words were attributed the tag <W>. Together with

the 56 instances of foreign words in the native LOCNEC+, they form an interesting category of (dis)fluency features that may reveal insightful perspectives on the use of "communication strategies" (Tarone 2005) in L1 and L2 speech.

In **LINDSEI-FR+**, foreign words have a mean frequency of **0.48 per hundred words**. In other words, learners use one word in another language than English every two hundred words. This mean rate, however, hides varied individual mean frequencies: the frequency of foreign words ranges between 0.00 and 2.87 foreign words phw. Surprisingly, native speakers also sometimes use non-English words, though much less frequently. In **LOCNEC+**, foreign words occur at a frequency of 0.04 phw, and the maximum frequency is also lower than the learners' (max: 0.49 phw). The boxplots and stripchart from Figure 5-25 provide a visual representation of the data. They highlight that the dispersion of the frequency of foreign words phw is far greater in LINDSEI-FR+ than in LOCNEC+ ($p < .000$). Note also that quite a few learners are plotted as outliers in the graphs, which emphasises the heterogeneity of the group of learners with respect to the use of this (dis)fluency feature.

The comparison of the means of foreign words in LINDSEI-FR+ and LOCNEC+ shows that there is a statistically significant difference ($t = 6.04$; $p < .000$) between the average frequency of foreign words in the two corpora: the **French learners in LINDSEI-FR+ use significantly more foreign words than native speakers**.

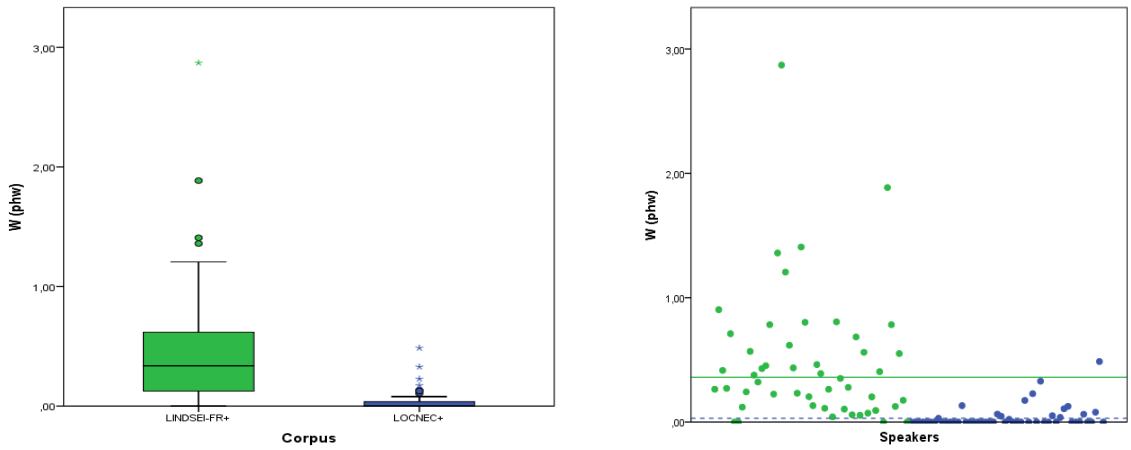


Figure 5-25: Boxplots and stripchart of foreign words phw in LINDSEI-FR+ and LOCNEC+

French is the most commonly used language for code-switching both in the learner (over 95%) **and in the native corpus** (50%) – see Table 5-15. For the NNSs, this is barely surprising as French is their mother tongue. Nacey and Graedler (2013:352) write that "[s]uccessful conveyance of the message and maintaining the flow of conversation would therefore seem to be natural priorities, more so than formal correctness: in some cases, resorting to the L1 may thus be a strategy to keep the flow going, thereby favouring a fluent delivery over the formal respect of the code". This hypothesis might be further supported by the fact that the interviewer in LINDSEI-FR+, although she was a native speaker of English, understood French too (De Cock 2015a; *cf.* also 2015b; 2017b). Consider examples 5-93 and 5-94 below, where the interviewer marks her understanding of FR. *braderie* and FR. *une section*

psychopédagogique with a backchannel, or does not react at all, as in 5-95. In the native corpus, French is typically used in discussions about Erasmus stays or internships in France. The native speakers thus use words related to the French educational system, as illustrated in 5-96. It is also interesting to see that only four of the French words in LOCNEC+ are preceded by another (dis)fluency feature, such as a lengthening of the determiner in 5-97, while the use of lengthenings and pauses is far more common in LINDSEI-FR+ (see for example 5-94).

5-93: FR029-F

B: clowns everywhere there were erm (0.580) some sort of **braderie** but
A: oh I see

5-94: FR043-F

B: there was a (0.280) psy= eh **une section psychopédagogique**
A: yeah

5-95: FR046-F

I have problem eh with the **bourse** so er I will have to (0.470) to give the money back

5-96: EN048-F - cultural/institutional bridge

I was lucky cos I was teaching in a **lycée** and a **collège** (0.370) so I had all ages

5-97: EN008-S - cultural/institutional bridge

I met some prople who worked at my school the: **surveillant** the the younger people (0.480) and they were brilliant

	LINDSEI-FR+	**LOCNEC+**
French	95.87% (418)	50.00% (28)
German	3.44% (15)	0% (0)
Italian	0.69% (3)	3.57% (2)
Russian	0% (0)	8.93% (5)
Spanish	0% (0)	7.14% (4)
Other	0% (0)	30.36% (17)
Totals	**100% (436)**	**100% (56)**

Table 5-15: The source language of foreign words in LINDSEI-FR+ and LOCNEC+

Although it is often assumed that learners only use words from their mother tongue when they code-switch, several other foreign languages can be found in LINDSEI-FR+. French-speaking learners sometimes make use of **German** (c. 3.5%) and **Italian** (c. 0.7%) lexicon. A closer look at the environment of these words indicates that they are used when the learner talks about trips in Germany and Italy, or about the German or Italian culture (e.g. It. *focola*,

Ge. *Das Boot*). Likewise, native speakers also sometimes use **Italian** (c. 3.6%), **Russian** (c. 8.9%) and **Spanish** (c. 7.4%), and those words also occur in travel-related topics (i.e. overwhelmingly in the set topic part of the interviews). They include words such as Sp. *mañana* or Ru. *dovotchka*. Two words in the "other" category are from **Indonesian** origin, and the 15 other, which are uttered by just one speaker, are "Nadsat" – an invented language in the book and film *A Clockwork Orange*. In the interview, speaker EN049 explains how he came to use some Nadsat words while speaking. Two of those words are reproduced in 5-98.

5-98: EN049-S - cultural/institutional bridge

for example em **rockers** is your hands (0.490) erm your **gollover** is your head

The use of foreign words as communication strategies will not be investigated in detail here as this aspect has already been thoroughly explored by De Cock (De Cock 2015a; 2015b; 2017b; 2017a; see also Nacey & Graedler 2013). In her studies, De Cock identified three main uses of foreign words, namely **lexical gaps, cultural/institutional bridges**, and **pragmatic/discourse bridges**. The proportion of L1 and L2 foreign words in these three categories is displayed in Table 5-16.

	Languages	**LINDSEI-FR+**	**LOCNEC+**
Lexical gaps	French	27.5% (120)	0.0% (0)
Cultural/institutional bridges	French Others	44.7% (195) 3.4% (15)	50.0% (28) 50.0% (28)
Pragmatic/discourse bridges	French	23.6% (103)	0% (0)
Totals		100% (436)	100% (56)

Table 5-16: The role of foreign words in LINDSEI-FR+ and LOCNEC+

Examples of foreign words used as **lexical gaps**, i.e. words or expressions that appear to be unknown or inaccessible to the speaker, are shown in 5-99 and 5-100. Note how, in 5-100, the learner, in addition to using many (dis)fluency features, explicitly expresses that he has problems formulating and appeals for assistance (*I wanna say*). He also approves the translation suggested by the interviewer (*to trust them yes yeah*).

5-99: FR002-F - lexical gap[131]

B: [after the Gilles got dressed] they go from place to place (0.410) **donc** (0.520) they er (2.840) **vont cherch= venir chercher** (0.330) they er

[131] In the excerpt, the learner talks about the traditions around the Carnaval de Binche. Here, she talks about the *ramassage des gilles* that happens on Shrove Tuesday. Instead of using the French word *ramassage* (as a cultural bridge), she attempts to explain the concept, but appears to have problems finding the right words. She first uses English (*they go from place to place)* then tries to reformulate herself (the reformulation is preceded by FR. *donc*), and, after a nearly 3-second long pause, she ends up using French. The French *vont cherch[er]* is actually

A: they (0.240) what they go and get or they go and look for are they looking for something or
B: er they go (0.100) and get their their friends

5-100: FR046-S - lexical gap

B: to be er er (1.840) I wanna say er (2.710) **je veux pas être je veux pas leur faire confiance**
A: oh you didn't want to trust them
B: to trust them yes yeah

The second category of foreign words identified by De Cock (*ibid*.) is that of **cultural and institutional bridges**, i.e. words or expressions that are culture-specific or that refer to a specific educational system. This category is illustrated in Examples 5-101 to 5-103 (for learners), and 5-96 and 5-97 (for native speakers).

Lastly, the category of **pragmatic and discourse bridges** includes words and expressions that have a pragmatic function, e.g. *allez, donc,* or *enfin*. This category is illustrated in 5-104, where *enfin* introduces a correction (*a student restaurant*).

5-101: FR004-F - cultural/institutional bridge

he's doing a **régendat** in er (1.240) erm (0.690) modern languages

5-102: FR009-S - cultural/institutional bridge

B: it was during the (0.460) **blocus** <laughing/> so the the period
A: the period when you were studying
B: yeah

5-103: FR028-F - cultural/institutional bridge

in fact it's in Italian it's it means […] eh in E= in English I don't know (0.390) in fact **focola**[132] means er (1.120) er (0.290) it it has to to do with fire

5-104: FR005-S - pragmatic/discourse bridge

I went to to a restaurant **enfin** a student restaurant and (0.570) I I

In sum, this section set out to examine and compare the use of foreign words by learners and native-speakers. Although native speakers also use foreign words, learners use on average more of them. Foreign words can fulfil several functions, but, while both native speakers and learners use foreign words as cultural/institutional bridges, only the learners use foreign words to bridge lexical gaps and as pragmatic/discourse bridges.

---

the exact form she is looking for, but she stops in the middle of *chercher* to reformulate her target verb in the infinitive, as if to appeal to the interviewer for help to find the correct English equivalent. After a restart and a filled pause (*they er*), the interviewer comes to the rescue to suggest two possible translations. Her intervention restores the communication flow.

[132] The learner probably meant It. *focolare*, which means hearth.

Overall, the findings appear to support Nacey and Graedler's (2013) claim that code-switching can be a highly effective communicative strategy. However, contrarily to cultural and pragmatic bridges, lexical gaps tend to co-occur with other (dis)fluency features, such as editing expressions, or appeals to the interviewer, which may be evidence of cognitive disfluency.

5.4 SUMMARY TABLE

For the sake of completeness, Table 5-17 displays the descriptive statistics for the ten annotated (dis)fluency features as well as the four temporal variables in each corpus. It includes: the mean and median frequency of each variable, the minimum and maximum frequencies, the range (i.e. the difference between the maximum and the minimum values), as well as the standard deviation, which is a dispersion measure. The results of the independent samples *t*-tests and Levene's test for homognenity of variance are also included.

(Dis)fluency variables		Corpus	Mean	Median	Stand. dev.	Min.	Max.	Range	Comparison of means			Levene's test of equality of variance
									t (indep. samples)	*p* value	*d* (Cohen)	
Temporal variables	Mean length of runs (words)	LINDSEI-FR+	5.64	5.48	1.02	3.85	8.16	4.31	*t*(70.34) = -7.121	< .000	1.424	Unequal variances $F = 12.723$; $p < .05$
		LOCNEC+	8.01	7.72	2.13	5.22	15.15	9.93				
	Mean UP length (sec)	LINDSEI-FR+	0.51	0.49	0.09	0.341	0.712	0.371	*t*(98) = -.799	n.s.	/	Equal variances $F = 0.037$; $p > .05$
		LOCNEC+	0.52	0.51	0.09	0.369	0.755	0.388				
	Phonation time ratio (%)	LINDSEI-FR+	82.75	83.66	4.52	69.87	91.62	21.74	*t*(98) = -4.755	< .000	0.951	Equal variances $F = 0.705$; $p > .05$
		LOCNEC+	86.78	87.25	3.93	78.68	92.76	14.07				
	Speech rate (wpm)	LINDSEI-FR+	162.61	159.94	15.68	131.63	206.48	74.86	*t*(88.134) = -15.486	< .000	3.097	Unequal variances $F = 6.40$; $p < .05$
		LOCNEC+	222.13	220.62	22.20	178.59	269.64	91.05				
Annotated (dis)fluency features	C (phw)	LINDSEI-FR+	5.15	5.32	1.26	0.48	7.20	6.72	*t*(98) = .771	n.s.	/	Equal variances $F = 2.78$; $p > .05$
		LOCNEC+	4.96	4.88	1.12	0.96	9.70	8.74				
	DM (phw)	LINDSEI-FR+	2.04	1.75	1.46	0.33	6.26	5.93	*t*(98) = -1.491	n.s.	/	Equal variances $F = 3.95$; $p > .05$
		LOCNEC+	2.42	2.34	1.10	0.14	6.11	5.98				
	FP (phw)	LINDSEI-FR+	7.80	7.60	2.78	2.61	13.61	11.00	*t*(70.50) = 12.409	< .000	2.482	Unequal variances $F = 27.26$; $p < .05$
		LOCNEC+	2.38	2.19	1.34	0.41	7.09	6.67				
	FS (phw)	LINDSEI-FR+	0.70	0.64	0.32	0.14	1.67	1.53	*t*(82.81) = 3.964	< .000	0.793	Unequal variances $F = 9.61$; $p < .05$
		LOCNEC+	0.48	0.49	0.21	0.00	0.83	0.83				
	L (phw)	LINDSEI-FR+	3.11	3.22	1.01	1.38	5.96	4.58	*t*(58.84) = 15.584	< .000	3.117	Unequal variances $F = 32.31$ $p < .05$
		LOCNEC+	0.77	0.70	0.32	0.09	1.56	1.47				
	Rep (phw)	LINDSEI-FR+	3.94	3.82	1.40	1.33	7.88	6.55	*t*(98) = 7.069	< .000	1.414	Equal variances $F = 3.89$; $p > .05$
		LOCNEC+	2.15	1.97	1.11	0.05	6.90	6.86				
	RS (phw)	LINDSEI-FR+	1.84	1.83	0.63	0.56	3.51	2.95	*t*(86.30) = 5.437	< .000	1.116	Unequal variances $F = 5.06$; $p < .05$
		LOCNEC+	1.25	1.14	0.43	0.23	2.55	2.32				
	T (phw)	LINDSEI-FR+	1.64	1.53	0.68	0.44	3.35	2.92	*t*(84.02) = 8.858	< .000	1.772	Unequal variances $F = 9.91$; $p < .05$
		LOCNEC+	0.63	0.55	0.44	0.04	2.42	2.38				
	UP (phw)	LINDSEI-FR+	12.69	12.86	2.77	7.68	20.20	12.52	*t*(87.36) = 12.052	< .000	2.411	Unequal variances $F = 5.13$; $p < .05$
		LOCNEC+	6.95	6.60	1.92	4.00	11.58	7.57				
	W (phw)	LINDSEI-FR+	0.48	0.34	0.53	0.00	2.87	2.87	*t*(51.97) = 5.755	< .000	1.156	Unequal variances $F = 28.66$; $p < .05$
		LOCNEC+	0.04	< 0.00	0.09	0.00	0.49	0.49				
	Total annotated features (phw)	LINDSEI-FR+	39.36	39.07	5.15	29.51	51.10	21.59	*t*(98) = 18.80	< .000	3.760	Equal variances $F = 1.919$; $p > .05$
		LOCNEC+	22.02	21.53	4.00	9.42	28.89	19.47				

Table 5-17: Descriptive statistics of (dis)fluency features in LINDSEI-FR+ and LOCNEC+

5.5 CONCLUSION

This chapter aimed to provide a descriptive overview of the fourteen (dis)fluency variables under investigation in this thesis. Each variable has been examined individually in LINDSEI-FR+ and LOCNEC+ and illustrated with corpus examples. In the light of previous studies, the findings reported in this chapter both point to similarities and differences with results reported in the literature.

A fairly constant picture emerges from the analyses of the fourteen (dis)fluency variables. Firstly, learners have significantly **lower temporal (dis)fluency measures** and produce significantly **more (dis)fluency features** (except for conjunctions and discourse markers, for which there are tendencies) than native speakers. Secondly, there is generally a fair degree of **overlap between the learner and the native distributions** (*cf.* also Osborne 2011a). While there are significant differences between learners and native speakers as a whole, when individual data are considered, it appears that some learners perform as "well" or "fluently" as native speakers. The reverse is also true for some native speakers: some of them sometimes appear to perform more poorly than learners on some variables.

One of the main findings pertains to the **mean length of unfilled pauses**. Contrarily to previous research, the mean length of UPs has been found to be lower in LINDSEI-FR+ than in LOCNEC+. Several hypotheses have been formulated to account for this finding, including the fact that UPs have been identified and measured automatically in the corpora (which is not often the case in other studies), and that learners might need extra planning time on a more regular basis, but not necessarily longer planning time. However, similar analyses into the length of unfilled pauses in other components of LINDSEI and in other corpora are needed to bring support to the findings reported here. Additionally, in follow-up studies, it might be important to relate the mean length of unfilled pauses with individual speakers' speech rate given that the latter has been shown to affect the perception of the former (see e.g. Duez 1985; Megyesi & Gustafson-Capkova 2002; Miller, Grosjean & Lomanto 1984).

Much research remains to be done for the other (dis)fluency features too. For the **temporal measures**, only a mean (speech rate, phonation-time ratio etc.) per speaker has been examined. There is some evidence, however, that there are "temporal cycles" in speech (Roberts & Kirsner 2000). A speaker's speech rate, for example, is not always constant, and it may vary as a function of the topic, the local cognitive demands etc. It would be particularly insightful to look at variations in temporal measures in each interview and to see the extent to which these variations might be revealing of local drawbacks in fluency. Furthermore, the idea of "cycle" need not be restricted to temporal measures: future investigations could also examine the extent to which (dis)fluency features are spread evenly in the discourse. The analysis of areas with a higher or lower concentration of (dis)fluency features could undoubtedly deepen our current knowledge of how fluency is constructed along the discourse.

Relatedly, this chapter has attempted to raise the veil on how (dis)fluency features are combined linearly into **clusters of (dis)fluency features**. The examination of corpus examples has enabled the identification of a number of recurrent clusters, and, more importantly perhaps, it has also emphasised how the **examination of the context of use** of (dis)fluency features may contribute to the researcher's interpretation.

Another key issue that has been brought to the foreground in this chapter is the **considerable variability** between the speakers within each group. In his article entitled "Exploring variability within and between corpora: some methodological considerations", Gries (2006:110) stresses that "corpora are inherently variable internally". He thereby means that most linguistic phenomena will yield slighty different results in different parts of the corpus. In this chapter, the analysis of each (dis)fluency variable has yielded slightly different results for each speaker. Although more research needs to be carried out to identify the reasons for the large variability within the two speaker groups, three factors can be pinpointed at this stage: first, the slightly varying proficiency level of the learners (this aspect will be investigated in Chapter 7), second, the fact that each interview contains three speaking tasks (this aspect has partially been analysed in Dumont 2017b), and, third, the fact that speakers may have different (dis)fluency profiles (Götz 2013a) – this aspect will be examined in Chapter 6. More generally, the issue of the variability in the data stresses the importance of **taking the individual into account** in addition to adopting a group perspective through the analysis of aggregate data.

Against the backdrop of this chapter, the next section delves into multivariate statistics and examines the relationship between (dis)fluency variables with a view to outlining (1) potential underlying dimensions of (dis)fluency, and (2) possible (dis)fluency profiles in learner and native speech.

Chapter 6

A MULTIVARIATE APPROACH TO LEARNER AND NATIVE SPEAKER PRODUCTIVE (DIS)FLUENCY

I say "Ummm…" *a lot. I mentioned this to Karla and she says it's a CPU word. "It means you're assembling data in your head – spooling."*

I also say the word "like" *too much, and Karla said there was no useful explanation for people saying this word. Her best guess was that saying* "like" *is the unused 97 percent of your brain trying to make its presence known. Not flattering.*

I think I'm going to try and do mental Find-and-Replace on myself to eliminate these two pesky words altogether. I'm trying to debug myself.

Microserfs (Coupland 1995:177)

Having contrasted, (dis)fluency variable per (dis)fluency variable, the French-speaking learners of LINDSEI-FR+ with the native speakers from LOCNEC+ in the previous chapter, this chapter now strives to capture the **interrelationships between (dis)fluency variables** using multivariate analyses.

The first section (Section 6.1) seeks to examine the relationship between (dis)fluency variables with a view to identifying the underlying dimensions of (dis)fluency in learner and native speech. The second half of the chapter (Section 6.2) then exploits clustering techniques to examine whether speakers can be grouped based on their performance across the various (dis)fluency variables. The two analyses are similar in that they both aim at identifying the structure of interwoven relationships, but whereas Section 6.1 is centred on the **grouping of (dis)fluency variables into latent (dis)fluency dimensions** based on aggregate data, Section 6.2 is centred on the **grouping of individual speakers** based on their similarities across (dis)fluency measures.

The analyses of this chapter are based on the **data** extracted from the time aligned and annotated interviews of LINDSEI-FR+ and LOCNEC+, as presented in Chapter 3, Section 3.3 (*cf*. the list of the 14 (dis)fluency variables in Table 3-6) and in Chapter 4 (especially in Section 4.2.5).

Before embarking into this chapter, three words of warning ought to be mentioned.

This chapter mostly deals with **quantitative data**, and terms like *p* value, *t*-test, factor rotation, factor loading, correlation matrix, hierarchical cluster, and other potentially obscure jargon are bound to sprout more or less prolifically across the sections. I have done my best to make all the statisticalese more accessible – for example by explaining the aim and working of statistical tests as clearly as I can – and to make all these number-crunching data as reader-friendly and eye-friendly as possible – for example by including only the essential figures within the text[133], and by integrating many coloured graphs and charts.

The second warning has to do with the slight imbalance between the attention devoted to **learner and native speaker data**. In accordance with the general objective of this dissertation, the main focus is on the learners' speaking performance, and the discussion of native speaker data is consequently slightly less lengthy.

Lastly, it has to be kept in mind that the two statistical techniques used in the following sections (i.e. Principal Components Analysis and Cluster Analysis) are **exploratory and descriptive** and that the results of such tests could, in this context, be affected by factors such as the nature of the speaking task, the level of the learners or their mother tongue background, the variables chosen for the analysis and their measurement, as well as all sorts of other test options. Comparisons with other studies that have used the same statistical tests on learner or native spoken data (which, incidentally, are in fact fairly scarce) thus need to be exercised with caution.

[133] Some additional tables and figures are included in Appendix 9.7 (for the Principal Components Analysis) and 9.8 (for the Cluster Analysis).

6.1 THE RELATIONSHIP BETWEEN (DIS)FLUENCY VARIABLES

Many researchers have wondered about the relationship between (dis)fluency variables, and the question has been raised whether (dis)fluency "dimensions" could be delineated (esp. Skehan 1999; also Foster & Skehan 1996; 1999; Grosjean 1980; Housen, Kuiken & Vedder 2012; Segalowitz 2010; Skehan & Foster 2012). Following this line of thought, separate Principal Components Analyses (PCA) were run on the learner and on the native speaker data with a view to identifying interrelationships between individual (dis)fluency variables and potential latent dimensions of L1 and L2 (dis)fluency. The identification of such sub-dimensions of (dis)fluency and the examination of how these inter-relate will expand our knowledge of (dis)fluency in general, and render the choice of particular variables or measures in future research studies more effective.

Before expounding the procedure and results, it might however be helpful to re-define some technical terms (see also Chapter 3).

Key statisticalese	
Principal components analysis	A factor model in which the components (factors) are based on the *total* variance.
Correlation matrix	A table showing the correlations between a set of variables. The correlations between the components/factors extracted with oblique rotation are displayed in the *factor correlation matrix*.
Eigenvalue	The amount of variance accounted for by a component or factor.
Scree plot	A plot of the eigenvalues associated with each component/factor ranked in decreasing order. It is used to visually assess which components/factors explain most of the variability in the data.
Factor/component	A linear combination of observed (i.e. original) variables. They represent underlying dimensions and summarise the original set of variables.
Factor/component loading	The correlation between the original variable and each factor/component. Variables with higher loadings (in absolute value, as loadings can be positive or negative) on a factor are better representatives of the dimension underlying that factor than variables with lower loadings. Factor loadings are displayed in a *factor matrix*.
Factor/component rotation	The process of adjusting the factor solution to achieve more meaningful results. The rotation may be orthogonal (the extracted factors are independent) or oblique (the extracted factors may be correlated).

Factor/component score	A single score from an individual representing their performance on some latent variable. Its calculation is based on the speaker's performance on the constituent variables of the factor/component and on the relative importance of each constituent variable (i.e. the factor/component loading).

6.1.1 (Dis)fluency dimensions in the learner corpus

In the preliminary steps of the analysis, the 14 learner (dis)fluency measures were integrated (Preliminary Analysis 1 – PA1). These measures encompass the following: speech rate, mean length of runs, phonation-time ratio, mean length of unfilled pauses, unfilled pauses, filled pauses, conjunctions, repetitions, lengthenings, discourse markers, restarts, truncations, false starts and foreign words (*cf*. Table 5-1). The screening of the correlation matrix between the variables indicated that one variable, lengthenings, did not correlate significantly with any other variable. Following Field's (2013:685–686) advice, this variable was removed from the dataset, and the analysis re-run with the remaining 13 variables. In this second pre-analysis (PA2), the suitability of the data for factor analysis was investigated. The value of the overall Kaiser-Meyer-Oklin (KMO) statistic, which assesses the sampling adequacy for the analysis, was below the minimum criterion of .5 and the examination of the KMO statistic for individual variables in the anti-image correlation matrix led me to exclude a second variable, mean length of unfilled pauses, because it had the lowest value. The analysis was re-run with the 12 remaining variables (PA3). This time, the KMO statistic was above .5. I also examined the KMO values for the individual variables to determine whether I should consider excluding another one. Although the KMO for conjunctions was below .5, its removal resulted in a sizeable loss of total explained variance of the resulting components (PA4), so I decided to keep this variable nonetheless. Besides, Bartlett test of sphericity was significant at the .0001 level, supporting the factoriability of the data with conjunctions included.

To extract the components, I adopted Kaiser's criterion (which retains only factors that have eigenvalues > 1 and is the default option in SPSS), and I also examined the scree plot. In PA5, oblique rotation (direct oblimin) was used to increase the interpretability of the components extracted from the 12 variables (*cf*. PA3). The factor correlation matrix was used to assess whether it was reasonable to assume independence of the factors. The matrix indicated that there were actually barely any correlations between the extracted components (all $r \leq .11$), so I followed Pedhazur and Schmelkin (1991, in Field 2013:680) and ran the final analysis using orthogonal rotation[134].

[134] Note that I also ran a Factor Analysis with orthogonal rotation on the 12 variables and compared the results with the PCA. For the most part, they were identical, with only small differences.

For the **final analysis**, a Principal Components Analysis was conducted on **12 (dis)fluency variables** (i.e. excluding lengthenings and mean length of unfilled pauses) with orthogonal rotation (varimax). The Kaiser-Meyer-Olkin statistic for the sampling adequacy reached .59, which is above the minimum limit suggested by Field (2013:685). Based on Kaiser's criterion (i.e. eigenvalues over 1), **5 components** were retained. Together, these five components explain 77.26% of the variance[135].

The **interpretation** of the principal components is based on the identification of which of the original (dis)fluency variables are most strongly correlated with each component. In other words, **factor loadings** are used to gauge the substantive importance of a given variable to a component: the higher in magnitude a loading is (in either the positive or the negative direction), the more important the variable is for the component. As a consequence, and as stressed by Hair *et al.* (1995:397), "variables with higher loadings influence to a greater extent the name or label selected to represent a factor" . Table 6-1 below provides an overview of the rotated factor loadings[136], and items with loadings higher than .40 (Field 2013:706) are flagged (Figure 6-1 offers a synthesised version of the same data). The interpretation of the components also partly relies on **component scores**, which summarise each individual's performance on the component based on their performance on the variables that load highly on that component. Such scores make it possible to identify which individuals possess a particular characteristic represented by the component to a high or low degree (high positive and high negative scores, respectively) (Field 2013:672–673; also DiStefano, Zhu & Mindrila 2009; Hair *et al.* 1995).

The preliminary examination of the (dis)fluency variables that cluster together suggests that Component 1 represents the temporal side of (dis)fluency and Component 2 represents repair (dis)fluency. Component 3 mainly loads on two types of fillers, Component 4 seems to represent a dimension that could be linked to discourse cohesion and, lastly, Component 5 groups together two variables typically associated with learner disfluency.

The last three components are more complex to interpret than the first two because of the grouping of (dis)fluency variables and/or because some variables have high loadings on several components at the same time. Also, no other factorial study of (dis)fluency has explicitly included discourse markers, conjunctions and foreign words as additional variables. Therefore, the interpretation of the last three components should be seen as tentative, and they would definitely need to be corroborated by similar analyses on other datasets as well as more fine-grained investigations of the constituent variables of each component.

---

[135] The scree plot (displayed in Figure 9-1 in Appendix 9.7) showed two inflexion points that could justify retaining either 3 or 5 components, but I retained 5 components because of the convergence between Kaiser's criterion and the scree plot on this value.

[136] The factor loadings before rotation are included in Appendix 9.7 – Table 9-8.

	Rotated factor loadings				
(Dis)fluency variables	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Unfilled pauses	**-.943**	.037	-.039	.101	-.082
Mean length of runs	**.891**	.173	-.048	.061	-.133
Phonation-time ratio	**.870**	-.178	-.184	.080	.018
Speech rate	**.744**	.048	**.412**	-.151	-.039
Truncations	-.032	**.830**	-.121	-.124	.152
Restarts	.089	**.759**	-.136	.384	.155
Repetitions	-.035	**.698**	.139	-.010	-.212
Discourse markers	.077	.132	**.795**	-.057	-.257
Filled pauses	.075	.308	**-.760**	-.055	-.264
Conjunctions	-.038	-.010	-.069	**.920**	-.104
Foreign words	-.010	-.012	-.103	-.106	**.866**
False starts	-.059	.210	.320	**.534**	**.566**
Eigenvalues	3.047	2.105	1.579	1.398	1.143
% of variance	25.39	17.54	13.16	11.65	9.52

Table 6-1: Summary of PCA for the LINDSEI-FR+ data
Note: Factor loadings over .40 appear in bold; the variables are ranked in decreasing order of loading

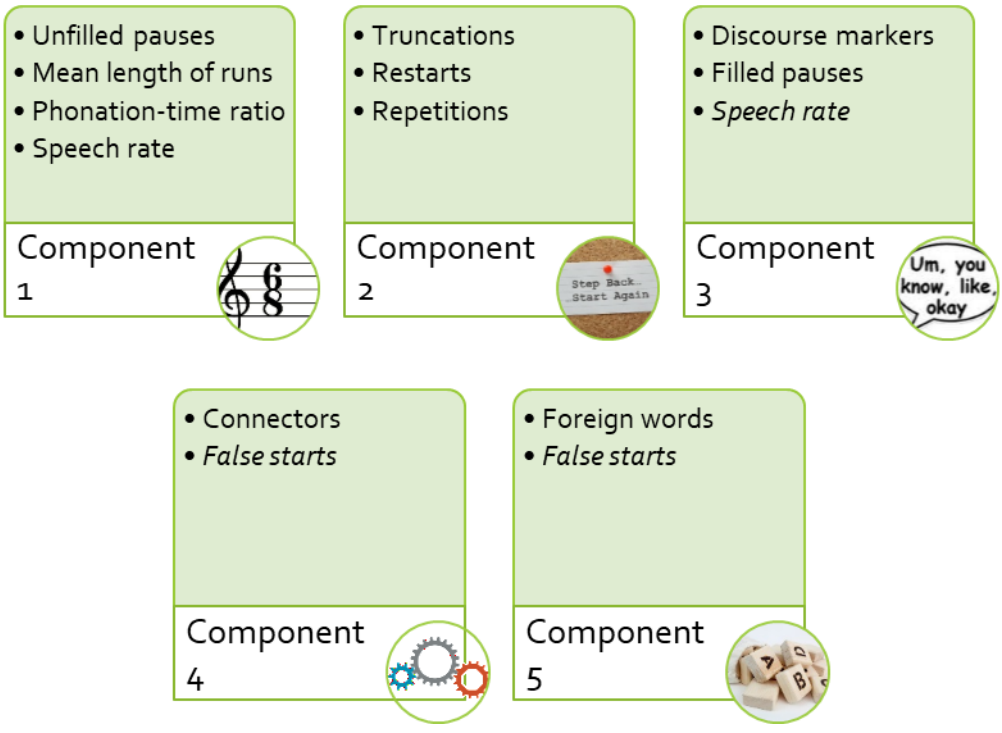


Figure 6-1: The 5 learner (dis)fluency components
Note: Italics show proportionally low loadings

The following sections review the five learner (dis)fluency components in greater details, including graphs displaying the relationship between the component scores and the constituent variables (expressed in *z*-scores for greater clarity).

6.1.1.1 Component 1

The first principal component that came up from the PCA is strongly correlated with four of the original (dis)fluency variables, namely the frequency of **unfilled pauses** (UP), the **mean length of runs** (MLR), **phonation-time ratio** (PTR) and **speech rate**[137] (SR). This indicates that these four variables are closely interrelated. Whilst this is actually not very surprising for UPs and MLR, given the fact that the rate of unfilled pauses is used directly to measure MLR, it is a little less obvious for PTR and SR. PTR and SR are measured by dividing the speaking time or the number of words by the total speaking time (i.e. including inter-turn pauses). Total speaking time is thus likely to vary as a function of the rate of unfilled pauses, which might explain their grouping with UPs and MLR.

Because the four variables that load highly on this Component refer to different aspects of temporal (dis)fluency, Component 1 could be viewed as a measure of overall **temporal (dis)fluency**. It accounts for a sizeable 25.4% of the variance.

Table 6-1 reveals that Component 1 correlates most strongly with the rate of unfilled pauses (-.943), very strongly with mean length of runs (.891) and with phonation-time ratio (.870), and strongly with speech rate (.744). However, whereas the variable UP has a negative loading on the component, the others all have positive loadings. This means that **Component 1 increases with decreasing UPs and increasing MLR, PTR and SR** – Figure 6-2 below provides a visual representation of this relationship. In other words, speakers with higher temporal (dis)fluency scores can be expected to use significantly fewer unfilled pauses, longer runs, a higher phonation-time ratio and a faster speech rate, and could be said to be temporally more fluent.

To better illustrate the linguistic reality behind high or low scores on the (dis)fluency dimensions, as well as their interpretation, two excerpts are presented in Examples 6-1 and 6-2 (unfilled pauses are in bold font). FR042 has the highest temporal (dis)fluency score among the learners (namely 2.06) and FR041 the lowest (-2.21). As can be seen in the excerpts, FR041's unfilled pauses are not only far more frequent (17.86 phw vs. 8.24 for FR042), but also much longer on average, which consequently leads to far lower values for mean length of runs (3.85 vs. 7.90), phonation-time ratio (69.89% vs. 91.62%) and speech rate (149.72 vs. 179.23 words per minute).

[137] Speech rate also loads slightly on Component 3 (.412).

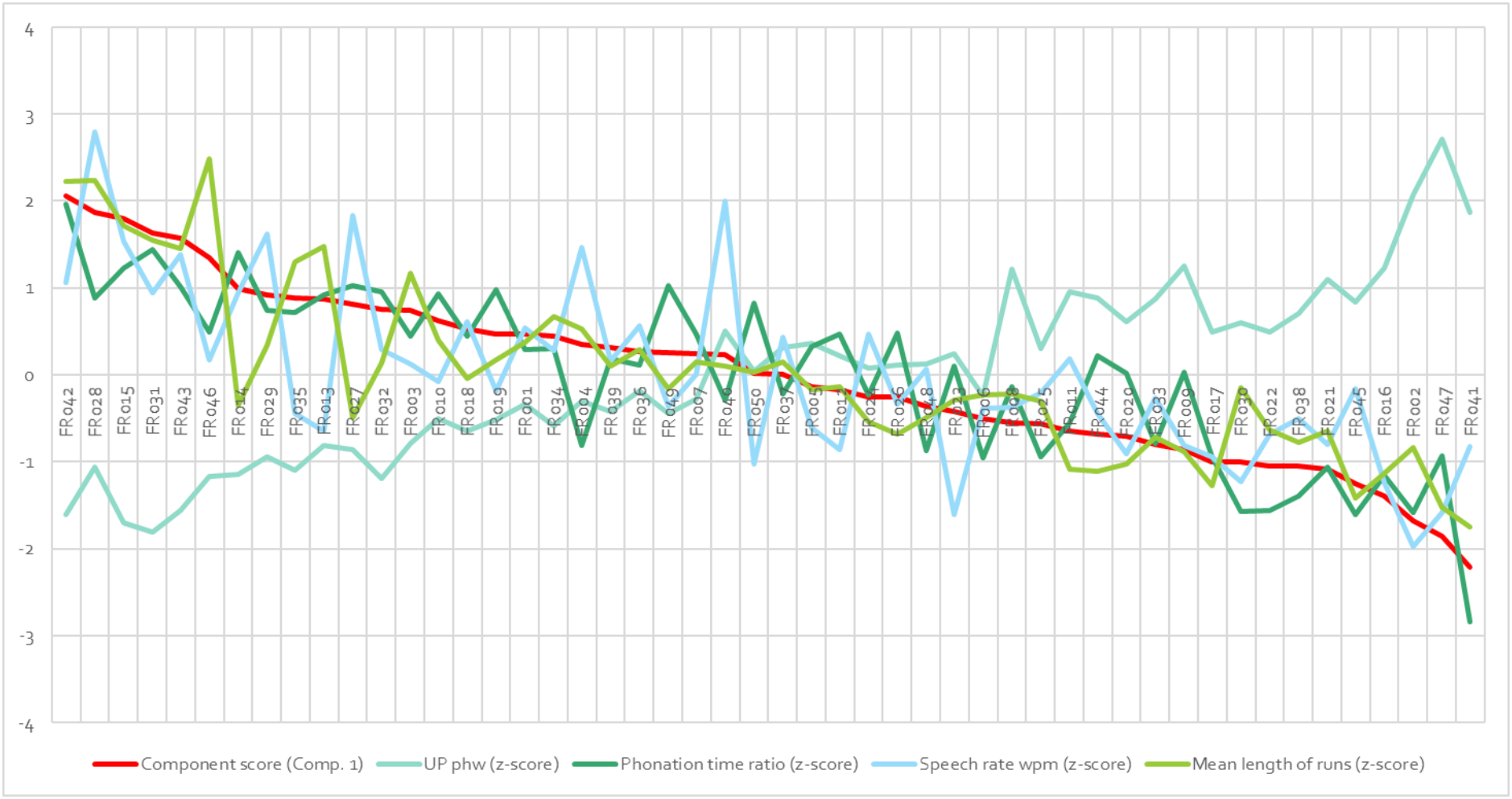


Figure 6-2: Constituent variables of Component 1 in LINDSEI-FR+

6-1: Positive Component 1 score (FR042)

FR042: mm yes also er there are eh also specialists er in Brussels at the K U B so erm **(0.730)** and **(0.380)** so **(0.230)** well I tried to go **(0.290)** once in er Brussels **(0.210)** to fi= to fetch books **(0.280)** then come here er to Louvain ouais **(0.230)** although mm my topic the passive is quite interesting I would say

A: and what exactly wha= what exactly are you focusing on

FR042: the passive

A: but but a particular (0.220) application of it or

FR042: mm er no really it's really technical you know <laughing/> and but eh I think that **(0.250)** erm I'm doing a lot of research now on er language acquisition **(0.190)** of the passive for example and there are quite erm er mm a lot to say I would say in in eh generative grammar eh starting from the perspective of inateness for example

6-2: Negative Component 1 score (FR041)

FR041: **(1.070)** at the beginning yes **(0.490)** but maybe in the licence **(0.300)** you **(0.890)** er the licence it's **(1.370)** quite **(0.220)** levelled I think the

A: but you're not so much older presumably (0.180) what are you

FR041: two or three years

A: two or three years so it's not er not as

FR041: yes in in fact it's well it **(0.630)** it took me three years to **(1.900)** before **(0.270)** er

A: and what made you to decide er what made you decide to (0.340) to do that

FR041: I I wanted more lit= lit= er literature courses <unknown/> yes **(1.330)** I was **(0.510)** interested in literature **(0.400)** and I **(0.850)** I felt **(0.900)** dissatisfied or or not **(0.190)** not er **(0.700)** satisfied enough with er the courses I had had before **(0.450)**

The results of two previous studies – on learners of Dutch and on English learners of German – square with the findings for this component. Tavakoli & Skehan (2005b) identified a factor made up of the following variables: time spent speaking (i.e. phonation time ratio), speech rate, rate of unfilled pauses and mean length of runs (plus total silence, which I did not investigate because it comes in complementary distribution with PTR, and mean length of unfilled pauses, which I excluded for methodological reasons, *cf*. *supra*). Likewise, Mehnert (1998) found a factor made up of speech rate, mean length of pauses and (filled and unfilled[138]) pauses (plus total pausing time). However, my results suggest that filled and unfilled pauses should not be grouped together in learner language (at least, in French-speaking learners' English), because, in my data, they clearly load on two different components. While unfilled pauses are associated with other temporal measures, filled pauses are strongly associated with discourse markers (which none of these two studies included) and load on Component 3.

6.1.1.2 Component 2

The second principal component is strongly correlated with three (dis)fluency variables: the rate of **truncations**, of **restarts** and of **repetitions**. The association of restarts and repetitions could be explained from a conceptual point of view. The two phenomena have in common a two-part structure of the type reparandum/repair (Levelt 1983; 1989; Shriberg 1994), and the

[138] The author only has one measure for pauses, which includes filled pauses and unfilled pauses of over 1 second, as measured by a stopwatch, in the first 3 minutes of speech.

difference between restarts and repetitions only lies in the presence (or absence) of some kind of modification in the second part (such as a substitution, a deletion or an insertion). Incidentally, most learner truncations also have a two-part structure with the word uttered in full after the truncation (*cf.* Section 5.3.8). Note also that when truncations are completed, some words may be repeated prior to the completion (such as a determiner), in which case, these were annotated as a restart (*cf.* Examples 4-26 and 4-27 in Chapter 4; Example 4-26 is included below for greater clarity). This may partly account for the association of truncations and restarts that was found in this component.

4-26: Annotation of completed truncations with restart - FR002-F

they	d	(0.310)	they	dance	around	the	fire
	<T	<UPA>	<RS	RS>+T>			
		<N>					

Based on the (dis)fluency variables it contains, this component could be interpreted as representing **repair (dis)fluency**. This component explains 17.54% of the variance.

The three (dis)fluency variables have positive loadings of .830, .759 and .698, respectively. This not only suggests that these three variables vary together, but that they also vary in the same direction: **Component 2 increases with increasing rate of truncations, restarts and repetitions** (see also Figure 6-3 for a visual representation). It follows that speakers who score highly on repair (dis)fluency produce a high number of truncations, restarts and repetitions, and speakers with a low score can be expected to produce a much smaller number of these. The latter group is arguably more fluent on this component. An illustration of a high positive score on Component 2 is shown in Example 6-3 and of a high negative score in Example 6-4. As can be observed, FR028 uses a notable number of truncations (3.07 phw), restarts (3.51 phw) and repetitions (7.88 phw), whereas these phenomena are barely present in the speech of FR009 (0.58; 0.87; 2.38 phw, respectively).

6-3: Positive Component 2 score (FR028)

FR028: I don't think **that that** people are in fact selfish in themselves (0.140) people **are** (0.400) in fact eh (0.270) **are a= al= all** everybody's eh **is** aimed **to** (0.550) **to** live with **o= other** people so (0.260) but **I I I** think that er (0.520) the kind of eh (0.760) yes **you you** said erm (0.340) the environment so all the

A: the sort of society you live in

FR028: **the sort of society** yes **the society** we live in (0.200) in fact **we are** <unknown/> sorry (0.280) eh <laughing/> […] **we are** conditioned by (0.400) external factors in fact I think and **it's it's** very hard **to** (0.630) **to to** go er against the

6-4: Negative Component 2 score (FR009)

FR009: and er I received the letter telling me that I had won the big prize

A: oh how amazing

FR009: yeah and I remember during the day I wasn't (0.420) able to study anything because **I was I was** so excited (0.130) I phoned to my mother at her office to my father and the whole family **was** er (0.480) **kn knew** about it (0.800) so

A: what about your sister

FR009: er well she helped me for that (0.220) radio play but (0.390) er she herself eh won another play she won a two

current in fact to (1.450) but **I I** think that (0.340) er	days trip to London in fact **to to** attend a concert

The work by Skehan and colleagues (Skehan 1999; Skehan & Foster 1999; Tavakoli & Skehan 2005b) also identified a "repair" dimension of fluency, which partially matches with the results for the French-speaking learner data as it also includes reformulations, replacements and repetitions. However, the present findings revealed that truncations also play an important role in this dimension of (dis)fluency, and they should therefore be considered as a fully-fledged (dis)fluency variable in future analyses. In contradiction to Skehan's work, however, the present findings disprove the association of false starts with restarts and repetitions in the repair component: in LINDSEI-FR+, false starts load with with conjunctions on Component 4 and with foreign words on Component 5. It is plausible that the selection of the variables included in our respective statistical analyses could have affected our results differently. In particular, the pooling of "replacements" and "reformulations" into a single "restart" category or the integration of additional categories as compared to previous work (e.g. conjunctions) may account for those diverging findings.

This notwithstanding, the results for Components 1 and 2 lend support to Ellis & Barkhuizen's (2005; in Guz 2015:234) call to keep a clear distinction between temporal variables such as speech rate, number of unfilled pauses, or length of runs on the one hand, and phenomena such as false starts, repetitions and reformulations on the other because they represent very different aspects of speech production.

6.1.1.3 Component 3

The third principal component shows a combination of three substantive loadings: **discourse markers**, **filled pauses** and **speech rate**. Whereas discourse markers and filled pauses load highly on the component (.795 and -.760, respectively), speech rate barely reaches the minimum acceptable level of .4 (.412), which means this measure it is a less good representative of the third underlying dimension than discourse markers and filled pauses (see also Figure 6-4 where the apple green line follows the component score line less closely than that of discourse markers and filled pauses). Additionally, speech rate loads more highly on Component 1 (.744), which implies that, although it is indeed related to discourse markers and filled pauses, it is more strongly associated with the other temporal (dis)fluency variables.

It is not immediately obvious which domain of (dis)fluency this component represents. No prior study has found a statistical association between discourse markers and filled pauses. In some studies, filled pauses are simply conflated with unfilled pauses (as in Mehnert 1998) on the grounds that, although the former are vocalised while the latter are not, they are pauses nonetheless. Yet, filled pauses and discourse markers are also sometimes referred to in the literature with the umbrella term "fillers", filled pauses being "non-lexical", and discourse markers being "lexical" (*cf.* e.g. Rohr 2017; Rose 1998; Stenström 1994).

Both discourse markers and filled pauses have sometimes been associated with planning difficulties due to higher cognitive load, and reflect the fact that speakers are, e.g., searching for a word, are in doubt, or are asking for help (Barr 2001; Crystal 1988; De Jong *et al.* 2012b). These two (dis)fluency features have also been argued to participate in the pragmatic aspect of an interaction: discourse markers have been shown to fulfil a variety of functions (see e.g. Denke 2009; Fraser 1990; Jucker & Ziv 1998; Mira 1998; Müller 2005; Schiffrin 1987), and so have filled pauses (e.g. Corley & Hartsuiker 2003; Swerts 1998; Tottie 2011; Watanabe *et al.* 2008). Tottie (2014:25), in the conclusion of her examination of *uhs* and *ums* in American English, explains that "[i]t is likely that *uh* and *um* originated in situations of cognitive load, where speakers needed time to pause to think and plan, but that **they – like the markers consisting of bona fide words such as *you know*, *I mean*, *well* etc. – have now also acquired pragmatic meanings**" (my emphasis). Incidentally, in Chapter 5, a recurrent association between discourse markers and filled pauses had been uncovered: they are regularly used conjointly to form a (dis)fluency cluster (DM+FP or FP+DM) (see also Crible, Degand & Gilquin 2017).

Bearing this in mind, and acknowledging that further investigations into the pragmatic and functional uses of discourse markers and filled pauses are definitely needed to establish a more coherent picture, Component 3 could be viewed as representing the area of **pragmatic (dis)fluency**. It accounts for 13.2% of the variance in the learner dataset.

As indicated by the component scores, discourse markers (and speech rate) and filled pauses move in opposite directions: **Component 3 increases with increasing rate of discourse markers** (and higher speech rate) **and decreasing rate of filled pauses**. Speakers who have a high score on this (dis)fluency component can thus be expected to produce proportionally more discourse markers, fewer filled pauses and speak more quickly than learners who have a lower score on this component. This component thus suggests a trade-off between discourse markers and filled pauses in learner language. Compare, for example, learner FR031 and learner FR002 (Examples 6-5 and 6-6, respectively). It is striking that FR031, who has the highest score on this component, repeatedly uses discourse markers (especially *well*) at the beginning of her utterances (in fact, as many as 6.23 phw) but proportionally few filled pauses (3.49 phw). It is the exact opposite for FR002, who uses filled pauses nearly compulsively (13.61 phw), but very few discourse markers (only 0.34 phw). Although a high use of discourse markers in learner speech could be assumed to be a sign of pragmatic fluency, Example 6-5 also stresses that qualitative aspects such as the appropriacy of use should definitely not be disregarded.

6-5: High positive Component 3 score (FR031)

FR031: **well** I'm going to talk to you about Russia (0.210)

A: oh

6-6: Low negative Component 3 score (FR002)

FR002: and **er in fact** (0.380) **er** the fight is very su= **er** successful there are a lot of people who come and see (0.320) that kind of **eh** celebration and (0.310) it's also **er** (0.590) a tradition who goes back to **er** (0.970) **er** (0.570) one great saint

FR031: **well** I went to Russia a few years ago when I was in the sixth form (0.320) and it was something organised by the school **in fact er well** the year before (0.220) a group of Russian students had come to our school and we (0.240) **well** and the the pupils who intended to go to Russia (0.360) the following year (0.190) had to take one of those **eh** Russian students at home […] and **er well** and the year (0.230) after I went to Russia […]

FR031: it was **well** it looks **eh well** at Saint Louis it looks a bi= it looked a bit like **eh** <sighing/> **well eh** (0.400) at s= secondary school **in fact** (0.770) we were twenty five in the first year and

A: oh quite a l= that's quite a lot

FR031: **well** at the beginning (0.290) **er** saint (0.860) **er** i= (0.410) it's a woman (0.210) it's **er** Sainte **er**

A: Sainte Catherine or something

FR002: **er**

A: no (0.630) somebody told me about her name

FR002: I don't remember (1.090) **er** it was **er** in relation (0.310) to (0.480) **er** (0.720) a disease (0.330) and there was **er** (0.250) a big disease in **er** Mons (0.360) in the fourteenth century I think (0.470) and **er** (0.860) they **er** (0.780) they to took out (0.290) **er** (0.300) the relics of **eh** that saint (0.510) out of the church and **er** (0.690) the disease was **er** (0.230) was finished after that

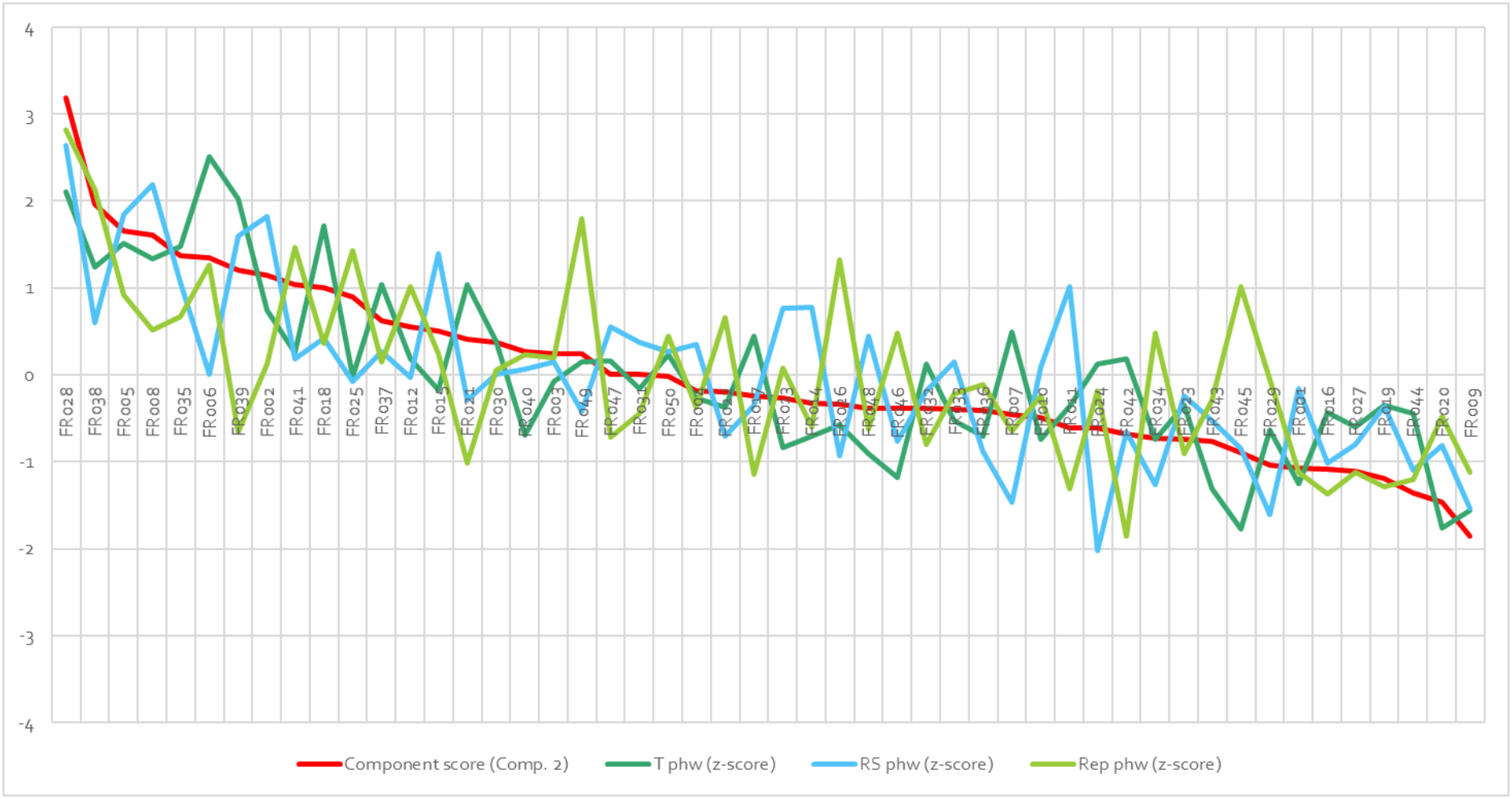


Figure 6-3: Constituent variables of Component 2 in LINDSEI-FR+

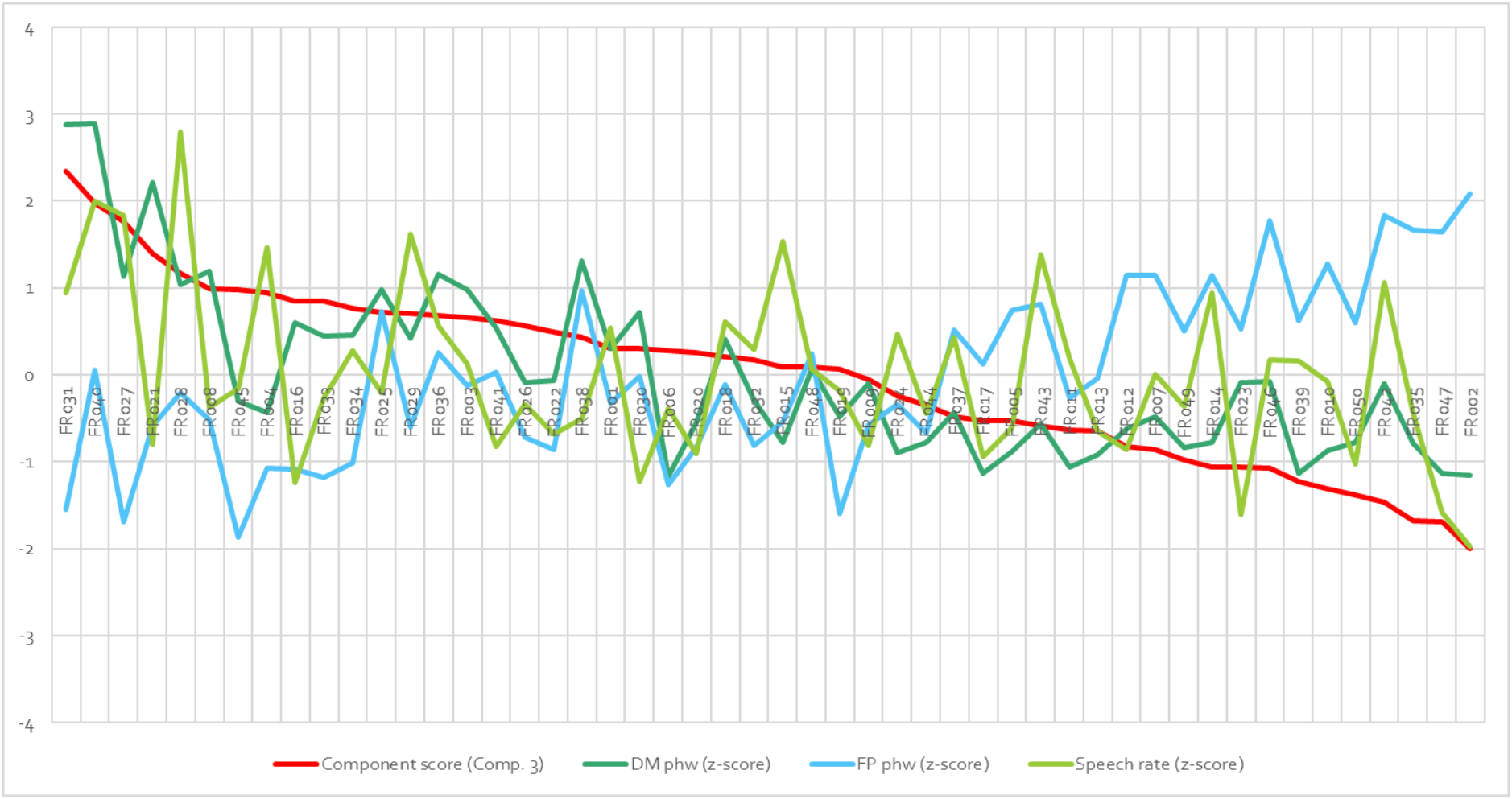


Figure 6-4: Constituent variables of Component 3 in LINDSEI-FR+

6.1.1.4 Component 4

Component 4 is very highly and positively correlated (.920) with the rate of **conjunctions** (*and, so, but*). It is also correlated with another variable – the rate of **false starts** –, although the loading (.534) suggests that this measure is less central to the dimension – Figure 6-5 offers a visual representation of the constituent variables of Component 4. Note also that false starts load nearly equally highly on Component 5, which renders the interpretation of these two components more complex. Given the difference in loadings, it could be tentatively suggested that this component is primarily a measure of **discourse cohesion**. This component accounts for 11.7% of the variance in the learner data.

The association of conjunctions and false starts seems somewhat perplexing at first sight. Conjunctions occur with a high frequency in the learner corpus; false starts are much rarer (*cf.* Section 5.3). Conjunctions are used to make explicit the links between utterances, units, or ideas, and are not always obligatory. False starts are a sign of lexico-grammatical breakdown, and speakers do not choose to produce (or not to produce) them. Another perplexing element is that both conjunctions and false starts have positive loadings on this component, which means that **Component 4 increases when the rate of conjunctions *and* of false starts increases**. In other words, learners who produce both more conjunctions and more false starts score highly on this component. For example, with 6.73 conjunctions and 1.67 false starts phw, FR005 (Example 6-7) has the highest score on this component, and FR024 (Example 6-8), who produces very few conjunctions and false starts (0.48 and 0.16 phw, respectively), has the lowest score. Compared with FR005, FR024 seems to use conjunctions not only less frequently, but also differently. While in the speech of FR024, conjunctions are used to mark continuity and to create links between utterances (Bestgen 1998; Schiffrin 1987), in Example 6-7, some conjunctions seem to be more closely related to planning difficulties or to appeals to the interviewer (Altenberg 1987; Peterson & McCabe 1987; Rose 1998; Spooren 1997). For example, the *so* after *we went* might arguably reflect the learner's production difficulties (the speaker is trying to find the proper wording).

6-7: Positive Component 4 score (FR005)

FR005: no **we** (0.540) in fact **it was th=** erm <sighs/> you have er Paris here **and** the Loire here **we went** (0.240) <sighs/> **so** (0.320)

A: oh I see

FR005: along enfin

A: so not <unknown/> along

FR005: mainly mainly the forests

A: yeah

FR005: **it was** er (0.830) **so and** we (0.880) we had some tents **and** (0.660) we pick up

6-8: Negative Component 4 score (FR024)

FR024: er (0.970) **and** why (0.970) does did (0.230) I love this film (0.530) that was the question I I couldn't answer it (0.710) **but** er (1.020) I find (0.390) I I think when (0.420) in a society now when we are thinking about (0.210) er (0.170) means of punishment (0.070) er capital punishment or (0.320) er (0.080) prison er (0.420) the whole life long erm (0.640) this film is proposing a (0.200) a s= solution a psychological solution (0.470) **and** er (1.490) **and** I I think (0.340) this solution could have been (1.640) yes (0.260) worth trying it

```
(0.360) we picked up (0.520) eh our tents
[…] and er they (0.550) so they they haven't
eh (0.390) all the eh comfort and (0.080)
so
```

Whereas a higher rate of false starts may be linked to a lower level of fluency, it is less obviously so for conjunctions, but several hypotheses could be proposed at this point to explain this association. One hypothesis is that less fluent learners may tend to make more explicit the links between utterances and/or that they tend to stick to the highly frequent conjunctions *and, so* and *but* at the expense, perhaps, of less frequent conjunctions or other means to establish discourse coherence (see e.g. Buysse 2014; Hasselgren 1994). Besides, given the fact that the French language, unlike English, "finds it difficult to manage without the connections they [conjunctions] can bring to the presentation of thought" (Vinay & Darbelnet 1995:234), it might also be possible that less fluent learners transfer their L1 French coherence strategy onto their L2 English to some degree, and that the latter is thus characterised by an overall higher rate of conjunctions. In this respect, in a study comparing the frequency of discourse connectives in native French and native English in eight communicative tasks, Crible (2017a) provided some convincing evidence supporting an overall higher rate of conjunctions in French as compared to English. Lastly, it can also conceivably be hypothesised that less fluent learners use conjunctions to a greater extent because these can also function as traces of difficulties speakers encounter when formulating (Bestgen 1998) or as some kind of filler to stall for time to plan forthcoming speech (Altenberg 1987; Rose 1998; Spooren 1997).

6.1.1.5 Component 5

The last component that emerged from the Principal Components Analysis on LINDSEI-FR+ data is highly correlated with two (dis)fluency variables: **foreign words** – with a loading of .866 – and **false starts** – with a loading of .566. False starts, given their lower loading, are slightly less good representatives of this underlying dimension of (dis)fluency than foreign words. Incidentally, false starts also load on the fourth component, and with about the same loading. Figure 6-6 provides a visual representation of the variables that load on Component 5.

Both foreign words and false starts are **very rare**, but their presence is strongly associated with learner and/or disfluent speech. The use of foreign words is generally argued to be a sign of great lexical difficulties (although they may have other uses too, as explained by, e.g., De Cock (2015a)), and false starts are indicators of lexico-grammatical breakdowns that cannot be resolved through, e.g., reformulation. Because of the negative connotation of these two (dis)fluency variables, Component 5 could perhaps be viewed as a measure of **lexico-grammatical disfluency**. It explains just under 10% of the variance in the data.

An illustration of high and low component scores is provided in Examples 6-9 and 6-10. FR018, who has the highest score on this component, produces a high rate of foreign words

(especially the French discourse marker *enfin*) and false starts (2.87 and 1.17 phw, respectively). FR040 produces very few of them (0.50 and 0.07 phw) and could arguably be said to be more fluent on this dimension.

6-9: Positive Component 5 score (FR018)

FR018: **it's enfin** I was afraid to live here because I I like to stay in my student room during the weekend

A: yeah I see

FR018: and er (0.840) **enfin** (0.250) I (0.540) I just thought yes in Louvain-la-N= in Louvain la Neuve when you stay during the weekend th= there is no **enfin** there is no one (0.450) and erm (0.450) and during the week you only see well er (0.230) teachers **enfin** professors

6-10: Negative Component 5 score (FR040)

FR040: what really impressed me in in China is that **they are** (0.310) well of course the number of people but everybody knows about that actually I've been to Hong Kong and then to (0.210) Kwangtung (0.540) Can= Canton […]

FR040: you have loads of people going er with the trolley (0.340) em (0.370) heating trolleys (0.250) and er with food in it and you had just a bill (0.230) and er they they put a stamp on it (1.310) do you see what I mean (0.320)

A: not exactly (0.380) no

FR040: erm well I'm afraid it's the Dutch word stamp er **tampon**

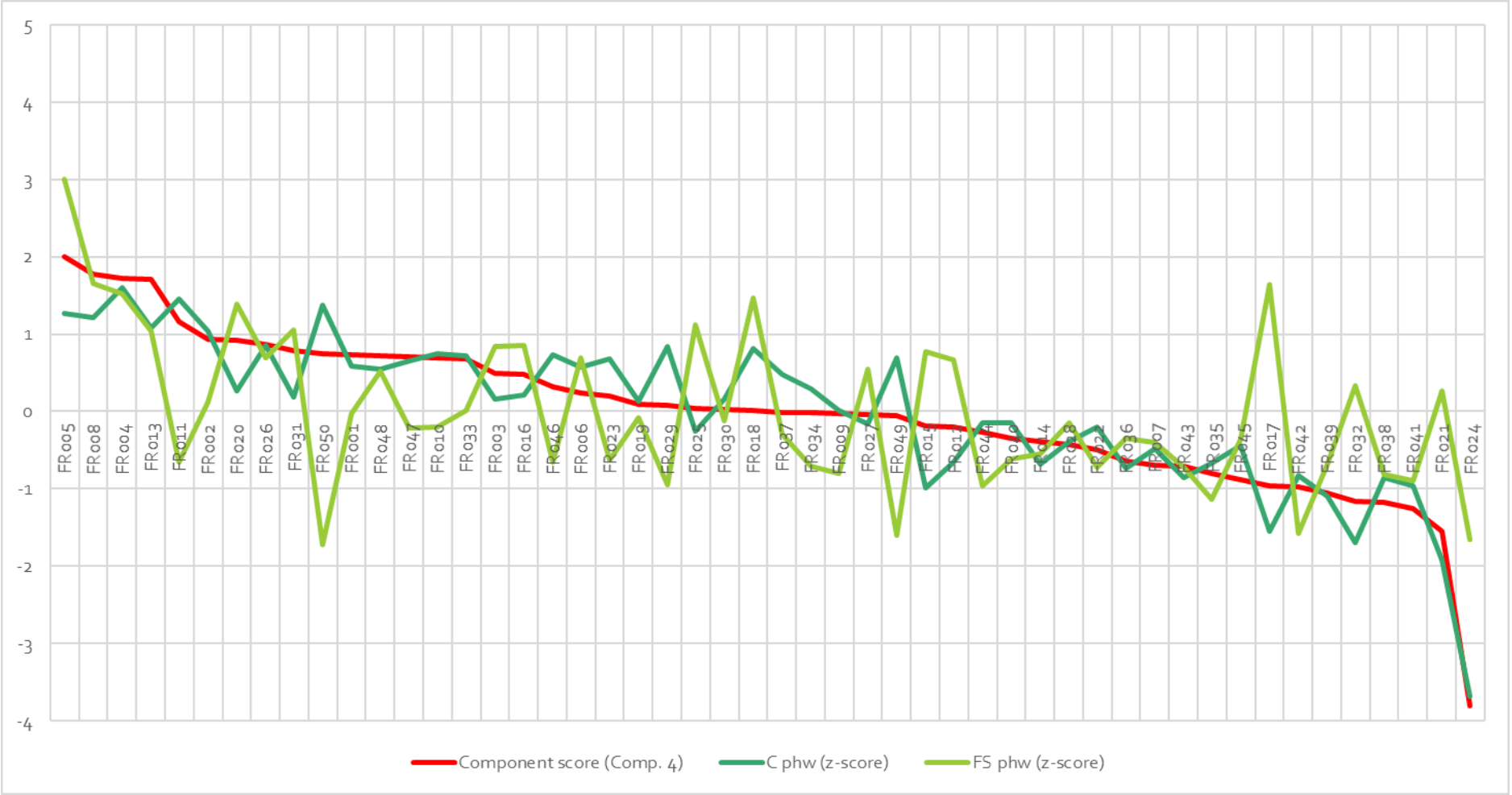


Figure 6-5: Constituent variables of Component 4 in LINDSEI-FR+

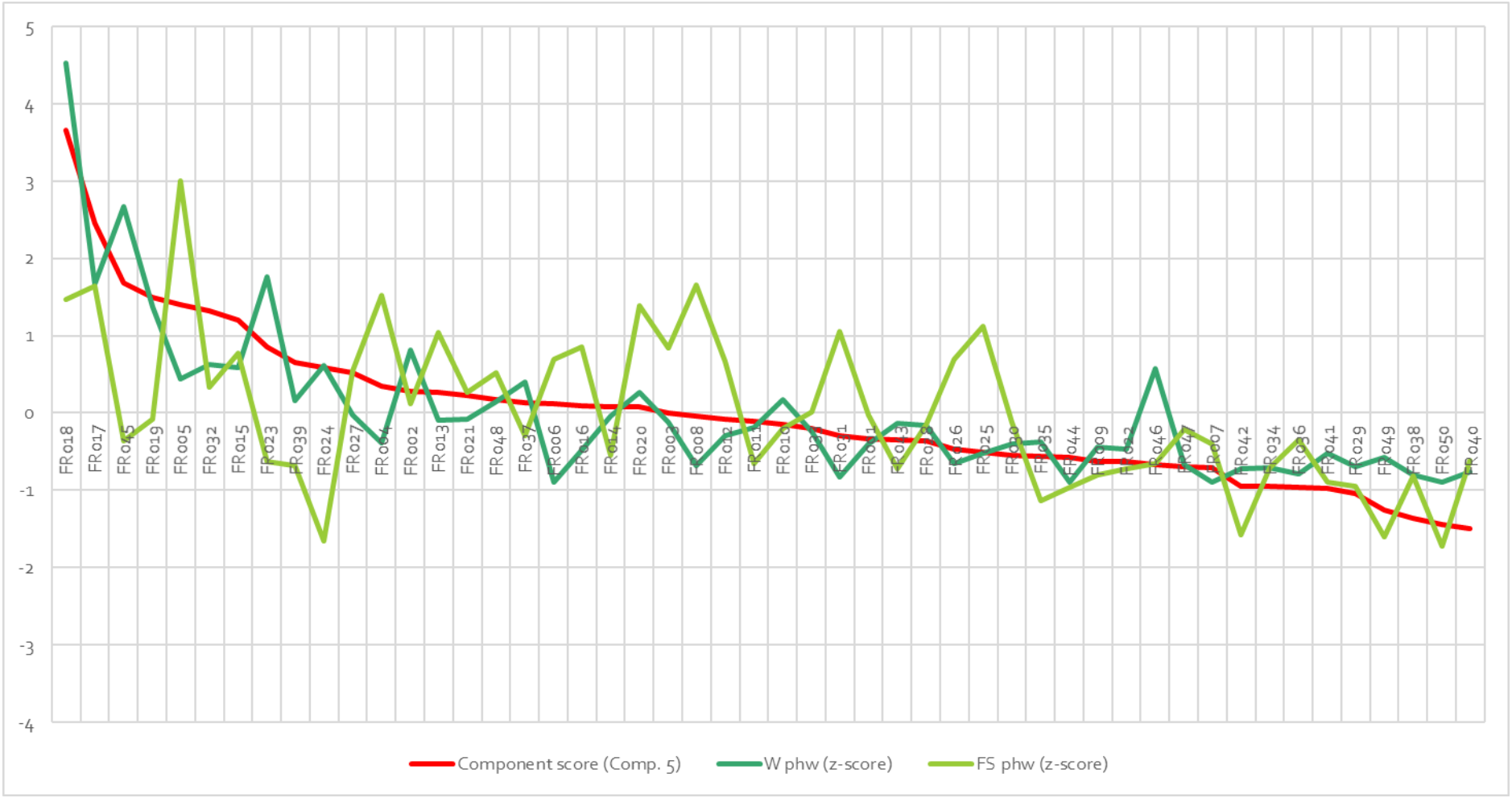


Figure 6-6: Constituent variables of Component 5 in LINDSEI-FR+

6.1.2 (Dis)fluency dimensions in the native corpus

In an attempt to see how L2 (dis)fluency compares with, and the extent to which it differs from, native English speech, a separate Principal Components Analysis was run on the native speaker data from LOCNEC+.

As with the learner data, I started by including all 14 (dis)fluency measures (PA1b). From the screening of the correlation matrix, it appeared that one variable, foreign words, did not significantly correlate with any other variable. It was thus excluded from the analysis. PA2b with the 13 remaining variables had an overall KMO statistic (the measure of sampling adequacy) below the minimum threshold. The analysis was re-run (PA3b) without the variable mean length of unfilled pauses (MLUP) (and without foreign words) because MLUP had the lowest KMO compared to the other (dis)fluency variables (incidentally, it is noteworthy that MLUP was also excluded from the learner PCA for the same reason). The preliminary analysis with the 12 remaining variables held a KMO statistic of > .5 and a significant Bartlett test of sphericity value (p = .000). Oblique rotation (direct oblimin) was used to increase the interpretability of the components but the factor correlation matrix, which indicates the correlations between the extracted rotated components, revealed that correlations between components were on average very small. For this reason, the final analysis was run with orthogonal rotation on the 12 (dis)fluency variables[139].

The final PCA was run on **12 NS (dis)fluency variables** (i.e. excluding foreign words and mean length of unfilled pauses) with **orthogonal rotation** (varimax). The KMO measure reached .66, which is within the acceptable range suggested by Field (2013:685). Based on Kaiser's criterion, eigenvalues over 1 were obtained for four components. Together, these 4 components explain 71.28% of the variance in the native speaker data, which is slightly less than what was obtained for the 5 NNS components but still very satisfactory. The scree plot (to be found in Figure 9-2 in Appendix 9.7) showed an inflexion point that would justify retaining three components, but I retained **4 components** because of the convergence between Kaiser's criterion and the scree plot on this value.

[139] As was the case for the learner data, I also performed two Factor Analyses on the 12 (dis)fluency variables with orthogonal rotation. The results were overall very similar to the PCA reported here. The main differences pertained to the loadings of lengthenings, false starts and (to a lesser extent) repetitions, which were noticeably lower.

	Rotated factor loadings			
(Dis)fluency variables	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Mean length of runs	**.876**	.077	-.022	.096
Unfilled pauses	**-.866**	.134	-.112	-.024
Phonation-time ratio	**.861**	-.193	-.043	.158
Speech rate	**.804**	.100	.157	-.062
Filled pauses	**-.547**	.231	**-.408**	.336
Truncations	-.043	**.916**	.116	-.078
Restarts	.021	**.875**	.182	.113
Repetitions	-.232	**.559**	-.380	.122
Lengthenings	.021	**.496**	-.379	**.457**
Discourse markers	.132	.080	**.823**	.063
Conjunctions	.134	-.027	.053	**.857**
False starts	-.075	.218	**.525**	**.530**
Eigenvalues	3.590	2.426	1.348	1.189
% of variance	29.920	20.216	11.233	9.910

Table 6-2: Summary of PCA for the LOCNEC+ data
Note: factor loadings over .40 appear in bold; the variables are ranked in decreasing order of loading

Table 6-2 below provides a detailed overview of the factor loadings of each (dis)fluency variable on the four components[140], and Figure 6-7 presents a summary of the constituent variables of each component. Figure 6-8 to Figure 6-11 below provide a graphical illustration of each speaker's performance for the constituent variables of the four components.

The inspection of the (dis)fluency variables that load on each component suggests that **the 4 NS components are comparable to the first four learner (dis)fluency components,** though there is **no perfect correspondence**.

[140] The factor loadings before rotation can be found in Appendix 9.7 – Table 9-9.

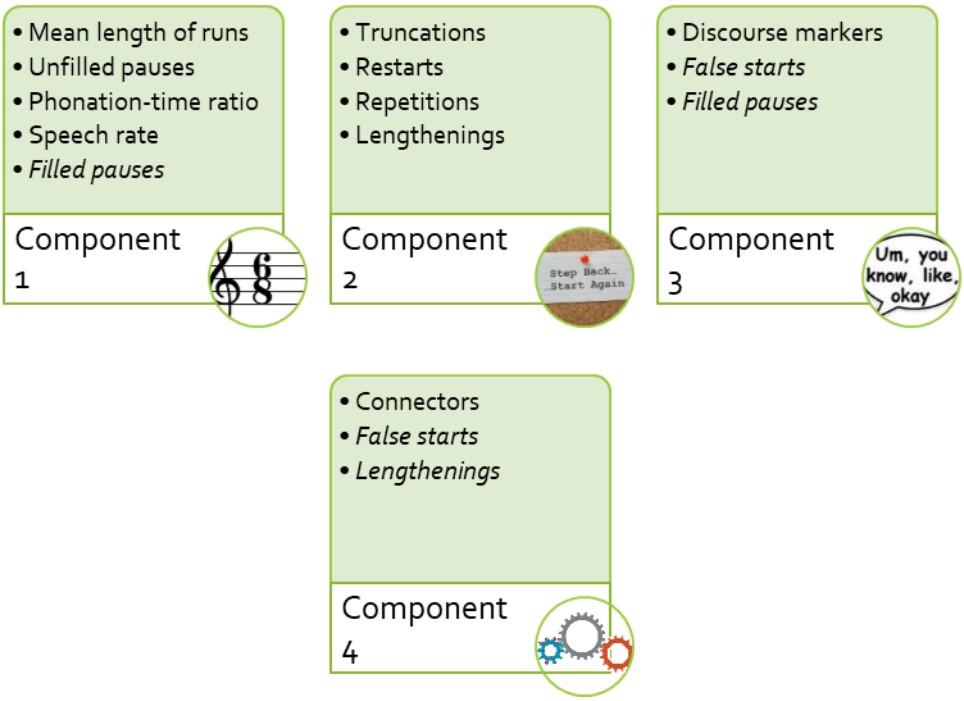


Figure 6-7: The 4 native (dis)fluency components
Note: Italics indicate lower loadings

As Table 6-2 shows, **Component 1** highly correlates with five (dis)fluency variables, namely (in decreasing order of absolute loading) **mean length of runs, rate of unfilled pauses, phonation-time ratio, speech rate and rate of filled pauses**. The examination of the loadings reveals that Component 1 increases with longer runs, higher phonation-time ratio and speech rate, and with decreasing rate of unfilled and filled pauses. It is worth underlining that the four variables that have the highest loadings on this component also made up NNS Component 1, but, whilst the learner temporal (dis)fluency component only had four constituent variables, this component also includes a significant loading for a fifth variable, namely filled pauses. This suggests that learners and native speakers differ in their use of filled pauses but it may also partly support the grouping of the two types of pauses in native speech. Note also that native filled pauses also slightly load on Component 3, together with discourse markers (as was the case in the learner corpus).

The second native speaker component is made up of four high loadings on measures of rate of **truncations, restarts, and repetitions** (the same variables as for NNS Component 2), and also on the measure of the rate of **lengthenings**[141]. It was underlined before that a common characteristic of the constituent variables of NNS Component 2 was that truncations, restarts and repetitions are all made up of two parts (the reparandum and the repair), with some exceptions for truncations. Lengthenings obviously do not match this two-part structure.

[141] Lengthenings were excluded in the learner data in the preliminary steps of the PCA because they did not correlate with any other variable.

However, it might be hypothesised that native lengthenings are used as (clues of) covert repairs, as defined by Levelt (1983).

Component 3 in the native speaker data is highly correlated with **discourse markers.** It is also correlated with **false starts** and with **filled pauses**. The association of discourse markers and false starts is quite surprising, and the positive loading of false starts is another perplexing factor for the interpretation of the underlying dimension represented by this component because it seems to suggest that, contrary to the commonly-held view, a high rate of discourse markers may not be such a good indicator of fluency. While it was argued that NNS Component 3, which also included discourse markers and filled pauses, might correspond to the domain of pragmatic (dis)fluency, the nature of false starts does not seem to fit with this interpretation. That being said, because false starts are extremely difficult (if not impossible) to elicit and particularly infrequent, very little research has been carried out on this phenomena so far, and whether they might have a pragmatic dimension too, like discourse markers and filled pauses, still remains uncharted territory.

The last dimension identified by the PCA in the native speaker data is made up of three (dis)fluency variables, namely **conjunctions**, **false starts** and **lengthenings** (but the last two also load on other components). This component partially corresponds to the "discourse cohesion" component in the learner data (which only included conjunctions and false starts).

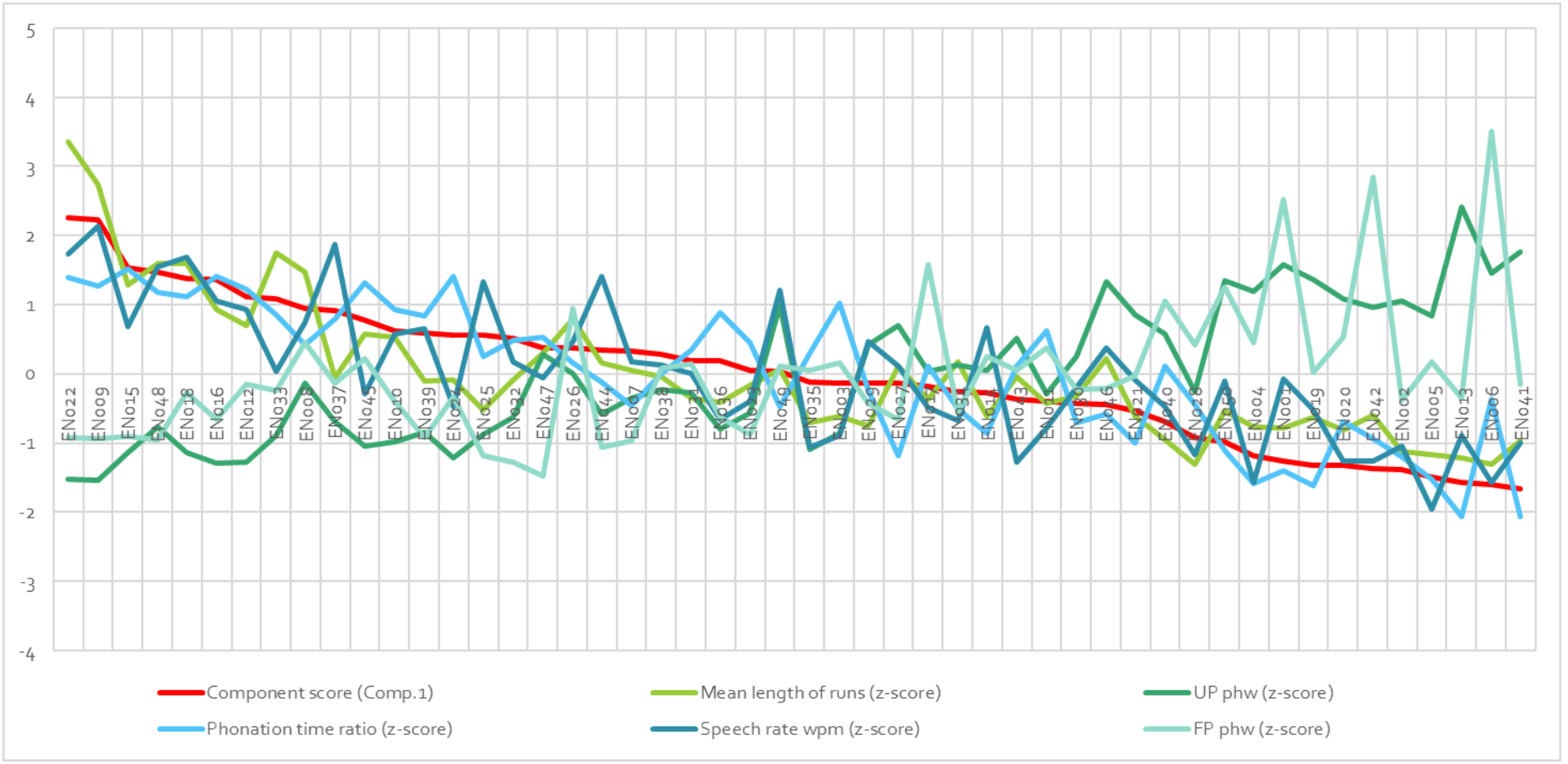


Figure 6-8: Constituent variables of Component 1 in LOCNEC+

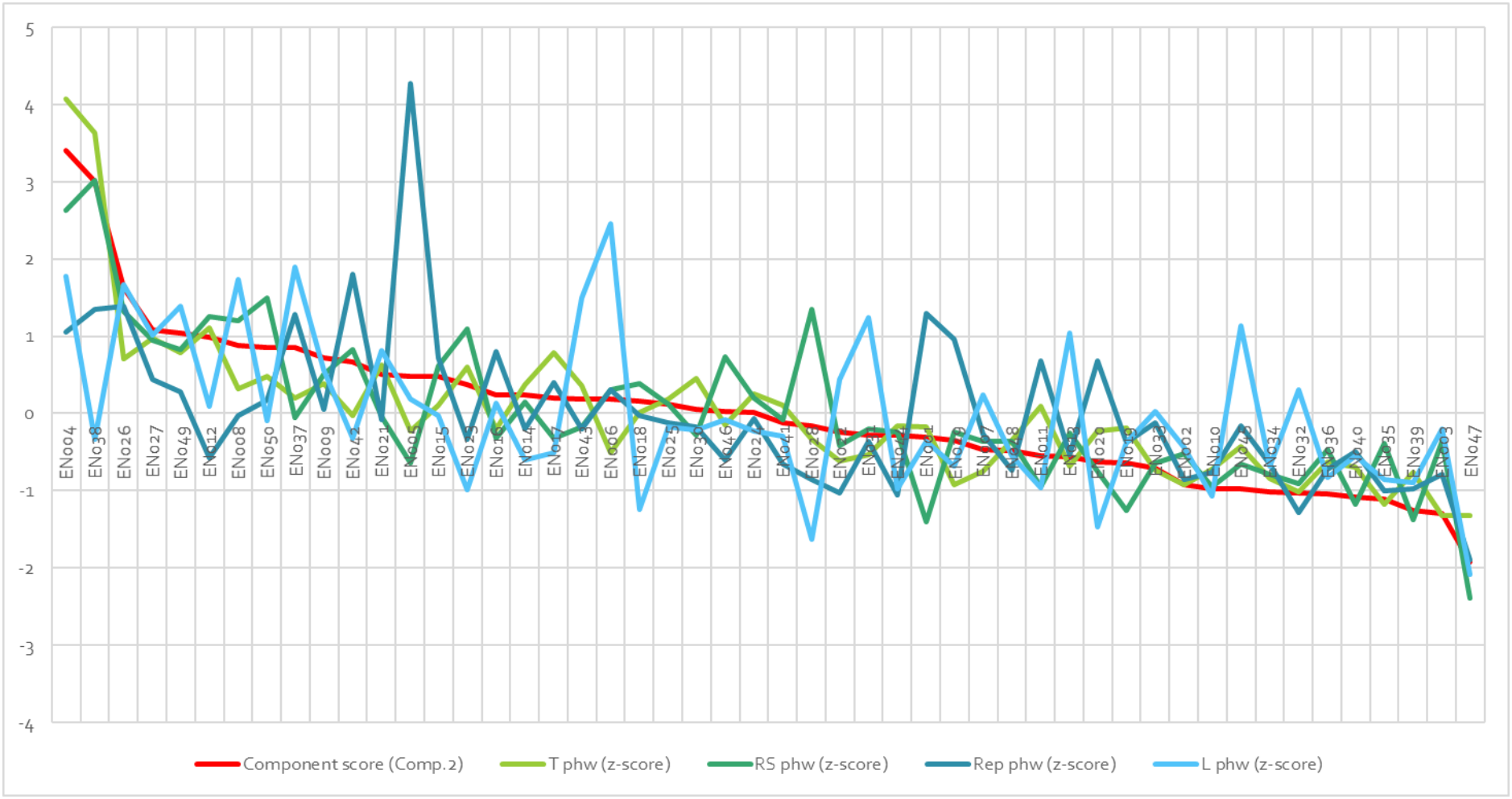


Figure 6-9: Constituent variables of Component 2 in LOCNEC+

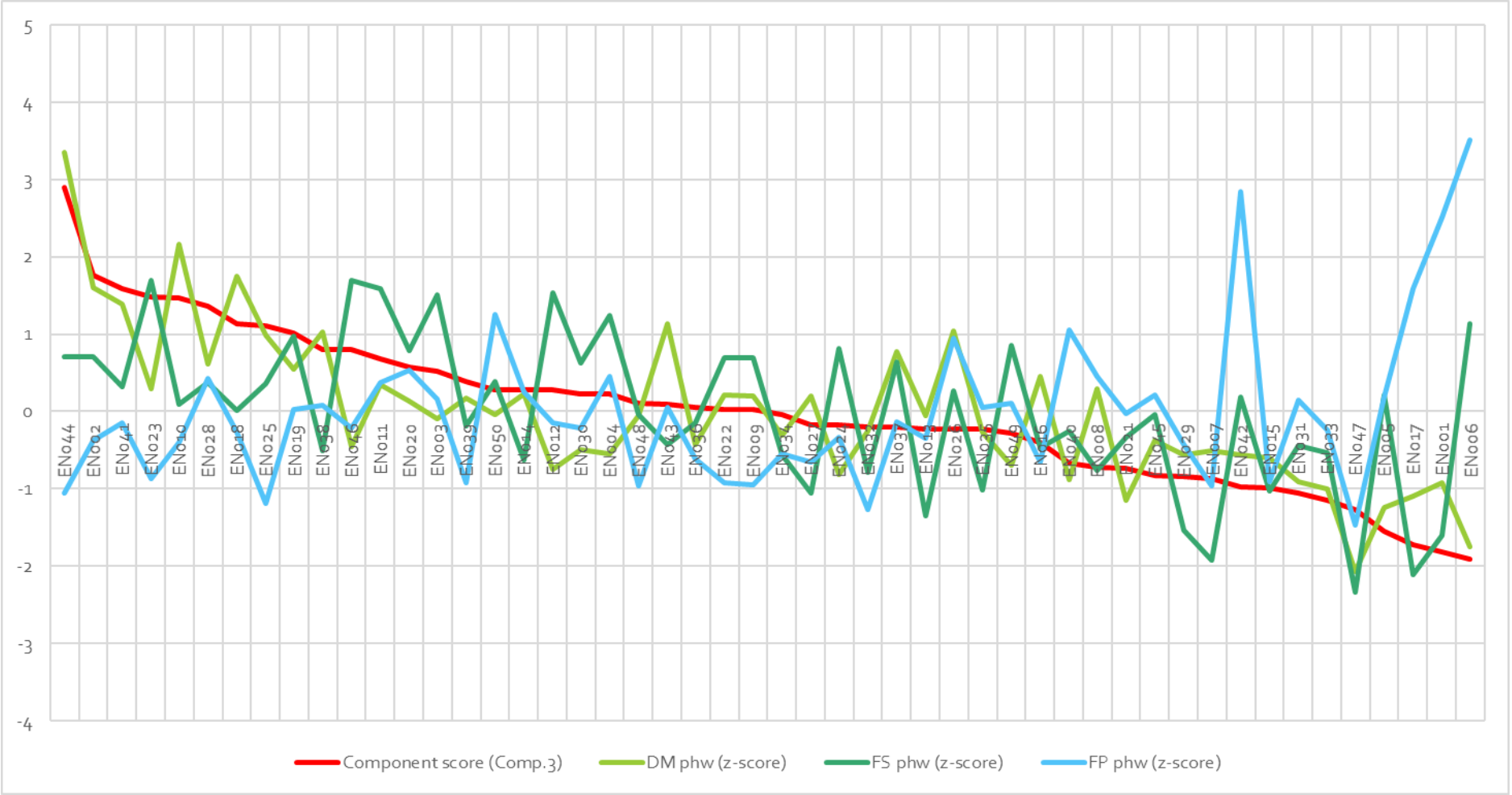


Figure 6-10: Component variables of Component 3 in LOCNEC+

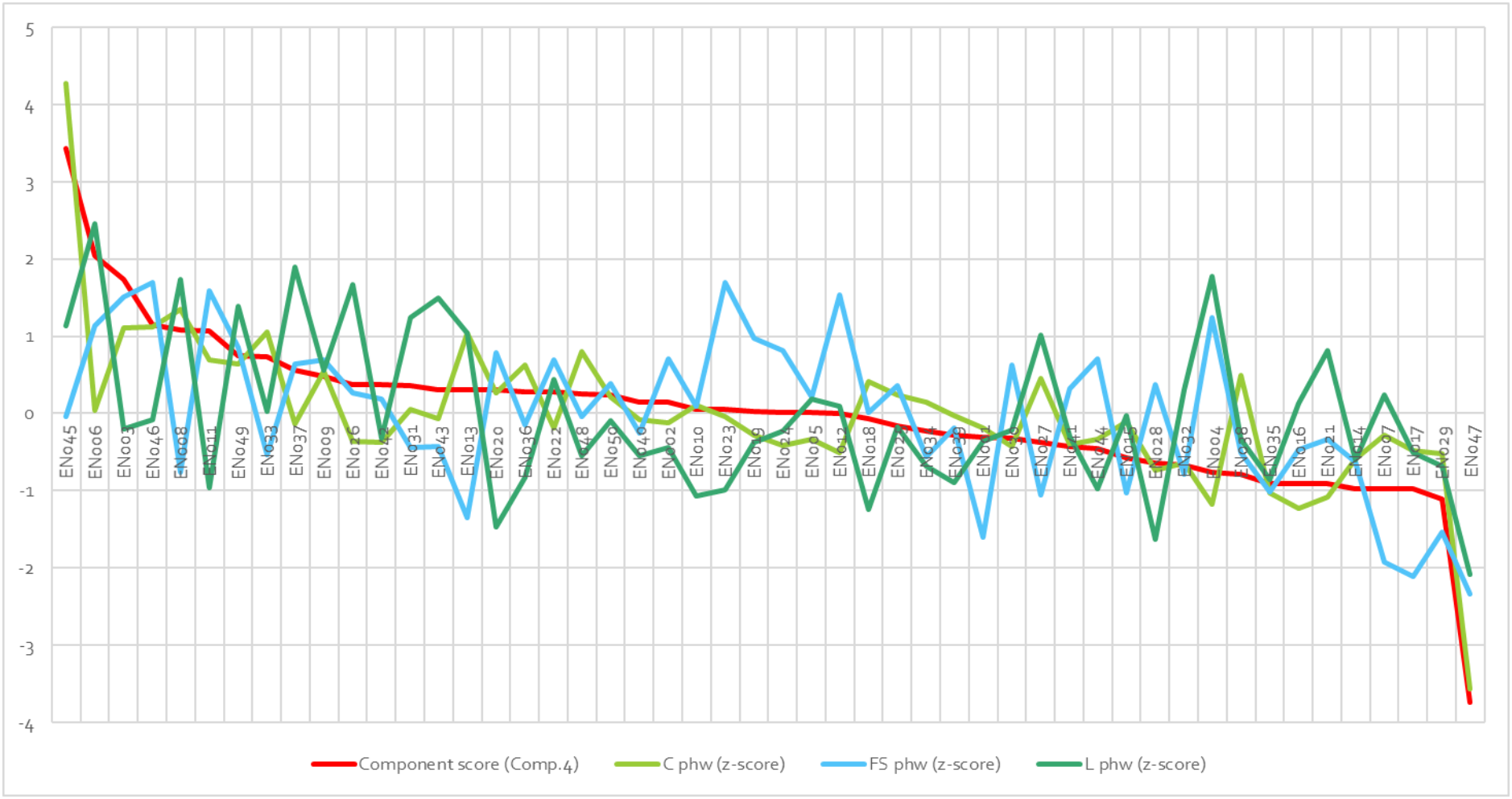


Figure 6-11: Component variables of Component 4 in LOCNEC+

6.1.3 Bringing together the (dis)fluency components

Some insightful similarities and differences come up from the **comparison of the learner and native speaker components**. Table 6-3 shows the constituent variables with a loading > .40 ("✓") for each component; similarities are highlighted with a dark background.

The most interesting difference pertains to **filled pauses** (FPs). The findings highlight that, unlike native FPs, learner FPs are *not* associated with unfilled pauses but are grouped together with discourse markers (this association also exists in LOCNEC+, but is far weaker). This seems to suggest a different use of *ers, uhs, uhms* and other filled pauses by learners as compared to native speakers, and further analyses could investigate, e.g., whether learner filled pauses are perhaps utilised in learner speech with similar functions as those traditionally given to discourse markers (in this respect, see e.g. Crible *et al.* (2017)). In addition, the dissociation of filled and unfilled pauses in the learner factor structure does not support the mingling of filled and unfilled pauses in a single macro category of pauses (as in e.g. Mehnert 1998) and, thereby, also argues against the use of misleading terms such as "pauses" or "pausing" without further specification.

Another difference between the learner and native factor structures pertains to **lengthenings**. In the learner data, lengthenings were excluded in the preliminary steps of the PCA because they did not correlate with any other variable – in a sense, they could also be seen as making up a component of (dis)fluency on their own. In the native data, lengthenings load on NS Components 2 and 4. It was suggested that lengthenings could be the trace of covert repairs (*cf.* Levelt 1983) in native language, unlike in learner speech. Lastly, Component 3 differs by two constituent variables in the learner and native speaker data: speech rate loads slightly on the learner version of the component, and false starts constitute the third constituent variable in the native counterpart.

Despite those dissimilarities, it is striking that **Components 1 to 4 are, overall, very similar in the learner and native speaker data**, which suggests that learner and native speaker (dis)fluency are not so different in essence. As can be observed in Table 6-3, which summarises the (dis)fluency variables that load on the learner and native components, the similarities are remarkable, with the majority of the variables being the same in the learner and native version of each component.

Table 6-4 then recapitulates the tentative interpretation of the component scores in terms of – arguably – higher (↗) or lower (↘) degree of fluency. Speakers with high positive scores on Components 1 and 3, and with high negative scores on Components 2, 4 and 5 are more fluent on those dimensions. These tentative interpretations merit further scrutiny, and Chapter 7 will probe further into the relationship between component scores and perceived CEFR fluency level, which is a hitherto an unexplored avenue of research.

(Dis)fluency variables	Comp. 1		Comp. 2		Comp. 3		Comp. 4		Comp. 5
	NNS	NS	NNS	NS	NNS	NS	NNS	NS	NNS
Conjunctions							✓	✓	
Discourse markers					✓	✓			
False starts						✓	✓	✓	✓
Filled pauses		✓			✓	✓			
Foreign words									✓
Lengthenings				✓				✓	
Mean length of runs	✓	✓							
Phonation-time ratio	✓	✓							
Repetitions			✓	✓					
Restarts			✓	✓					
Speech rate	✓	✓			✓				
Truncations			✓	✓					
Unfilled pauses	✓	✓							

Table 6-3: Bringing together the learner and native speaker components

Component scores	Comp. 1		Comp. 2		Comp. 3		Comp. 4		Comp. 5
	NNS	NS	NNS	NS	NNS	NS	NNS	NS	NNS
High positive score	↗	↗	↘	↘	↗	↗	↘	↘	↘
High negative score	↘	↘	↗	↗	↘	↘	↗	↗	↗

Table 6-4: The interpretation of the components scores

6.1.4 Discussion and limitations

From a methodological point of view, it is important to underline the **consistency between** the results obtained with **the Principal Components Analysis and the Factor Analysis**: the two methods resulted in the same groupings of variables in both the learner and the native speaker data, and, overall, the loadings differed by (very) little. This provides some confidence in the **robustness of the factor structures** that were found.

The factorial analysis of learner and native speaker data highlighted some key underlying dimensions of (dis)fluency. Two dimensions in particular – **temporal and repair (dis)fluency** – **corroborate previous findings** (e.g. Mehnert 1998; Skehan 1999; Tavakoli & Skehan 2005b) and showed that these two dimensions can (and should) be kept separate in future investigations. The first component, both in the learner and the native speaker data, loads on measures of speed and silence. This component seems to capture the degree and extent to which time is filled with talk (*cf.* also Fillmore 1979). The other component, termed repair (dis)fluency, is tapped by measures of truncations, restarts, and repetitions. This component seems more "connected to moment-by-moment decisions during performance, reflecting adjustments and improvements that are feasible within the pressure of real-time communication" (Foster & Skehan 1999:229).

The analysis also revealed that the learner and the native speaker components are orthogonal to one another (see also the scatterplots presented in Figure 9-3 and Figure 9.4 in Appendix 9.7). This means that the **learner components do not correlate with one another**, and **neither do the native speaker components**. However, in line with Götz's (2013a) idea of fluency profiles, a closer examination of the component scores and/or the constituent variables could reveal groups of speakers performing very similarly, i.e. speakers who have a similar (dis)fluency profile. The next section (Section 6.2) will go a step further in the identification of such profiles in the learner and native speaker data by presenting the results of a Cluster Analysis.

At the onset of the analysis, some (dis)fluency **variables had to be excluded**. The **mean length of unfilled pauses** was excluded from both the NNS and the NS PCA. This could be interpreted as evidence that the rate of unfilled pauses is more important than their actual length, as claimed by Cucchiarini *et al.* (2000; 2002) and Kormos and Dénes (2004). However, as some studies (esp. Campione & Véronis 2002) pointed to a trimodal distribution of the length of unfilled pauses, it would also be interesting to re-run an analysis with the separate rates of short, medium and long unfilled pauses. In the learner data, **lengthenings** did not correlate with any other (dis)fluency variable and were therefore also excluded from the factorial analysis. The same procedure was followed for their detection and annotation in the two corpora, but it might be that the category was too broadly defined. More specifically, in accordance with the main hypothesis of the project, no distinction was made in the

(dis)fluency annotation between lengthenings occurring as a result of phonological processes (e.g. stressed articles pronounced /ði:/ and /ei/ before a vowel or for a pragmatic purpose) and "unnatural" lengthenings that are used to buy additional time (Clark & Fox Tree 2002; Grosjean 1980c). Maybe these two subcategories should have been kept separate to reach more meaningful results, as was originally done in the LINDSEI-FR and LOCNEC transcription files. In the native speaker data, **foreign words** were excluded from the PCA. While it seems fair to think of foreign words as a strategy typical of learner (dis)fluency, in native speaker speech, they should perhaps rather be approached from the perspective of communication strategies (e.g. as cultural bridges; *cf.* De Cock 2015a; Nacey & Graedler 2013), which are perhaps only peripheral to (dis)fluency.

More generally, the interpretation of the components identified in learner and native speech could benefit from further investigations into under-studied (dis)fluency variables such as false starts, lengthenings and conjunctions. Further research could also seek corroboration for the findings presented for LINDSEI-FR+ and LOCNEC+ with (dis)fluency components in other learner (and native speaker) populations and, thereby, also provide new insights into the cross-linguistic validity of these (dis)fluency dimensions.

6.2 TOWARDS HOLISTIC PRODUCTIVE (DIS)FLUENCY PROFILES

Section 6.1 aimed to get a better understanding of the (dis)fluency construct by examining how the variables that are constitutive of (dis)fluency are structured. The present section takes the perspective of the **individual performances of learners and native speakers** and, rather than assuming a linear relationship between (dis)fluency features, examines whether oral interviews produced by a group of comparable learners (or native speakers) fall into **multiple and significantly different clusters based on the use of (dis)fluency variables**. By "profiles" or "clusters", I mean groups of speakers that display high within-group similarities with respect to the (dis)fluency variables while simultaneously showing large between-group differences with respect to one or more of these variables.

To the best of my knowledge, only one large-scale study has previously investigated learner and native speaker (dis)fluency profiles. Götz (2011; 2013a), on the basis of an operationalisation of (dis)fluency in 8 variables, found 3 clusters of native speakers, and 3 clusters of German-speaking learners of English (see Section 1.2.5 for a full report of her results). In line with her study, **two Agglomerative Hierarchical Cluster Analyses** were run to explore whether multiple profiles (i.e. clusters of speakers) emerge among the speakers of LINDSEI-FR+ and of LOCNEC+ (Sections 6.2.1 and 6.2.2). Learner and native-speaker profiles are discussed contrastively in Section 6.2.3. Although it might have been possible to run such an analysis based on component scores, I decided to use the observed (dis)fluency variables instead. Not only does this ease the interpretation considerably, but it also prevents several problems, such as the loss of information due to the use of pre-processed data (see e.g. Dolnicar & Grün 2008; Mooi & Sarstedt 2010).

Before engaging in this section, it is important to underline that Cluster Analysis is a largely **exploratory and descriptive technique** and that the number and composition of the clusters inevitably varies depending on the variables included, their measurement, the clustering algorithm, the distance measure, etc. Different methodological choices might thus result in slightly different findings.

The key statistical terms are summarised below.

Key statisticalese	
Agglomerative hierarchical cluster analysis	Type of cluster analysis where cases (here, speakers) are aggregated into increasingly large clusters based on a measure of similarity.
Squared Euclidean distance	One of the measures of similarity that estimates the distance between pairs of individuals.
Dendrogram	Also called tree graph. A graphical representation of the results of a clustering procedure. The vertical axis shows the individuals and the

	horizontal axis consists of the number of clusters formed at each step.
Fusion coefficient	A measure of the relative distance between the clusters formed at each step in a hierarchical cluster analysis. The larger the distance, the less similarity there is between cases that have been clustered together. The coefficients are displayed in the agglomeration schedule.
Ward's method	An agglomerative hierarchical clustering procedure whereby clusters with the greatest similarity are combined.

6.2.1 Learners' (dis)fluency profiles

A pre-test revealed that all 14 (dis)fluency variables (*cf*. Table 3-6; transformed into *z*-scores (Staples & Biber 2015:253–254)) could be subjected to the Cluster Analysis as there was no correlation > .9 (*ibid*.). The Cluster Analysis with Ward's clustering procedure (*cf*. Section 3.4.3.3) starts with each speaker being a singleton cluster. At the first step, the two learners who have the smallest distance measure are combined in a single cluster. At the second step, either a third speaker is added to that cluster, or two other speakers are joined into a new cluster. The process is repeated until each speaker is added to an existing cluster and, eventually, all clusters are aggregated into one final, all-encompassing, cluster. The process is illustrated in the dendrogram displayed in Figure 6-12: in the first step, the algorithm built a cluster from FR022, FR030 and FR041. The second step resulted in the clustering of FR025 and FR038 etc.

The identification of the appropriate number of clusters is based on the observation of the fusion coefficients in the agglomeration schedule output of SPSS [142] as well as on the examination of the distribution and distances of cluster combinations in the dendrogram[143]. The fusion coefficients show an abrupt discontinuity in the composition of clusters for a 2, 3 and 6-cluster solution. Likewise, the examination of the cluster combinations in the dendrogram reveals that there are two main clusters, and possible further sub-divisions.

[142] This is also known as the "elbow method". There also exist more complex statistical procedures and indices to help choose the optimal number of clusters, such as the average silhouette width (Rousseeuw 1987; Kaufman & Rousseeuw 1990) or gap statistics (Tibshirani, Walther & Hastie 2001) see Charrad et al. (2014) for a more complete review.

[143] As clearly explained by Divjak & Gries (2006:38), a dendrogam "provides two types of information that can be read off the tree plot from bottom to top. On the one hand, the tree plot shows what is similar and what is different: items that are clustered or amalgamated early are similar, and items that are amalgamated late are rather dissimilar […]. On the other hand, the tree plot gives an indication of how independent the clusters are: the larger the distance between different points of amalgamation, the more autonomous the earlier [cluster] is from the [cluster] with which it is merged later".

Based on these observations, the **2-cluster solution** was deemed optimal for the data at this stage.

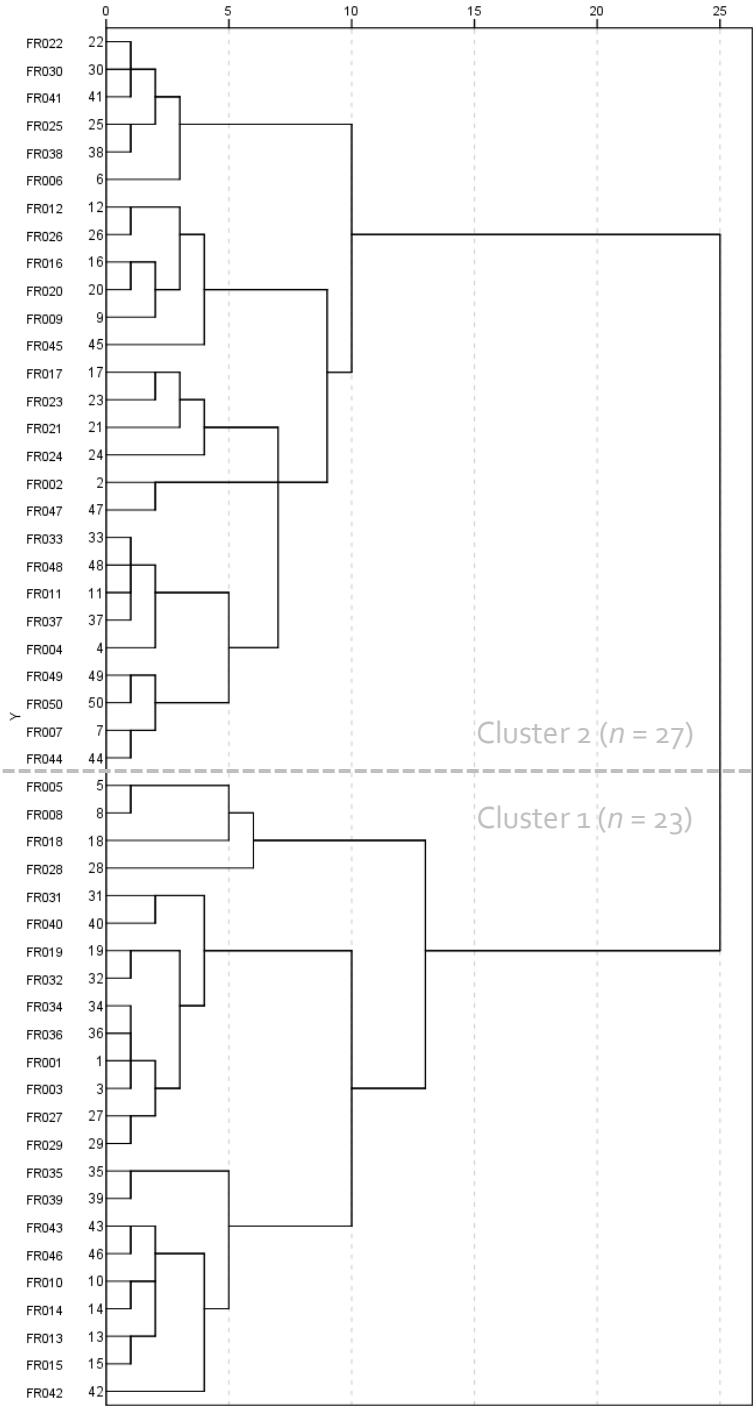


Figure 6-12: Dendrogram obtained from Hierarchical Cluster Analysis: Speaker performances across (dis)fluency variables in LINDSEI-FR+ (Ward's method, Squared Euclidean Distance)

6.2.1.1 The 2-cluster solution

The 2-cluster solution distinguished Cluster 1 (in the lower half of the dendrogram displayed in Figure 6-12), which includes 23 learners from Cluster 2 (in the top part of the graph), which is slightly larger and includes 27 learners[144]. Table 6-5 provides a summary of the mean *z*-scores as well as the standard deviation (*sd*) for each of the 14 (dis)fluency variables in the two clusters. Figure 6-13 is a graphical illustration of the *z*-scores, with Cluster 1 means represented in light green and Cluster 2 in dark green. As can be observed, the two **clusters tend to function as mirror images**: for each (dis)fluency variable, while one cluster has a positive mean *z*-score, the other has a negative mean. For example, Cluster 1 has a mean of 0.74 for the variable mean length of runs and Cluster 2 has a mean of -0.63 for the same measure.

Cluster 1 is characterised by positive scores for nine (dis)fluency variables (discourse markers, false starts, filled pauses, foreign words, mean length of runs, phonation-time ratio, restarts, speech rate and truncations) and negative scores for five variables (conjunctions, lengthenings, mean length of unfilled pauses, repetitions, and unfilled pauses). The opposite is true for Cluster 2. In other words, on average, the 23 learners in **Cluster 1** tend to produce a higher rate of discourse markers, filled pauses etc. and speak faster and in longer speech runs than the speakers in **Cluster 2**, who generally speak more slowly, with fewer discourse markers, filled pauses etc., but with more (and longer) unfilled pauses, repetitions, conjunctions and lengthenings.

(Dis)fluency variables	**Cluster 1 (*n* = 23)**		**Cluster 2 (*n* = 27)**		***t*-test**	**Cohen's *d***
	Mean *z*-score	***sd***	**Mean *z*-score**	***sd***		
Conjunctions	-0.01	0.84	0.01	1.14	t = -.05; p = .963	
Discourse markers	0.23	1.14	-0.19	0.84	t = 1.50; p = .141	
False starts	0.10	1.07	-0.09	0.95	t = .65; p = .516	
Filled pauses	0.05	1.06	-0.04	0.97	t = .30; p = .767	
Foreign words	0.09	1.12	-0.07	0.90	t = .55; p = .584	
Lengthenings	-0.20	0.88	0.17	1.08	t = -1.32; p = .194	
Mean length of runs	0.74	0.90	-0.63	0.56	**t = 6.30; p = .000**	d = 1.83
Mean UP length	-0.49	0.67	0.42	1.05	**t = -3.70; p = .000**	d = 1.03
Phonation time ratio	0.71	0.54	-0.61	0.90	**t = 6.41; p = .000**	d = 1.78
Repetitions	-0.07	0.94	0.06	1.06	t = -.45; p = .656	
Restarts	0.27	1.12	-0.23	0.84	t = 1.79; p = .079	
Speech rate	0.63	0.91	-0.54	0.73	**t = 4.96; p = .000**	d = 1.42

[144] Table 9-10 in Appendix 9.8 provides the make-up of each cluster.

Truncations	0.02	1.10	-0.01	0.93	t = .10; p = .924	
Unfilled pauses	-0.73	0.73	0.62	0.75	**t = -6.49; p = .000**	d = 1.82

Table 6-5: Mean z-score and standard deviation (sd) per (dis)fluency variable and independent samples t-test results for the 2-cluster solution in LINDSEI-FR+

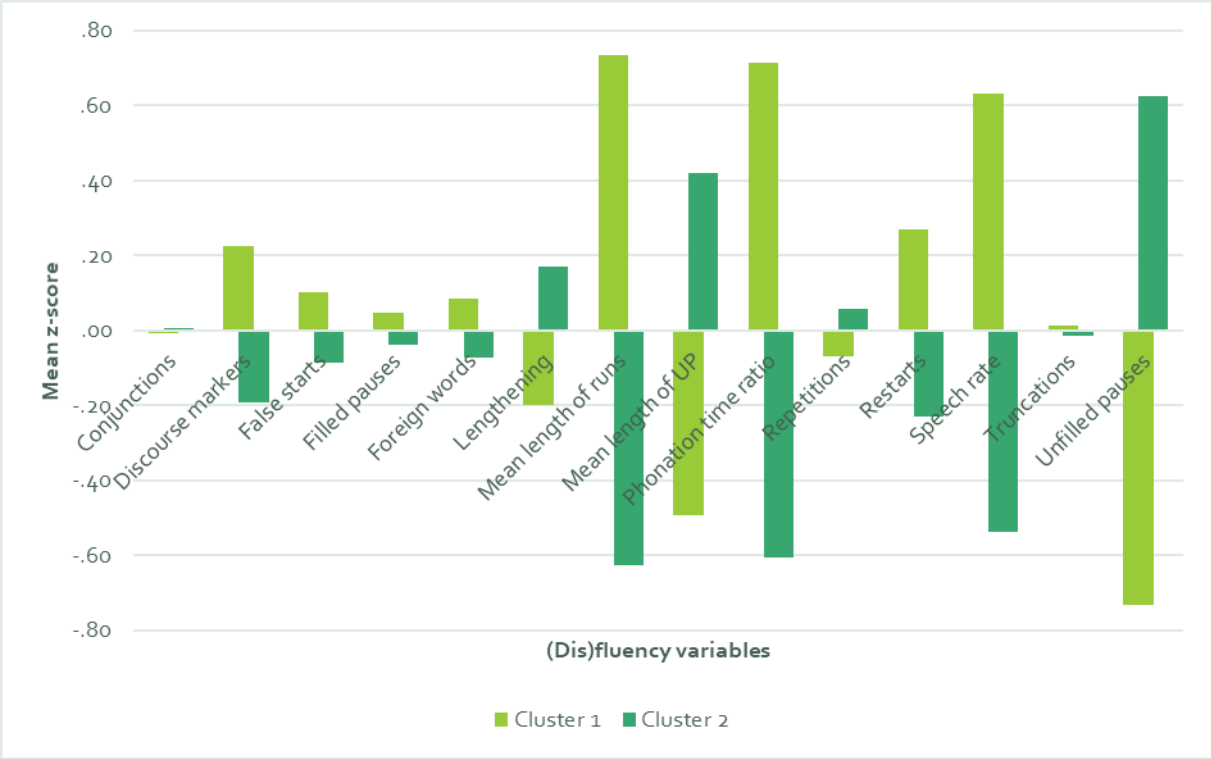


Figure 6-13: Mean z-scores per (dis)fluency variable for the 2 clusters in LINDSEI-FR+

Independent samples *t*-tests revealed statistically **significant differences** between the two clusters for 5 out of the 14 variables, namely the frequency of **unfilled pauses**, **speech rate**, **phonation-time ratio**, **mean length of runs** and **mean length of unfilled pauses** (see bold figures in Table 6-5). This implies that the 23 learners in Cluster 1 produce highly significantly fewer and shorter unfilled pauses, have a higher speech rate, a higher phonation-time ratio, and a higher mean length of runs than the speakers in Cluster 2. Moreover, the effect sizes (Cohens' *d*) of those differences are large to very large ($1.03 \leq d \leq 1.83$; Cohen (1977)). The mean differences for the other 9 (dis)fluency variables did not turn up to be statistically significant.

It is actually not very surprising that it is precisely those variables that discriminate the two main clusters of learners. From the Principal Components Analysis in the first section of this chapter, it already appeared that these variables (except mean length of unfilled pauses, but this variable is clearly closely related to the others) made up the component that explained the largest part of the variance in LINDSEI-FR+, i.e. **Component 1**, which represents the area of temporal (dis)fluency (*cf.* also Table 6-1). As could have been expected, further independent samples *t*-tests revealed that the two clusters of speakers (only) significantly differ with respect to their mean score on Component 1 (see Table 6-6).

(Dis)fluency components	Cluster 1 (*n* = 23)		Cluster 2 (*n* = 27)		*t*-test	Cohen's *d*
	Mean component score	*sd*	Mean component score	*sd*		
Component 1	0.82	0.65	-0.70	0.65	***t* = 8.28; *p* = .000**	*d* = 2.35
Component 2	0.10	1.12	-0.8	0.90	*t* = .62; *p* = .537	
Component 3	0.10	1.13	-0.9	0.88	*t* = .65; *p* = .521	
Component 4	0.08	0.89	-0.07	1.10	*t* = .54; *p* = .594	
Component 5	0.11	1.14	-0.12	0.86	*t* = .79; *p* = .433	

Table 6-6: Mean and standard deviation (sd) per (dis)fluency component and independent samples t-test for the 2-cluster solution in LINDSEI-FR+

In sum, the two main clusters in LINDSEI-FR+ appear to correspond to learners who are temporally more fluent (Cluster 1, with a positive mean score on Component 1) and temporally less fluent (Cluster 2, with a negative mean score on the first component). However, the **large standard deviations** around the means in Table 6-5 and Table 6-6 (even for the 5 significant temporal variables and Component 1) indicate that the within-cluster variance is still very high, which means that the clusters in the 2-cluster solution are still fairly heterogeneous. Therefore, in the next step of the analysis, I also examined the **6-cluster solution** to examine whether more fine-grained usage patterns of (dis)fluency variables could be successfully delineated in LINDSEI-FR+.

6.2.1.2 The 6-cluster solution

Each cluster from the 2-cluster solution is subdivided into three sub-clusters in the 6-cluster solution. Cluster 1 from the 2-cluster solution is subdivided into Cluster A (*n* = 10), Cluster B (*n* = 4) and Cluster C (*n* = 9). Cluster 2 from the 2-cluster solution includes Cluster D (*n* = 15), Cluster E (*n* = 6) and Cluster F (*n* = 6).

The mean *z*-scores of the six clusters[145] for the 14 (dis)fluency variables are displayed in Table 6-7 and graphically illustrated in Figure 6-14 through Figure 6-19 (a summative graph is shown in Figure 6-20). For the sake of completeness, the mean score per (dis)fluency component for the six clusters is included in Table 6-8 and illustrated in Figure 6-21 (more graphs are included in Appendix 9.8).

[145] The speaker IDs per cluster can be found in Appendix 9.8, Table 9-11.

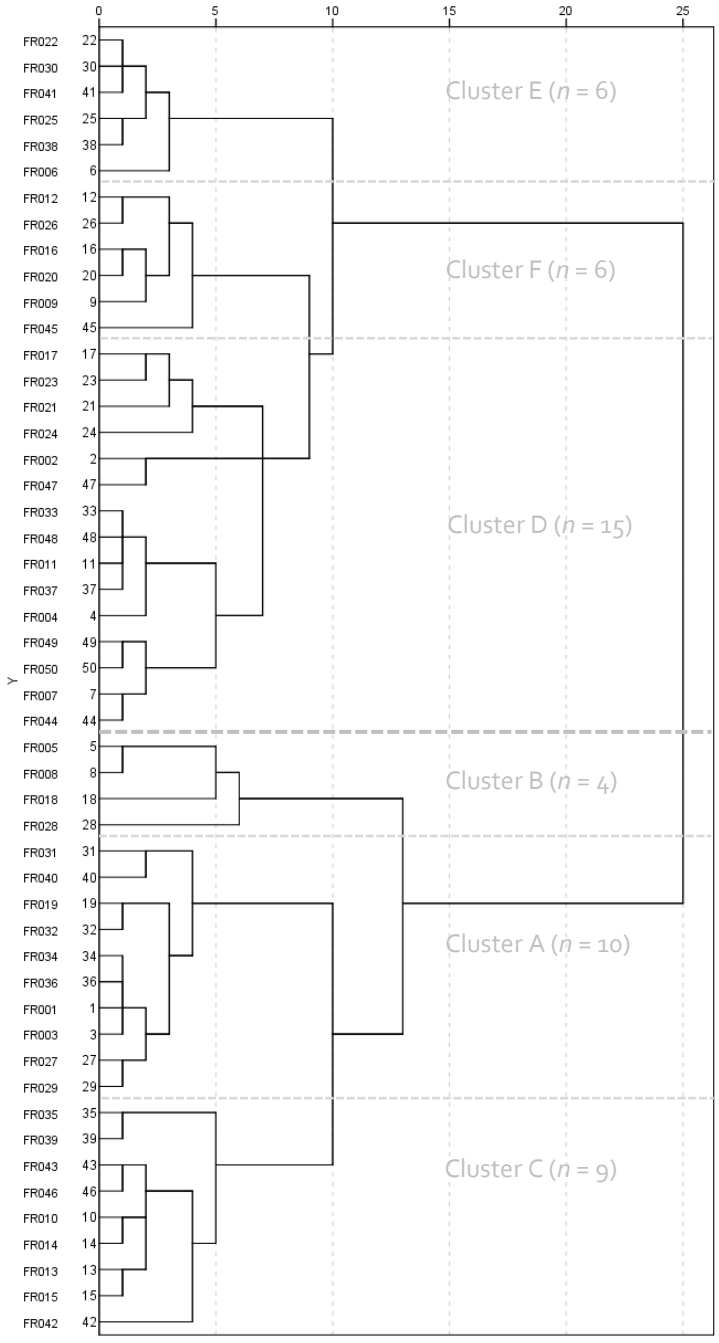


Figure 6-12 (reproduced): Dendrogram obtained from Hierarchical Cluster Analysis: Speaker performances across (dis)fluency variables in LINDSEI-FR+ (Ward's method, Squared Euclidean Distance)

(Dis)fluency variables	Cluster A (*n* = 10)		Cluster B (*n* = 4)		Cluster C (*n* = 9)		Cluster D (*n* = 15)		Cluster E (*n* = 6)		Cluster F (*n* = 6)	
	Mean *z*-score	*sd*	Mean *z*-score	*sd*	Mean *z*-score	*sd*	Mean *z*-score	*sd*	Mean *z*-score	*sd*	Mean *z*-score	*sd*
Conjunctions	-.05	.72	.73	.77	-.28	.87	.10	1.46	-.26	.59	.04	.54
Discourse markers	.95	1.16	.44	.94	-.67	.36	-.43	.87	.39	.89	-.18	.44
False starts	.00	.68	1.50	1.29	-.41	.84	-.27	1.02	-.12	.85	.41	.82
Filled pauses	-.74	.71	-.02	.54	.95	.82	.22	.94	-.07	.87	-.66	.99
Foreign words	-.23	.73	1.03	2.38	.01	.42	.04	.87	-.60	.20	.18	1.26
Lengthenings	-.62	.81	-.37	.58	.34	.83	-.55	.61	.41	.26	1.74	.63
Mean length of runs	.43	.57	.45	1.19	1.20	.96	-.52	.59	-.64	.60	-.88	.44
Mean UP length	-.49	.54	-.76	.44	-.37	.88	.05	.79	1.80	.49	-.06	.90
Phonation time ratio	.60	.52	.38	.42	.99	.52	-.36	.75	-1.55	.70	-.29	.88
Repetitions	-.40	.64	1.15	1.13	-.24	.76	-.39	.81	1.17	.72	.07	1.20
Restarts	-.47	.65	1.78	.96	.41	.91	-.07	.96	.00	.42	-.86	.49
Speech rate	.80	.77	.60	1.55	.46	.80	-.43	.92	-.64	.36	-.72	.40
Truncations	-.51	.40	1.67	.33	-.14	1.17	.10	.58	.67	1.05	-.99	.83
Unfilled pauses	-.67	.62	-.03	1.03	-1.11	.46	.59	.88	.62	.70	.71	.49

Table 6-7: Mean z-scores and standard deviations (sd) per (dis)fluency variable for the 6-cluster solution in LINDSEI-FR+

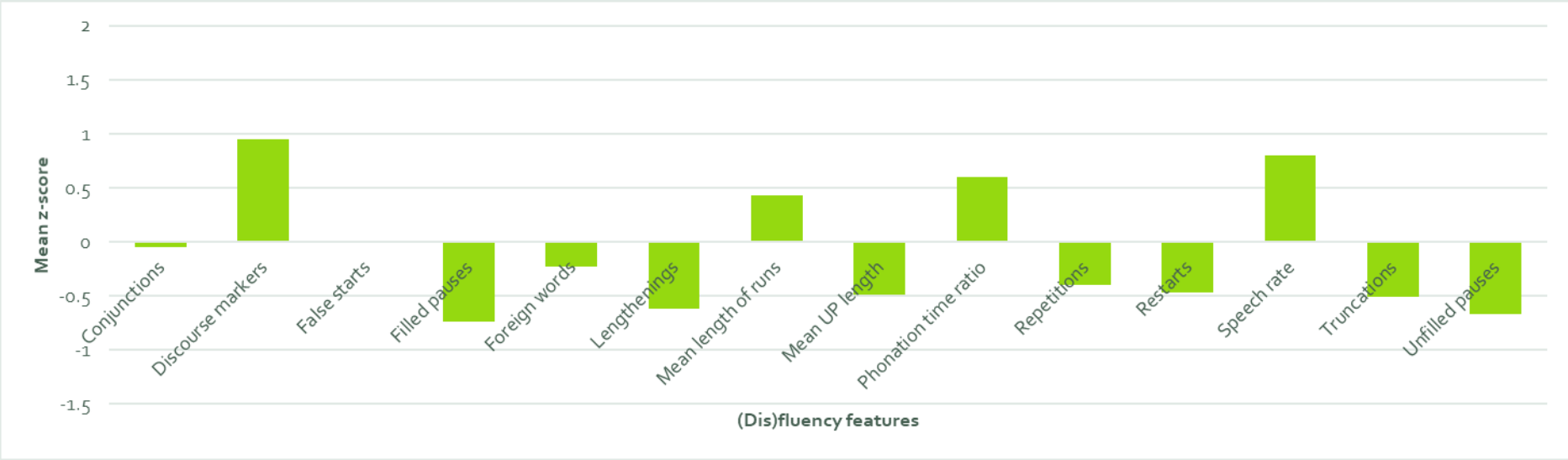


Figure 6-14: Cluster A profile in LINDSEI-FR+

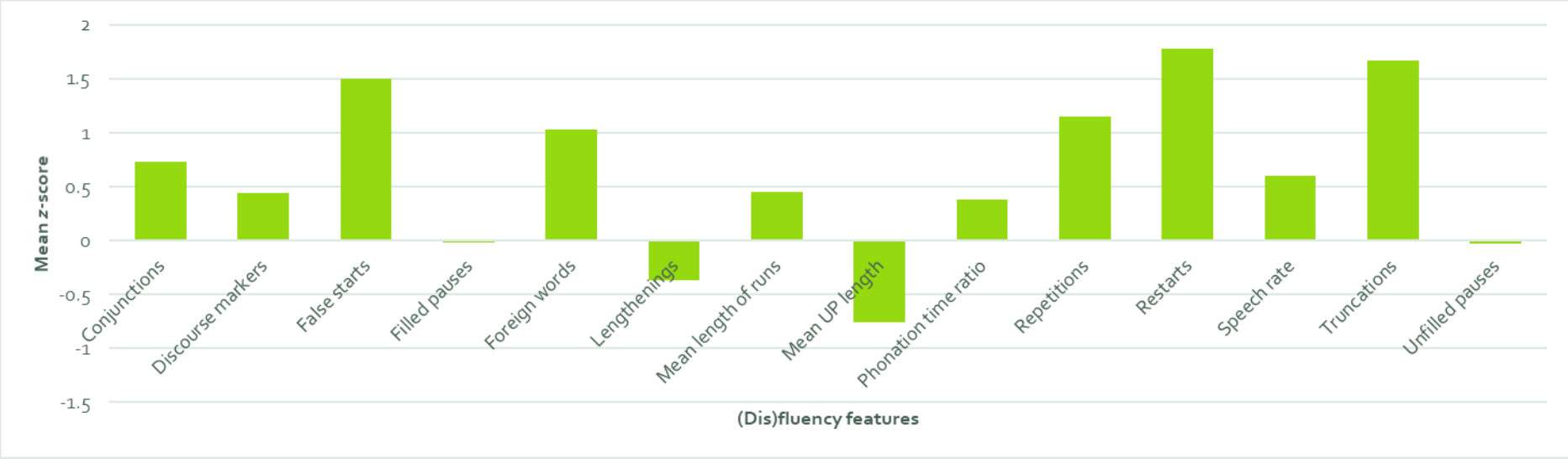


Figure 6-15: Cluster B profile in LINDSEI-FR+

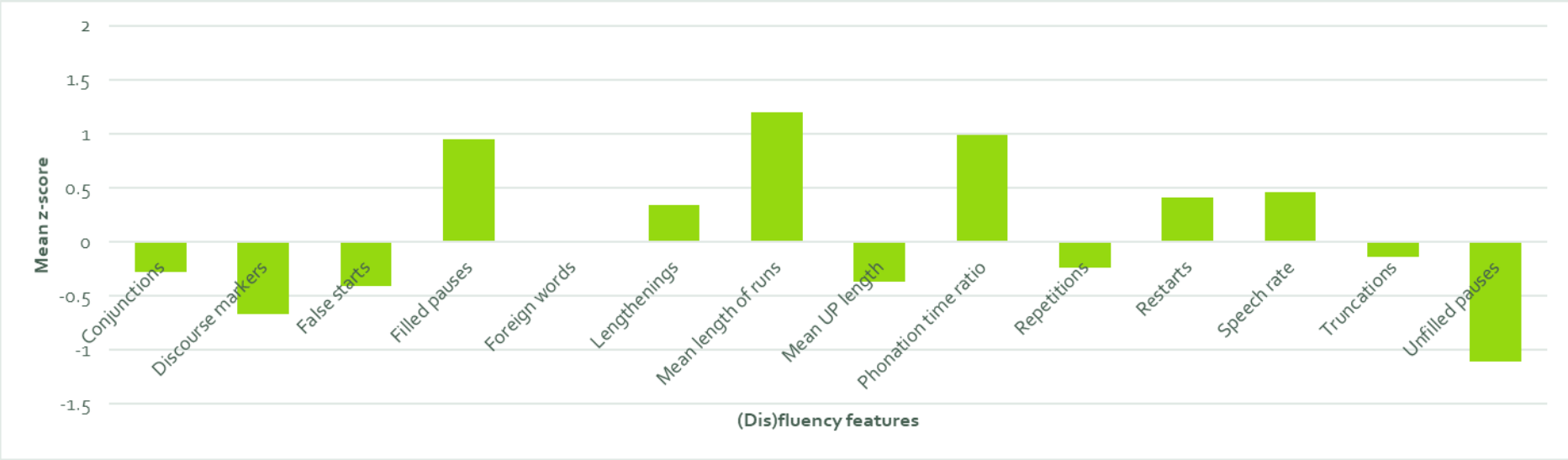


Figure 6-16: Cluster C profile in LINDSEI-FR+

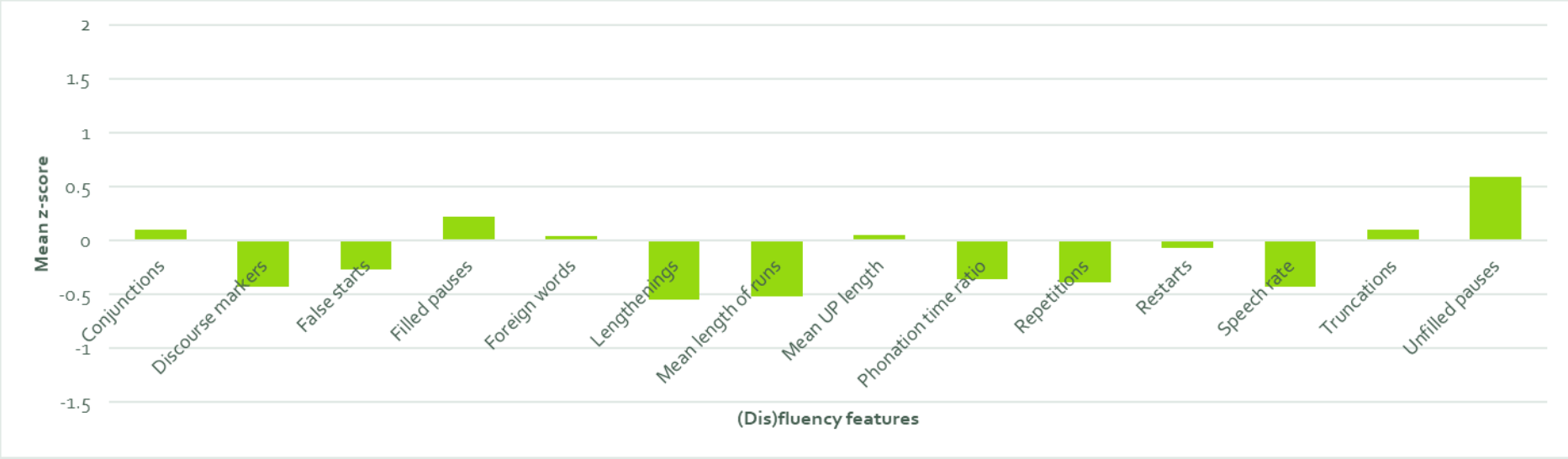


Figure 6-17: Cluster D profile in LINDSEI-FR+

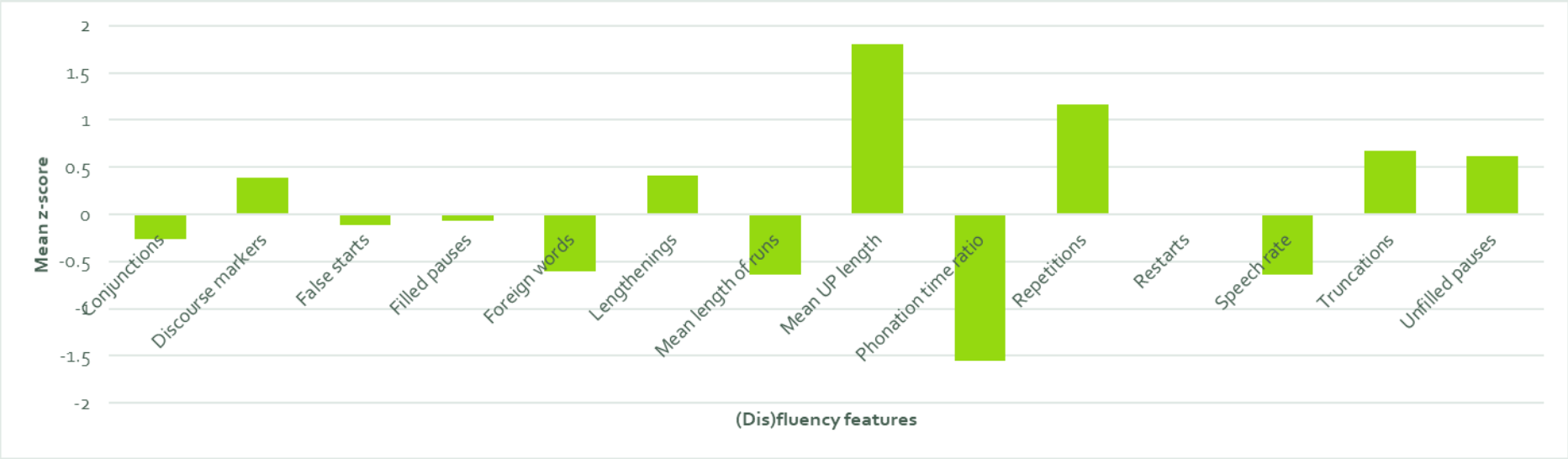


Figure 6-18: Cluster E profile in LINDSEI-FR+

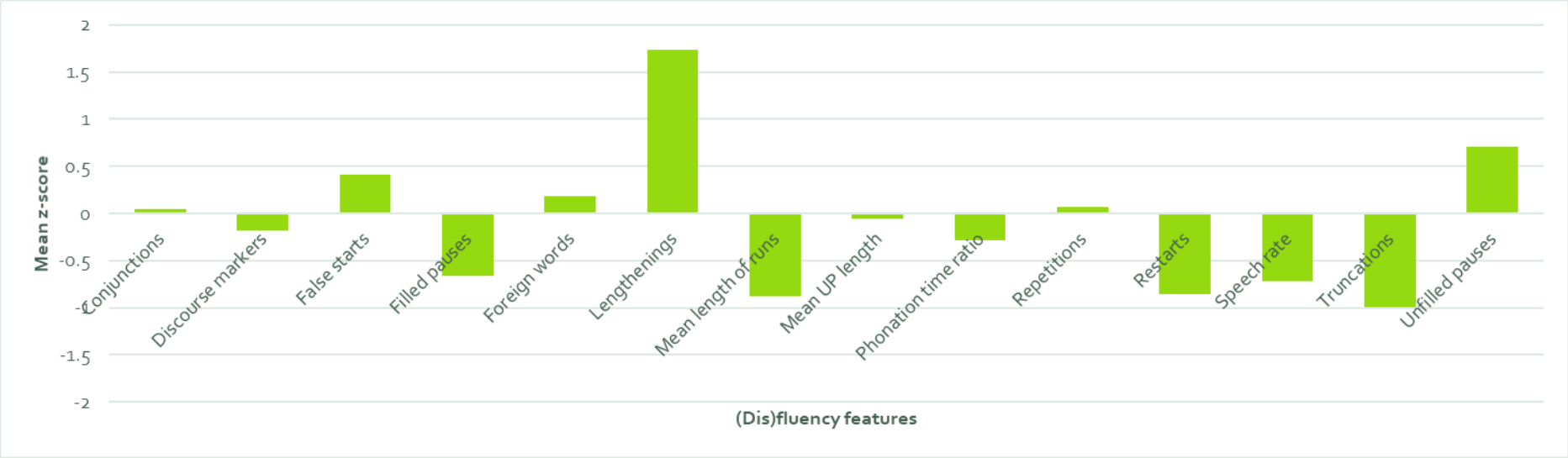


Figure 6-19: Cluster F profile in LINDSEI-FR+

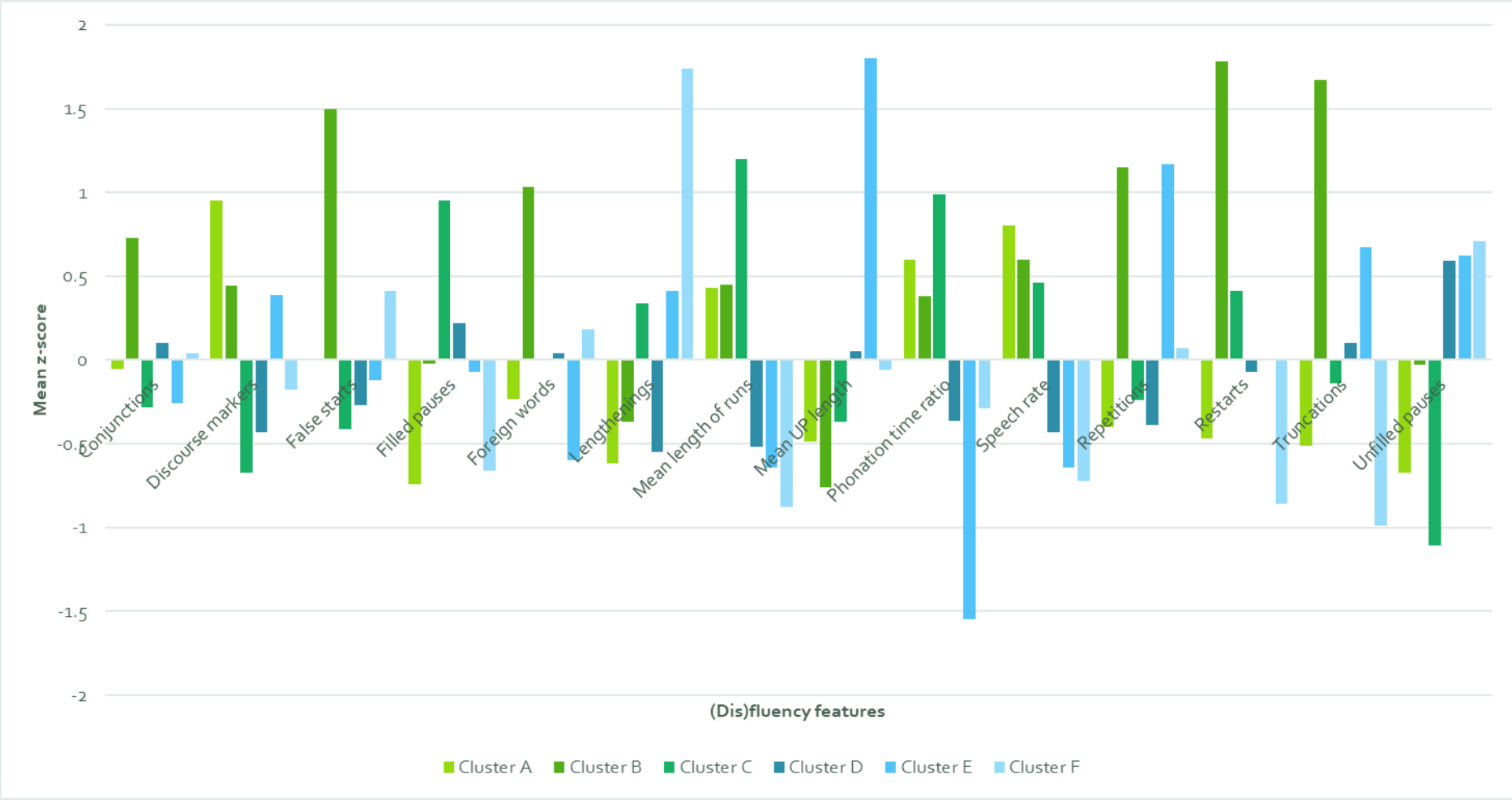


Figure 6-20: Mean z-scores per (dis)fluency variable for the 6-cluster solution in LINDSEI-FR+

(Dis)fluency components	Cluster A (n = 10)		Cluster B (n = 4)		Cluster C (*n* = 9)		Cluster D (*n* = 15)		Cluster E (*n* = 6)		Cluster F (*n* = 6)	
	Mean score	*sd*	Mean score	*sd*	Mean score	*sd*	Mean score	*sd*	Mean score	*sd*	Mean score	*sd*
Component 1	.67	.41	.43	1.06	1.16	.57	-.53	.68	-1.06	.62	-.77	.50
Component 2	-.54	.57	1.86	.61	0.2	.81	-.17	.61	.91	.75	-.85	.86
Component 3	.94	.80	.46	.78	-1.0	.54	-.46	.96	.48	.17	.30	.67
Component 4	-.00	.61	.84	1.22	-.16	.91	-.03	1.35	-.44	.65	.19	.69
Component 5	-.17	1.01	1.16	1.83	-.05	.68	.00	.94	-.65	.50	.12	.82

Table 6-8: Mean component scores and standard deviations (sd) per (dis)fluency component for the 6-cluster solution in LINDSEI-FR+

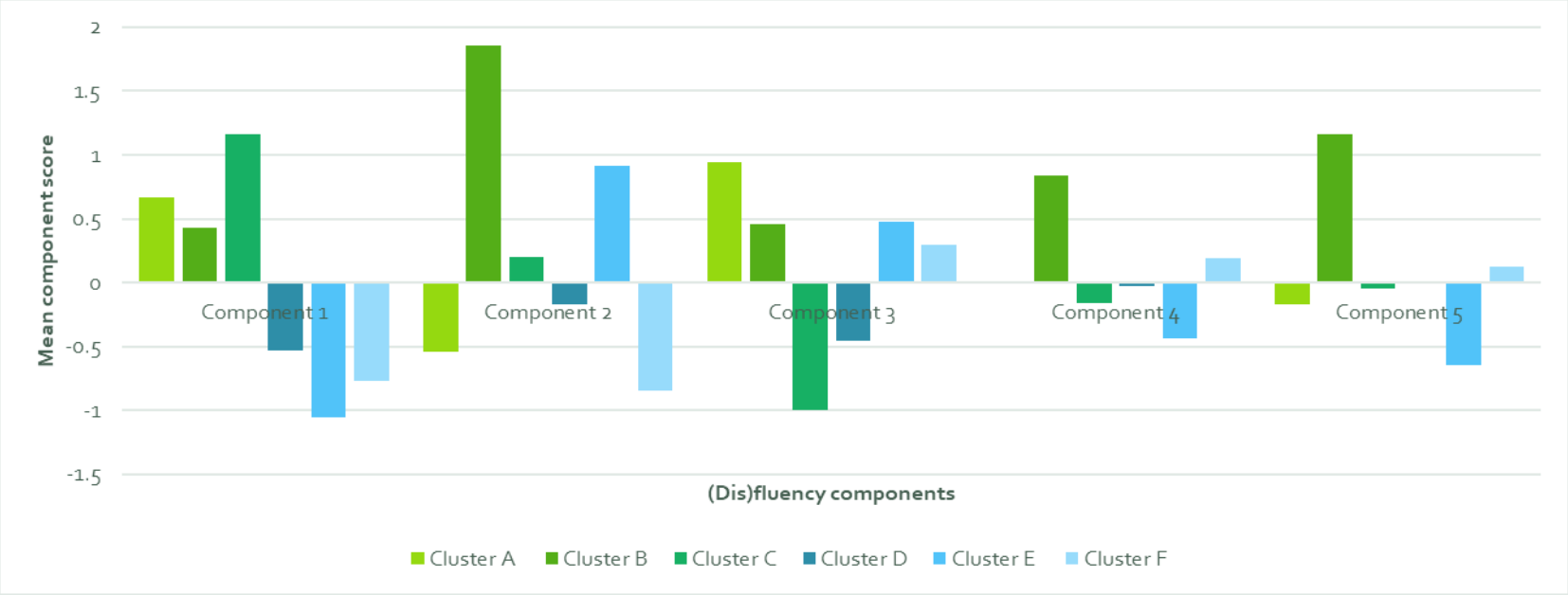


Figure 6-21: Mean component scores for the 6-cluster solution in LINDSEI-FR+

Following Jarvis *et al.* (2003; also Friginal, Li & Weigle 2014), the characterisation of each cluster (i.e. profile) is based on the level of magnitude of the mean *z*-score per (dis)fluency variable: a +/- .5 mean *z*-score is designated as cut-off, with *z*-scores above .5 ("+") and below -.5 ("-") representing noteworthy high ($Z > .5$) or low ($Z < -.5$) use of the various (dis)fluency features. On this basis, Table 6-9 offers an overview of the major characteristics of each profile in terms of individual (dis)fluency variables and Table 6-10 in terms of (dis)fluency components.

- **Profile A** is characterised by a high temporal fluency (few unfilled pauses, a high phonation-time ratio and speech rate), a high use of discourse markers, and a low use of filled pauses, lengthenings and truncations. All other (dis)fluency features in the dataset for this cluster have mean *z*-scores that fall within the -0.5 and +0.5 range.

- **Profile B** speakers have a fast speech rate and produce short unfilled pauses. They also show a marked preference for truncations, restarts and repetitions (i.e. Component 2 (dis)fluency variables), as well as conjunctions, false starts and foreign words (i.e. Components 4 and 5).

- **Profile C** is characterised by a very high temporal fluency (very few unfilled pauses, very long runs and high phonation-time ratio). It also shows a high use of filled pauses and a low use of discourse markers (i.e. the opposite of profile A, i.e. a low mean Component 3 score).

(Dis)fluency variables	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F
Conjunctions		+				
Discourse markers	+		-			
False starts		+++				
Filled pauses	-		+			-
Foreign words		++			-	
Lengthenings	-			-		+++
Mean length of runs			++	-	-	-
Mean UP length		-			+++	
Phonation-time ratio	+		+		---	
Repetitions		++			++	
Restarts		+++				-
Speech rate	+	+			-	-
Truncations	-	+++			+	-
Unfilled pauses	-		--	+	+	+

Table 6-9: Summary of the major characteristics of the 6 clusters in LINDSEI-FR+ per (dis)fluency feature
Note: '+' and '-' represent mean z-scores > ± .5; '++' and '- -' scores > ± 1.0;
'+++' and '- - -' scores > ± 1.5 (adapted from Jarvis et al. (2003))

(Dis)fluency components	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F
Component 1 Temporal (dis)fluency	+		++	-	--	-
Component 2 Repair (dis)fluency	-	+++			+	-
Component 3 Pragmatic (dis)fluency	+		--			
Component 4 Discourse cohesion		+				
Component 5 Lexico-gram. (dis)fluency		++			-	

Table 6-10: Summary of the major characteristics of the 6 clusters in LINDSEI-FR+ per (dis)fluency component
Note: '+' and '-' represent mean component scores > ± .5; '++' and '- -' scores > ± 1.0;
'+++' scores > 1.5

- **Profile D** speakers pause a lot and produce short speech runs. They use few lengthenings, but have no marked preference or dispreference for any other (dis)fluency feature.

- **Profile E** is characterised by a particularly low temporal fluency: profile E speakers produce many and long unfilled pauses, speak in short runs, have a slow speech rate and a very low phonation-time ratio. This profile is also characterised by a high use of truncations and a very high use of repetitions (two variables included in repair fluency), as well as few foreign words.

- **Profile F** is characterised by a high use of unfilled pauses, short runs and a slow speech rate (i.e. a low temporal fluency). Profile F speakers use few truncations, restarts and filled pauses (i.e. repair fluency variables), but have a marked preference for lengthenings.

Although there is some overlap in terms of temporal variables between the sub-clusters of Cluster 1 (Clusters A to C) and of Cluster 2 (Cluster D to F), it is quite striking that **each cluster exhibits a unique pattern** across the (dis)fluency variables (and components). The uniqueness of the patterns is further confirmed by a series of **one-way ANOVAs**, with cluster membership as fixed factor and the 14 (dis)fluency measures as dependent variables. Since the assumption of homogeneity of variance was not met for foreign words and truncations, Welch's adjusted *F*-ratio was calculated instead for those two variables (Field 2013:443; Howell 2013:343). As revealed by Table 6-11, statistically **significant differences were found for 13 variables** out of 14. In other words, the six clusters differ significantly on their average rate of discourse markers, filled pauses, false starts, foreign words, lengthenings, mean length of runs and of unfilled pauses, phonation-time ratio, repetitions, restarts, speech rate, truncations, and unfilled pauses (i.e. all variables but conjunctions).

Post hoc tests were carried out to compare the six groups with each other using Gabriel's procedure, which is advised for unbalanced designs (i.e. when group sizes are different)[146] or with Games-Howell (for truncations and foreign words, as the assumption of homogeneity of variance was not met for those variables) (Field 2013:458–460; Howell 2013:343). Table 6-12 below provides a summary of the significant differences between groups (the detailed results of the ANOVA and of the post hoc tests are included in Appendix 9.8, Table 9-12 to Table 9-16). A quick inspection of Table 6-12 shows that, as could be expected, **Clusters A to C** (i.e. the sub-clusters of Cluster 1) generally **widely differ from Clusters D, E and F** (the sub-clusters of Cluster 2 in the 2-cluster solution). The differences, however, go beyond the temporal measures and also pertain to the rate of restarts, repetitions or foreign words for example. Clusters A, B and C (the sub-clusters of Cluster 1) differ by the rate of **truncations, repetitions, restarts, discourse markers, filled pauses, lengthenings and false starts** but are homogeneous with respect to the other (dis)fluency variables. The three sub-clusters of Cluster 2 (i.e. Clusters D to F) can be differentiated by temporal and non-temporal variables: the number of **lengthenings**, the **length of unfilled pauses**, the **phonation-time ratio**, the **rate of truncations and of repetitions** all significantly differ between those three clusters.

(Dis)fluency variables	***Levene's test***	***F***	**Sign.**
Conjunctions	$F = 2.12$; $p = .081$	.66	$p = .658$ (n.s.)
Discourse markers	$F = 1.42$; $p = .235$	4.93	***p* =. 001**
False starts	$F = .279$; $p = .922$	3.06	***p* = .019**
Filled pauses	$F = .351$; $p = .879$	4.66	***p* = .002**
Foreign words	$F = 4.82$; ***p* = .001**	3.65*	***p* = .024**
Lengthenings	$F = 1.57$; $p = .188$	12.91	***p* = .000**
Mean length of runs	$F = 1.96$; $p = .104$	10.63	***p* = .000**
Mean UP length	$F = 1.651$; $p = .167$	9.60	***p* = .000**
Phonation-time ratio	$F = 1.46$; $p = .223$	13.57	***p* = .000**
Repetitions	$F = 1.43$; $p = .234$	5.08	***p* = .001**
Restarts	$F = 1.74$; $p = .146$	6.53	***p* = .000**
Speech rate	$F = 2.36$; $p = .056$	5.12	***p* = .001**
Truncations	$F = 3.20$; ***p* = .015**	21.17*	***p* = .000**
Unfilled pauses	$F = 1.24$; $p = .309$	10.14	***p* = .000**

Table 6-11: Results of one-way ANOVA per (dis)fluency variable
*Note: * Welch's adjusted F-ratio; significant results are in bold font*

[146] For all (dis)fluency variables except truncations and foreign words, I compared the results from Gabriel's procedure with Hochberg's GT2, which is also advised when group sizes are different, as well as with the Games-Howell post-hoc test (Field 2013:458–460; Howell 2013:343). Gabriel and Hochberg provided perfectly identical results, and Games-Howell was overall very similar too.

	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F
Cluster A	REP RS T*	DM FP L	DM MLR PTR SR T* UP	MLUP PTR REP SR UP	L MLR SR UP
Cluster B		FS T*	FS REP RS T*	MLUP PTR RS	L RS T*
Cluster C			L MLR PTR UP	MLR PTR REP UP W*	FP L MLR PTR UP
Cluster D				MLUP PTR REP	L
Cluster E					L PTR T*

Table 6-12: Summary of post hoc tests
Pairwise comparisons with Gabriel's procedure (Games-Howell)* [147]

Further **one-way ANOVAs** with cluster membership as fixed factor and the 5 (dis)fluency components as dependent variables (see Table 6-13) revealed that the 6 learner clusters differ significantly along the first three components, namely the temporal, repair and pragmatic components. A summary of the results from the pairwise comparisons with Gabriel's procedure is provided in Table 6-14: these results generally confirm those presented per (dis)fluency variable in Table 6-12.

(Dis)fluency components	***Levene's test***	***F***	**Sign.**
Component 1: Temporal (dis)fluency	$F = 1.30$; $p = .282$	16.520	***p* = .000**
Component 2: Repair (dis)fluency	$F = .50$; $p = .772$	10.391	***p* = .000**
Component 3: Pragmatic (dis)fluency	$F = 1.88$; $p = .117$	8.262	***p* = .000**
Component 4: Discourse cohesion	$F = 1.26$; $p = .298$	.871	$p = .508$
Component 5: Lexico-grammatical (dis)fluency	$F = 1.84$; $p = .126$	1.824	$p = .128$

Table 6-13: Results of one-way ANOVA per (dis)fluency component

[147] As a reminder (in alphabetical order): DM = discourse markers; FP = filled pauses; L = lengthenings; MLR = mean length of runs; MLUP = mean length of unfilled pauses; PTR = photation-time ratio; Rep = repetitions; RS = restarts; SR = speech rate; T = truncations; UP = unfilled pauses; W = foreign words.

*Note: * Welch's adjusted F-ratio; significant results are in bold font*

	Cluster B	**Cluster C**	**Cluster D**	**Cluster E**	**Cluster F**
Cluster A	Comp. 2	Comp. 3	Comp. 1 Comp. 3	Comp. 1 Comp. 2	Comp. 1
Cluster B		Comp. 2 Comp. 3	Comp. 2	Comp. 1	Comp. 2
Cluster C			Comp. 1	Comp. 1 Comp. 3	Comp. 1 Comp. 3
Cluster D				Comp. 2	/
Cluster E					Comp. 2

Table 6-14: Summary of post hoc test with Gabriel's procedure
Pairwise comparisons with Gabriel's procedure

To sum up, the Cluster Analysis indicates that there are, indeed, multiple (dis)fluency profiles even among (what was thought to be) a homogeneous group of learners. Two main groups can be distinguished based on temporal (dis)fluency variables, but even among these two groups, smaller clusters of speakers could be identified who all have a distinctive pattern across (dis)fluency variables (and components).

To see if native speakers also differ from one another along the same variables, a second Cluster Analysis was run on the LOCNEC+ data.

6.2.2 Native speakers' (dis)fluency profiles

As was the case for the clustering of LINDSEI-FR+ learners, the Cluster Analysis of the native speakers of LOCNEC+ was carried out using the 14 (dis)fluency variables and the same clustering algorithm (i.e. Ward's). A pre-test showed that there were four outliers (EN004, EN038, EN047, and EN049) who only clustered very late with the other speakers and they were consequently excluded from the final analysis. The dendrogram illustrating the procedure is displayed in Figure 6-22. The examination of the fusion coefficients indicates that there could be 2 or 5 clusters. Likewise, the dendrogram also points to a 2- or 5-cluster solution. The 2-cluster solution is examined first.

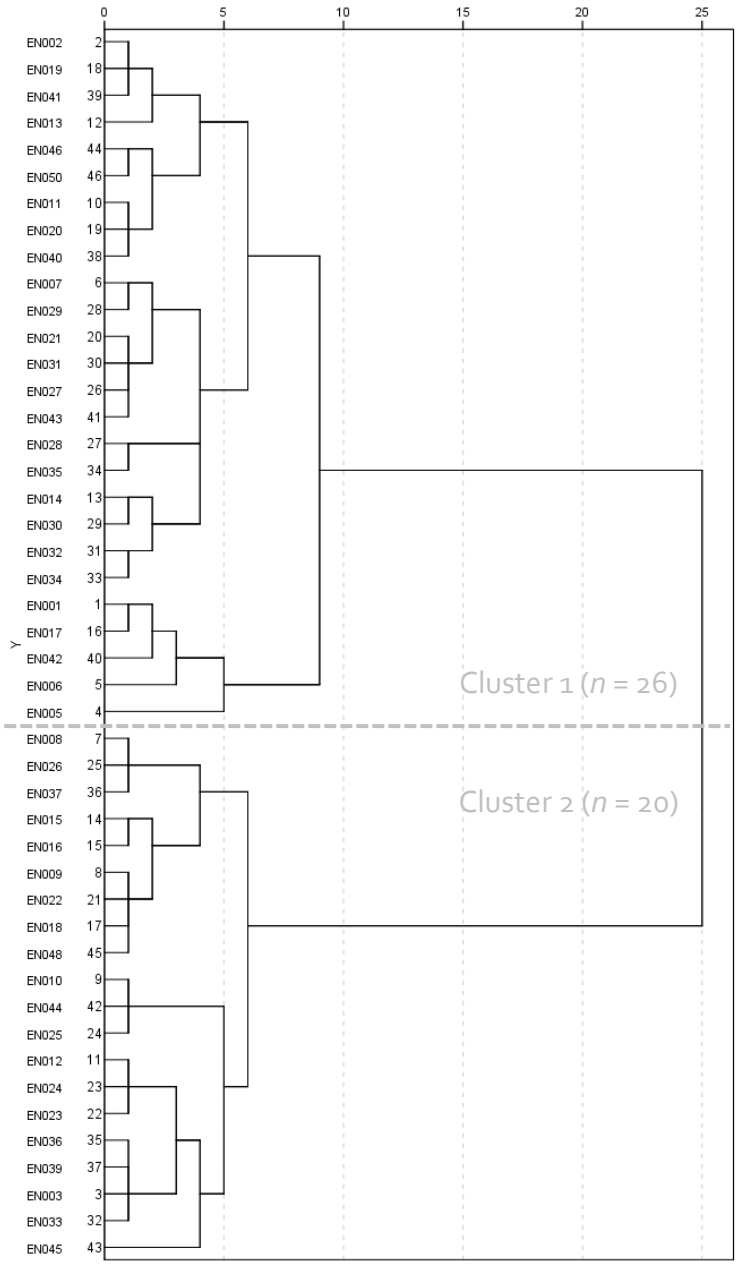


Figure 6-22: Dendrogram obtained from Hierarchical Cluster Analysis: Speaker performances across (dis)fluency variables in LOCNEC+ (Ward's method, Squared Euclidean Distance)

6.2.2.1 The 2-cluster solution

The two main clusters in LOCNEC+ include 26 and 20 speakers, respectively. The mean *z*-scores for each (dis)fluency variable, which are presented in Table 6-15, are represented graphically in Figure 6-23.

(Dis)fluency variables	Cluster 1 (n = 26)		Cluster 2 (n = 20)		t-test	Cohen's d
	Mean z-score	sd	Mean z-score	sd		
Conjunctions	-0.14	0.55	0.36	1.12	t = -1.99; p = .053	
Discourse markers	-0.20	0.82	0.38	1.08	**t = -2.08; p = .044**	d = 0.61
False starts	-0.19	1.05	0.29	0.75	t = -1.74; p = .088	
Filled pauses	0.37	1.15	-0.44	0.57	**t = 2.89; p = .006**	d = 0.89
Foreign words	0.03	0.94	-0.22	0.30	t = 1.25; p = .221	
Lengthenings	-0.05	0.93	0.03	0.97	t = -0.27; p = .786	
Mean length of runs	-0.58	0.46	0.78	1.08	**t = -5.26; p = .000**	d = 1.63
Mean UP length	0.48	0.96	-0.59	0.60	**t = 4.36; p = .000**	d = 1.33
Phonation-time ratio	-0.65	0.75	0.92	0.47	**t = -8.66; p = .000**	d = 2.50
Repetitions	0.09	1.12	-0.16	0.73	t = 0.88; p = .382	
Restarts	-0.19	0.76	0.04	0.77	t = -1.03; p = .310	
Speech rate	-0.53	0.69	0.71	0.91	**t = -5.28; p = .000**	d = 1.54
Truncations	-0.20	0.60	-0.10	0.60	t = -0.55; p = .587	
Unfilled pauses	0.61	0.83	-0.91	0.40	**t = 7.50; p = .000**	d = 2.32

Table 6-15: Mean z-scores and standard deviations (sd) per (dis)fluency variable and independent samples t-tests results for the 2-cluster solution in LOCNEC+

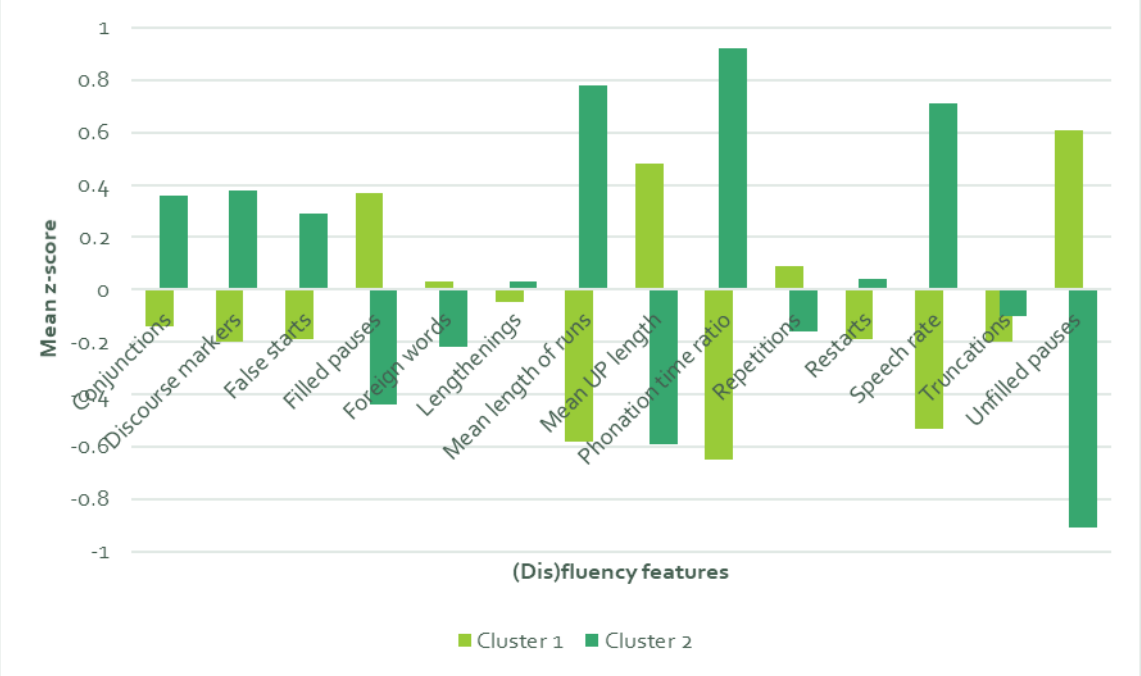


Figure 6-23: Mean z-scores per (dis)fluency variable for the 2 clusters in LOCNEC+

As was the case for the LINDSEI-FR+ 2-cluster solution, the two clusters tend to function as mirror images, with, for each variable, one cluster having a positive mean z-score while the other has a negative mean. **Cluster 1** (n = 26) is characterised by positive mean z-scores on 5

variables (filled pauses, foreign words, mean length of unfilled pauses, repetitions, and unfilled pauses), and negative means for the other variables (i.e. conjunctions, discourse markers, false starts, lengthenings, mean length of runs, phonation-time ratio, restarts, speech rate and truncations). The opposite is true for **Cluster 2** (*n* = 20), except for truncations, the mean of which is also negative. In other words, on average, the 26 speakers of Cluster 1 tend to produce more filled and unfilled pauses, as well as more foreign words and repetitions than the 20 speakers of Cluster 2, who tend to produce more conjunctions, discourse markers, false starts and to have a higher speech rate than Cluster 1.

Independent samples *t*-tests revealed that there was a **statistically significant difference** between the two clusters for **seven (dis)fluency variables**, namely discourse markers, filled pauses, mean length of runs, mean length of unfilled pauses, phonation-time ratio, speech rate and unfilled pauses (see bold figures in Table 6-15) and the effect sizes of those differences is large to very large ($0.61 \leq d \leq 2.50$; Cohen (1977)). These results imply that Cluster 2 speakers produce significantly fewer and shorter unfilled pauses than Cluster 1 speakers; their speech rate is thus faster, their phonation-time ratio higher and the length of their speech runs longer. The 20 speakers in Cluster 2 also produce more discourse markers and fewer filled pauses than the speakers in Cluster 1.

(Dis)fluency components	Cluster 1 (*n* = 26)		Cluster 2 (*n* = 20)		*t*-test	Cohen's *d*
	Mean component score	*sd*	Mean component score	*sd*		
Component 1	-.68	.65	.91	.65	**$t = -8.23$; $p = .000$**	$d = 2.46$
Component 2	-.18	.62	-.04	.84	$t = -.67$; $p = .507$	
Component 3	-.17	1.03	.24	.98	$t = -1.36$; $p = .182$	
Component 4	-.09	.75	.34	.93	$t = -1.74$; $p = .089$	

Table 6-16: Mean score and standard deviation (sd) per (dis)fluency component and independent samples t-test for the 2-cluster solution in LOCNEC+

As was the case for the 2-cluster solution in LINDSEI-FR+, it is thus the **temporal (dis)fluency variables** (i.e. 6 of the 7 significant variables) that distinguish the two main clusters in LOCNEC+ – and it logically follows that there is also a significant difference between the two clusters with respect to Component 1, i.e. the temporal (dis)fluency component (see Table 6-16). What is more surprising, however, is that the rate of **discourse markers** also significantly differs between the two groups of native speakers. In the Principal Components Analysis, discourse markers were found to exclusively load on Component 3. Although filled pauses primarily load on the temporal dimension of (dis)fluency in the native corpus (i.e. Component 1), they also slightly load on Component 3: this association may perhaps be a first clue to explain the significance of discourse markers in the 2-cluster solution. Note, however, that Component 3 did not turn out to be significant in the 2-cluster solution. In any case, this finding also highlights the fact that native speakers greatly differ from each other with

respect to the frequency of discourse markers, which has not been underlined very frequently in L1-L2 contrastive analyses.

6.2.2.2 The 5-cluster solution

To get better insights into finer-grained usage patterns of native speakers, the 5-cluster solution was also investigated.

Cluster 1 from the 2-cluster solution is further subdivided into three sub-clusters: Cluster A (*n* = 5), Cluster B (*n* = 9) and Cluster C (*n* = 12). Cluster 2 is made up of two sub-clusters of about equal size: Cluster D (*n* = 11) and Cluster E (*n* = 9). Table 6-17 below displays the mean z-scores for each (dis)fluency variable in each of the five clusters (see also a visual representation of each profile in Figure 6-24 to Figure 6-29). The mean component scores per cluster are presented in Table 6-18 (see also Figure 6-30) and illustrated in Appendix 9.8, Figure 9-11 to Figure 9-15.

(Dis)fluency variables	Cluster A (*n* = 5)		Cluster B (*n* = 9)		Cluster C (*n* = 12)		Cluster D (*n* = 11)		Cluster E (*n* = 9)	
	Mean *z*-score	*sd*	Mean *z*-score	*sd*	Mean *z*-score	*sd*	Mean *z*-score	*sd*	Mean *z*-score	*sd*
Conjunctions	-0.27	0.20	0.27	0.56	-0.39	0.47	0.55	1.35	0.12	0.75
Discourse markers	-1.12	0.44	0.29	0.81	-0.19	0.64	0.32	1.36	0.45	0.68
False starts	-0.44	1.37	0.54	0.93	-0.64	0.71	0.53	0.78	0.00	0.64
Filled pauses	2.13	1.29	0.24	0.61	-0.27	0.52	-0.49	0.48	-0.38	0.70
Foreign words	-0.43	0.00	-0.15	0.56	0.35	1.25	-0.22	0.30	-0.21	0.31
Lengthenings	0.29	1.24	-0.36	0.68	0.04	0.97	-0.37	0.66	0.51	1.9
Mean length of runs	-0.84	0.40	-0.71	0.44	-0.37	0.44	0.17	0.69	1.52	1.02
Mean UP length	0.58	1.24	0.10	0.76	0.72	0.96	-0.18	0.42	-1.09	0.36
Phonation-time ratio	-0.82	0.69	-0.95	0.92	-0.35	0.55	0.82	0.46	1.03	0.47
Repetitions	1.62	1.61	-0.22	0.58	-0.31	0.61	-0.51	0.37	0.27	0.84
Restarts	-0.25	0.86	-0.31	0.91	-0.08	0.65	-.018	0.81	0.32	0.66
Speech rate	-1.07	0.78	-0.63	0.52	-0.24	0.65	0.20	0.82	1.33	0.60
Truncations	-0.03	0.49	-0.24	0.45	-0.24	0.74	-0.24	0.71	0.06	0.40
Unfilled pauses	0.98	0.62	1.18	0.75	0.03	0.57	-0.90	0.22	-0.91	0.56

Table 6-17: Mean z-scores per (dis)fluency variable for the 5-cluster solution in LOCNEC+

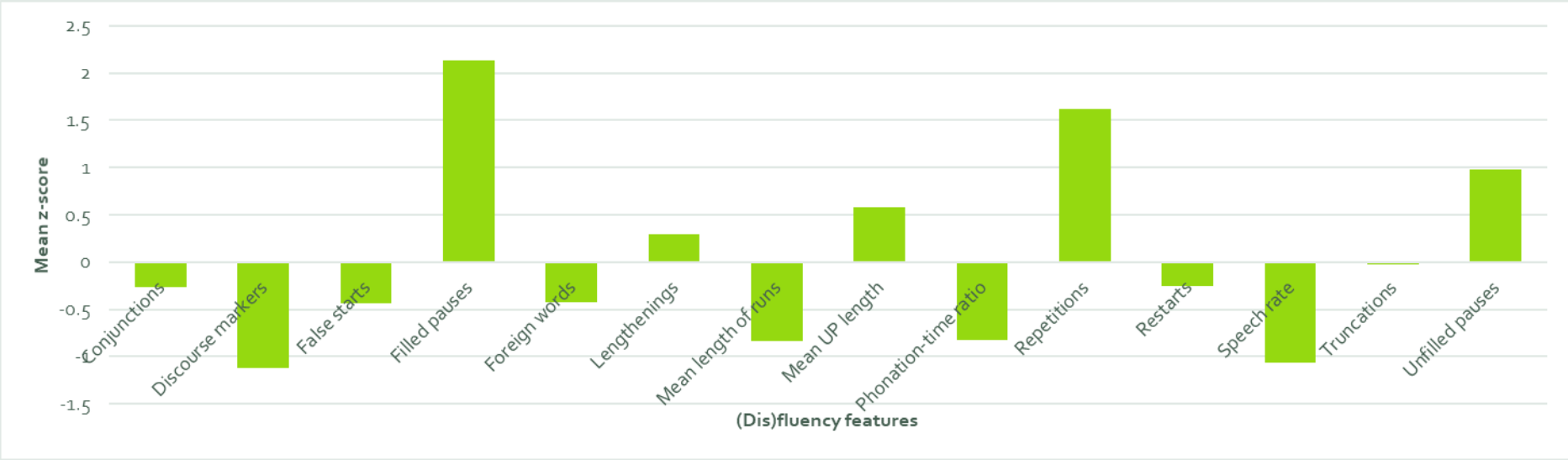


Figure 6-24: Cluster A profile (LOCNEC+)

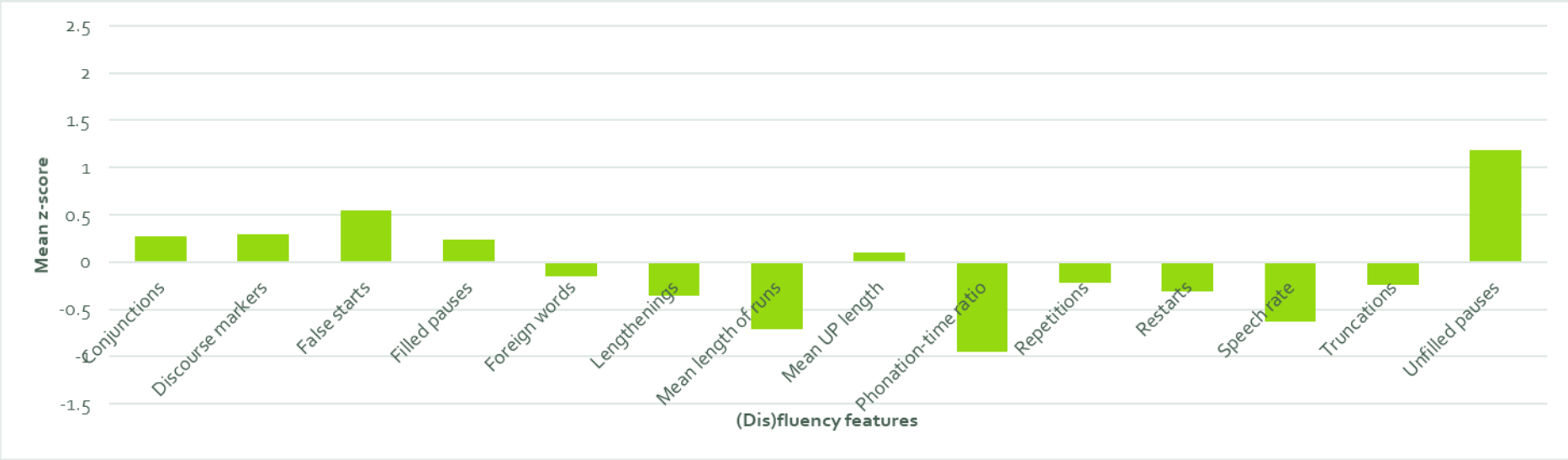


Figure 6-25: Cluster B profile (LOCNEC+)

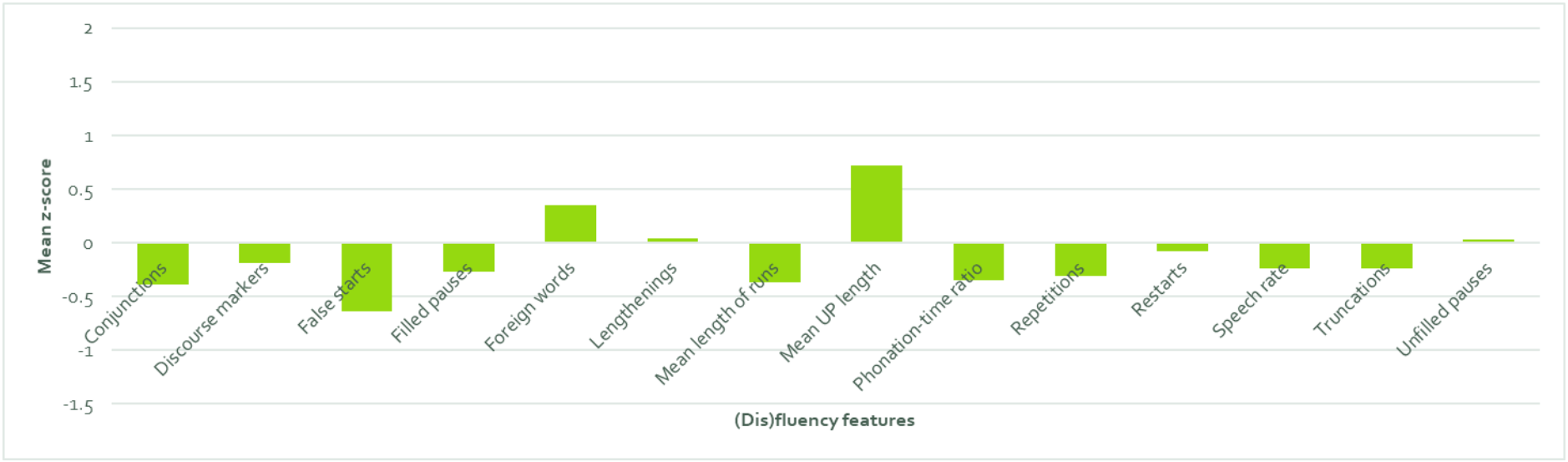


Figure 6-26: Cluster C profile (LOCNEC+)

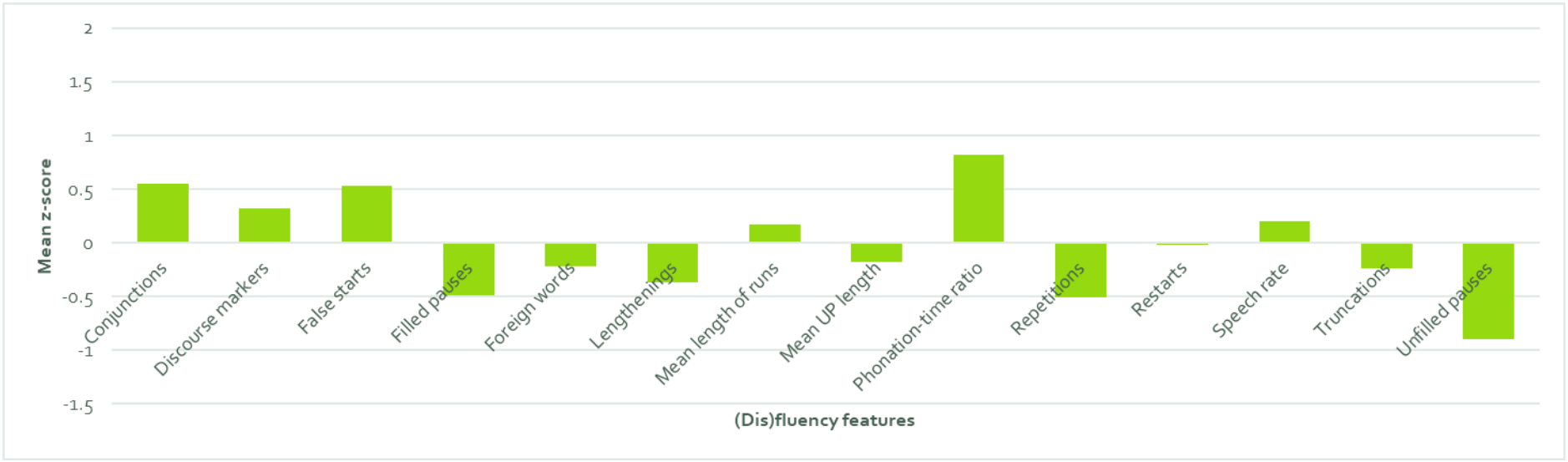


Figure 6-27: Cluster D profile (LOCNEC+)

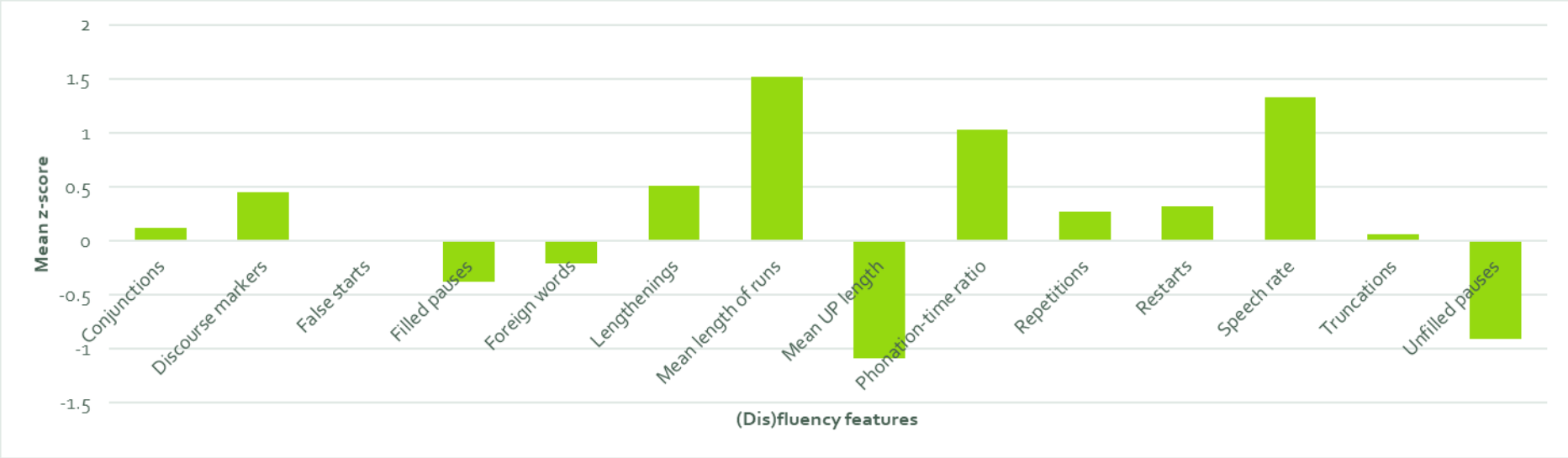


Figure 6-28: Cluster E profile (LOCNEC+)

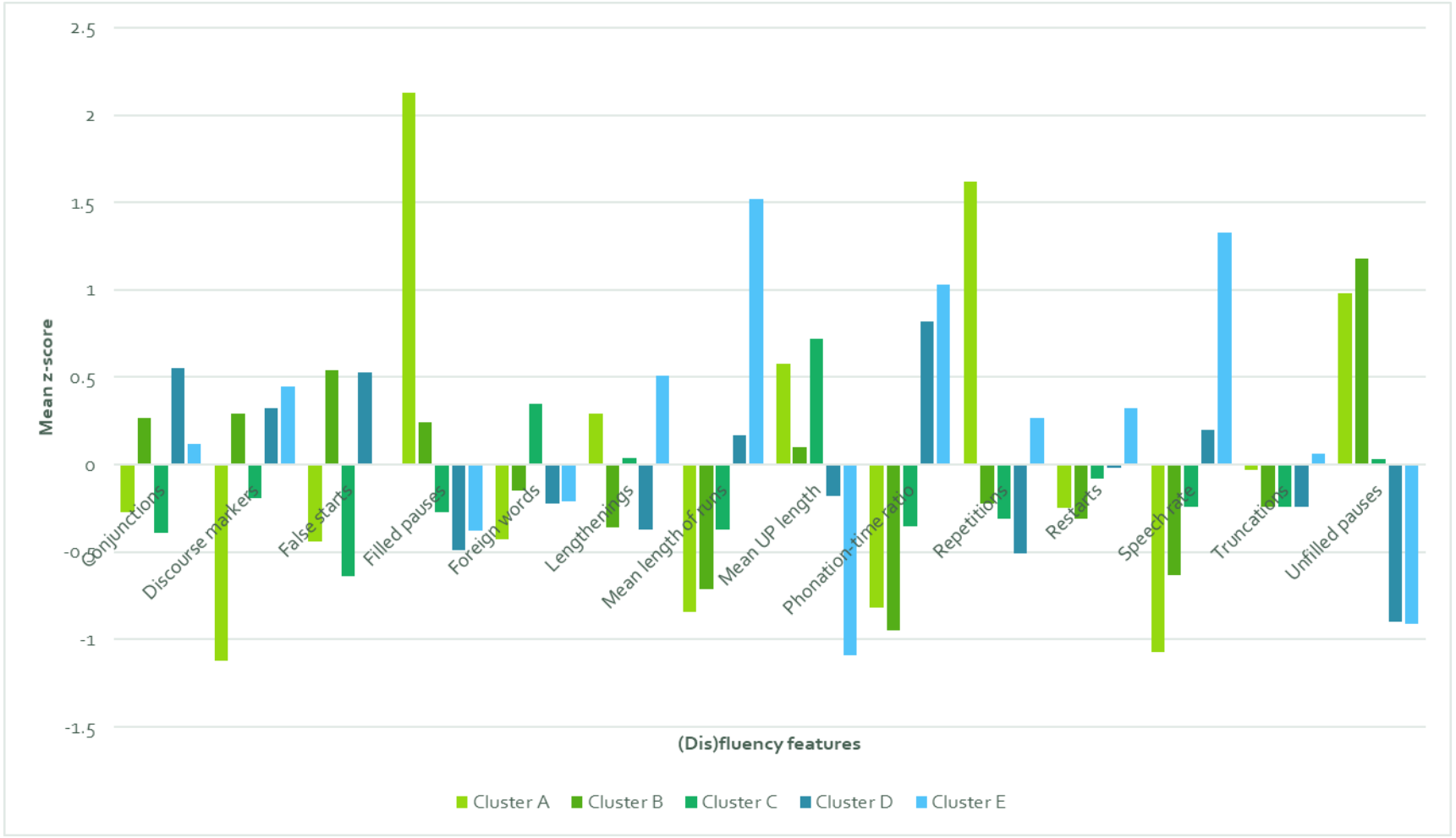


Figure 6-29: Mean z-scores per (dis)fluency variable for the 5-cluster solution in LOCNEC+

(Dis)fluency components	Cluster A (*n* = 5)		Cluster B (*n* = 9)		Cluster C (*n* = 12)		Cluster D (*n* = 11)		Cluster E (*n* = 9)	
	Mean component score	*sd*	Mean component score	*sd*	Mean component score	*sd*	Mean component score	*sd*	Mean component score	*sd*
Component 1 Temporal (dis)fluency	-1.18	.57	-1.08	.48	-.18	.39	.53	.39	1.38	.61
Component 2 Repair (dis)fluency	.25	.37	-.40	.59	-.19	.66	-.46	.75	.47	.64
Component 3 Pragmatic (dis)fluency	-1.59	.37	.64	.78	-.18	.66	.55	1.15	-.14	.60
Component 4 Discourse cohesion	.23	1.13	.33	.49	-.54	.49	.49	1.14	.16	.60

Table 6-18: Mean component scores and standard deviations (sd) per (dis)fluency component for the 5-cluster solution in LOCNEC+

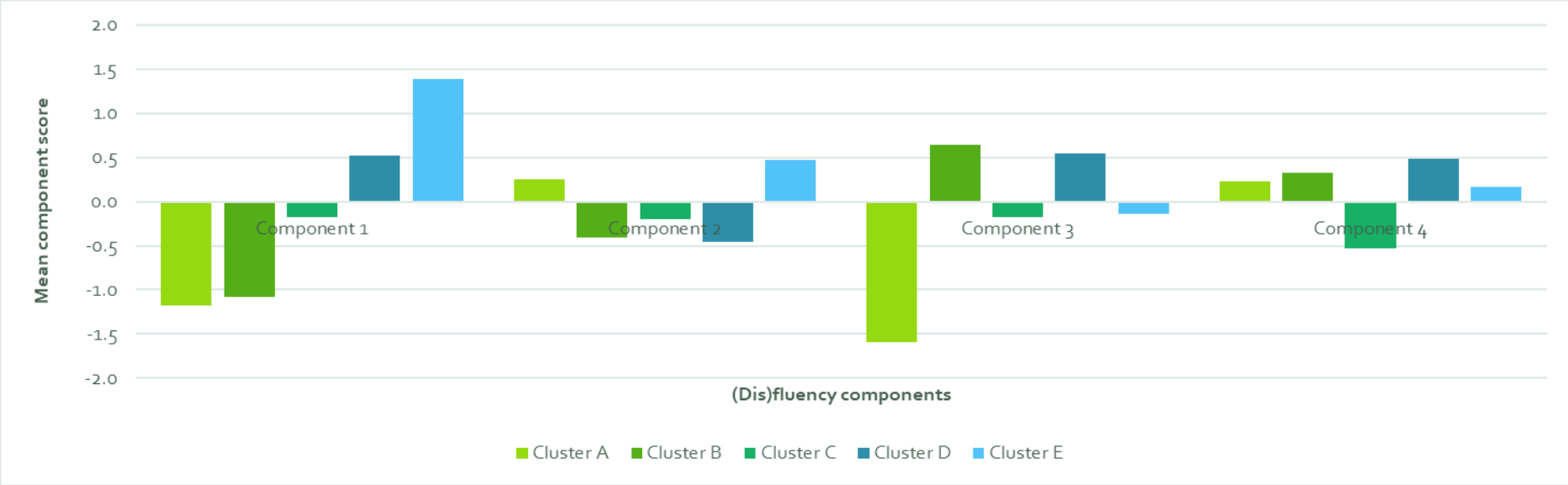


Figure 6-30: Mean component scores for the 5-cluster solution in LOCNEC+

As was the case for the learner profiles, the characterisation of each native profile is based on the level of magnitude of the mean *z*-score per (dis)fluency variable (i.e. a +/- .5 mean *z*-score as cut-off, *cf*. Jarvis *et al.* (2003)). Table 6-19 summarises the major characteristics of each of the five native (dis)fluency profiles in terms of separate disfluency variables and Table 6-20 is a synthesis per (dis)fluency component.

- **Profile A** is characterised by a low temporal fluency: unfilled pauses occur frequently and are quite long, speech runs are short, the phonation-time ratio is low, and the overall speech rate is very slow. Filled pauses also occur particularly frequently. Two other distinctive characteristics of the discourse of Profile A speakers are the very high rate of repetitions, and the very low frequency of discourse markers (i.e. a low pragmatic fluency).

- The major characteristic of **Profile B** speakers is their very frequent use of unfilled pauses. As a consequence, the mean length of speech runs and the phonation-time ratio are low, and the speech rate is slow. False starts are also typical of this profile.

- **Profile C** is characterised by long unfilled pauses and the rarity of false starts. The mean *z*-score of all other variables falls within the -.5 and +.5 band.

(Dis)fluency variables	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
Conjunctions				+	
Discourse markers	--				
False starts		+	-	+	
Filled pauses	++++				
Foreign words					
Lengthenings					+
Mean length of runs	-	-			+++
Mean UP length	+		+		--
Phonation time ratio	-	-		+	++
Repetitions	+++			-	
Restarts					
Speech rate	--	-			++
Truncations					
Unfilled pauses	+	++		-	-

Table 6-19: Summary of the major characteristics of the 5 clusters in LOCNEC+ per (dis)fluency variable
Note: '+' and '-' represent mean z-scores > ± .5; '++' and '- -' scores > ± 1.0;
'+++' and '- - -' scores > ± 1.5 (adapted from Jarvis et al. (2003))

(Dis)fluency components	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
Component 1 : Temporal (dis)fluency	--	--		+	++
Component 2 : Repair (dis)fluency					
Component 3 : Pragmatic (dis)fluency	---	+		+	
Component 4: Discourse cohesion			-		

Table 6-20: Summary of the major characteristics of the 5 clusters in LOCNEC+ per (dis)fluency component
Note: '+' and '-' represent mean component scores > ± .5; '++' and '- -' scores > ± 1.0;
'- - -' scores > - 1.5

- A low frequency of unfilled pauses and repetitions, a high phonation-time ratio, as well as a high frequency of conjunctions and false starts are defining characteristics of **Profile D** speakers.

- **Profile E** speakers have the highest temporal fluency: unfilled pauses are not only rare, but they are also very short, the phonation-time ratio and speech rate are thus very high. However, Profile E speakers also produce a lot of lengthenings.

Again, each cluster seems to exhibit a unique pattern across the (dis)fluency variables (and components), which was also confirmed by one-way ANOVAs with cluster membership as fixed factor and the (dis)fluency measures as dependent variables (Table 6-21). The omnibus statistic revealed **significant differences for 10 variables** out of 14 (i.e. all but conjunctions, lengthenings, restarts and truncations), which means that the five clusters significantly differ with respect to the mean of the following variables: discourse markers, false starts, filled pauses, foreign words, mean length of runs, mean UP length, phonation-time ratio, repetitions, speech rate and unfilled pauses. The examination of mean differences between clusters in terms of (dis)fluency components (Table 6-23) highlights that all four components significantly differ between the 5 native clusters.

A summary of the results of the **post hoc tests** to compare the five groups with each other using the Gabriel's[148] (or Games-Howell's) procedure (Field 2013:458–460; Howell 2013) is provided in Table 6-22 – the detailed results are included in Appendix 9.8. The inspection of Table 6-22 indicates that Clusters A, B and C (i.e. the sub-clusters of Cluster 1) differ from one another with respect to false starts and unfilled pauses, whereas the sub-clusters of Cluster 2 (D and E) differ with respect to three temporal variables: mean length of runs, mean length of unfilled pauses and speech rate. Clusters A, B and C and Clusters D and E differ not only in terms of temporal variables (including filled pauses), but also by the rate of discourse markers and false starts. Note also that, although the omnibus ANOVA was significant for **repetitions** and **foreign words**, the post hoc tests did not reveal significant between-group differences,

[148] I compared the results from Gabriel's procedure with Hochberg GT2 and the results were identical.

thereby suggesting that the size of the differences is very small and that the 5 clusters only *tend to differ* with respect to those (dis)fluency variables.

All in all, the 5 native clusters can thus clearly be distinguished by 8 variables: the 6 temporal variables, discourse markers and false starts. Interestingly, these variables perfectly match **Component 1 and Component 3** (*cf.* Section 6.1.2) – but the two other components also significantly differenciate the clusters (see Table 6-23 and Table 6-24).

(Dis)fluency variables	Levene's test	*F*	Sign.
Conjunctions	$F = 1.97$; $p = .117$	*2.23*	$p = .082$ (n.s.)
Discourse markers	$F = 2.05$; $p = .105$	*3.23*	***p* = .021**
False starts	$F = 1.93$; $p = .123$	*4.03*	***p* = .008**
Filled pauses	$F = 3.08$; ***p* = .026**	*5.75**	***p* = .005**
Foreign words	$F = 11.44$; ***p* = .000**	*3.77**	***p* = .021**
Lengthenings	$F = 1.31$; $p = .284$	*1.64*	$p = .182$ (n.s.)
Mean length of runs	$F = 1.50$; $p = .222$	*18.35*	***p* = .000**
Mean UP length	$F = 3.04$; ***p* = .028**	*12.90**	***p* = .000**
Phonation time ratio	$F = 1.96$; $p = .119$	*19.24*	***p* = .000**
Repetitions	$F = 3.43$; ***p* = .017**	*3.32**	***p* = .038**
Restarts	$F = .45$; $p = .773$	*.918*	$p = .463$ (n.s.)
Speech rate	$F = 1.35$; $p = .268$	*14.37*	***p* = .000**
Truncations	$F = 2.80$; ***p* = .038**	*.73**	$p = .582$ (n.s.)
Unfilled pauses	$F = 1.77$; $p = .153$	*27.36*	***p* = .000**

*Table 6-21: Results of one-way ANOVA (*Welch's F)*

	Cluster B	Cluster C	Cluster D	Cluster E
Cluster A		UP	DM FP* PTR SR UP	DM FP* MLR PTR SR UP
Cluster B		FS UP	MLR PTR UP	MLR MLUP* PTR SR UP
Cluster C			FS PTR UP	MLR MLUP* PTR SR UP
Cluster D				MLR MLUP* SR

Table 6-22: Summary of post hoc tests
*Pairwise comparisons with Gabriel's procedure (*Games-Howell)*

(Dis)fluency variables	Levene's test	*F*	Sign.
Component 1: Temporal (dis)fluency	$F = .561$; $p = .692$	42.07	**$p = .000$**
Component 2: Repair (dis)fluency	$F = 1.02$; $p = .409$	3.53	**$p = .015$**
Component 3: Pragmatic (dis)fluency	$F = 1.70$; $p = .168$	7.88	**$p = .000$**
Component 4: Discourse cohesion	$F = 1.64$; $p = .182$	2.83	**$p = .037$**

Table 6-23: Results of one-way ANOVA

	Cluster B	Cluster C	Cluster D	Cluster E
Cluster A	Component 3	Component 1 Component 3	Component 1 Component 3	Component 1 Component 3
Cluster B		Component 1	Component 1	Component 1
Cluster C			Component 1 Component 4	Component 1
Cluster D				Component 1 Component 2

Table 6-24: Summary of post hoc tests
*Pairwise comparison with Gabriel's procedure (*Games-Howell)*

To sum up, the Cluster Analysis revealed that there are two main groups in the native speaker data, which can be distinguished based on temporal (dis)fluency variables as well as the rate of discourse markers. The same variables, together with false starts, delineate more fine-grained clusters of speakers. All in all, the results of the analysis provide empirical support to Lennon's (1990:392) comment that "the idea of monolithic and unitary fluency for native speakers is mythical" and that, indeed, "[n]ative speakers clearly differ among themselves in fluency".

6.2.3 Discussion

The Cluster Analyses were set out with the aim of investigating whether multiple clusters of speakers (or (dis)fluency profiles) would emerge among two supposedly homogeneous groups of learners and of native speakers.

In each group, the **two-cluster solution** was examined first. In **LINDSEI-FR+, 5 (dis)fluency variables**, namely unfilled pauses, speech rate, phonation-time ratio, mean length of runs and mean length of unfilled pauses, distinguished the two clusters. Incidentally, these five variables correspond to the learner Component 1 (i.e. temporal (dis)fluency), thereby also corroborating the findings from the Principal Components Analysis presented in Section 6.1.1. In **LOCNEC+**, the two main (dis)fluency profiles were found to differ along the **6 temporal (dis)fluency variables** that made up the native Component 1 (i.e. filled and unfilled pauses, mean length of runs, mean length of unfilled pauses, phonation-time ratio and

speech rate), but also with respect to **discourse markers**. This latter finding sheds new light on the use of discourse markers in L1 speech and opens the way to further investigations into the variability of discourse markers in the speech of native speakers (and contrastive L1-L2 studies).

In her investigation of **German-speaking learners of English**, Götz (2013a) found **three learner (dis)fluency profiles**. The first cluster included learners with a high temporal fluency, a high frequency of formulaic language (i.e. 3- and 4-grams), but a low proportion of filled pauses and of repetitions. The other two clusters were characterised by an average temporal fluency, but whereas it was characteristic of one group to use a high proportion of formulaic language (and a low proportion of other (dis)fluency features), the reverse was observed for the other group, which was characterised by a comparatively low proportion of 3- and 4-grams and a high frequency of filled pauses, discourse markers and repetitions. Although straightforward comparisons with LINDSEI-FR+ are restricted due to a different operationalisation of (dis)fluency, it is important to underline that the temporal variables also seem to play an important role in the clustering of German-speaking learners, albeit formulaic language also seems to successfully distinguish groups. Götz (*ibid*.) also found **three clusters of native speakers** whose characterisation, overall, corroborates the findings of the present analysis in terms of distinguishing variables.

The **second step of the Cluster Analysis** then sought to determine whether more fine-grained usage patterns existed in the learner and native speaker groups. In the **learner** data, **6 (dis)fluency profiles** could be identified: these differed along 13 of the 14 (dis)fluency variables (all variables but conjunctions) and the first three (dis)fluency components. In the **native speaker** data, **5 (dis)fluency profiles** were distinguished by 10 (dis)fluency variables: the same variables as for the 2-cluster solution as well as false starts, foreign words and repetitions. The fact that the learner profiles differ along proportionally more (dis)fluency variables than the native profiles confirms that heterogeneity is a level of magnitude higher in LINDSEI-FR+ than in the native corpus.

An interesting finding was also that the 6 learner clusters, unlike the native speaker clusters, differ with respect to restarts, repetitions *and* truncations, i.e. the three **repair (dis)fluency** variables of Component 2. LOCNEC+ speakers only differ in their use of *repetitions* (statistical analyses did not reveal significant between-group differences). These results stress the **importance of the repair variables in learner (dis)fluency**, as compared to its relatively less salient role in native (dis)fluency.

More generally, by crossing the results of the Principal Components Analyses with the results of the Cluster Analyses (see Table 6-25 and Table 6-26 below), it becomes apparent that **learner and native (dis)fluency depends less on the use of individual (dis)fluency variables or components than on how these are used in combination with one another**. There seem to be at least three different ways (dis)fluency variables and components are combined:

- The **compensators**. In LINDSEI-FR+, the speakers of Profiles B, C and F seem to be "able to compensate for potential deficiencies [...] by capitalising on a few of their strengths" (Jarvis *et al.* 2003:399). Brand & Götz (2011:267; see also Götz 2013a) explain that:

  > the learners may not have internalised the complete nativelike variety of variables that contribute to fluency and, as a result, may use one variable much more frequently than another to establish their spoken fluency. In other words, they may show a very poor performance concerning one fluency variable, but "make up for that", as it were, by a very good performance in another and thus may establish their overall fluency performance through different means.

  For example, in Profile C, the learners could be argued to make up for a weakness in pragmatic (dis)fluency by a very good performance on the temporal variables. In Profile F, a weakness in temporal (dis)fluency seems to be counterbalanced by a good (i.e. non-excessive) use of repair variables. To some extent, the native speakers corresponding to Profiles D and E are also examples of "compensators" because such speakers seem to make up for a somewhat weaker performance on Component 2 (Profile D) or Component 4 (Profile E) by a (very) good performance on the temporal variables.

- The **(over and under) performers**. Some profiles display a **flawlessly good (or bad) performance** across the various (dis)fluency variables and dimensions. Learner Profile A, for example, gathers good performances on the variables of Component 1 (high temporal (dis)fluency), of Component 2 (low use of repair variables) *and* of Component 3 (good use of pragmatic (dis)fluency). By contrast, Profile E accumulates weaker performances on the variables of Component 1, 2 *and* 3. In the native speaker corpus, Profile A is also a performer as it cumulates weaker performances on the variables of Component 1 (a lower temporal fluency), Component 2 (a high rate of repetitions) and Component 3 (a lower pragmatic (dis)fluency). To a lesser extent, Profile B is also a gatherer of weaker performances. Interestingly, it is worth underlining that the profile "over-performer" does not seem to have a native counterpart and future investigations could probe further into this profile to examine, for example, whether these learners with excellent productive fluency are also perceived as particularly fluent.

- The **averagers**. The learner Profile D and the native Profile C have average performances on the majority of (dis)fluency variables and dimensions. In a sense, they are a special type of performers, who display an average performance across (nearly) all (dis)fluency variables and components.

From the above discussion, it seems possible to suggest a **tentative ranking of the learner and native profiles along a (dis)fluency scale**. The learner profiles could be ranked as follows, from the most fluent to the least fluent:

Profile A (over-performers) > Profiles C, B and F (compensators) >
Profile D (averagers) > Profile E (under-performers)

Likewise, the native profiles could be ranked in the following order:

Profile E and D (compensators) > Profile C (averagers) >
Profiles B and A (under-performers)

Such rankings are obviously still highly speculative at this stage, and further analyses would be needed to corroborate both the profiles and their ranking. In Chapter 7, I will take a first step in addressing this aspect by relating the profiles with CEFR levels and descriptors.

(Dis)fluency variables	Cluster A	Cluster C	Cluster B	Cluster F	Cluster D	Cluster E
Component 1 Temporal (dis)fluency	☺	☺☺	☺	☹	☹	☹☹
Unfilled pauses	-	--		+	+	+
Phonation time ratio	+	+				---
Mean length of runs		++		-	-	-
Speech rate	+		+	-		-
Component 2 Repair (dis)fluency	☺		☹☹	☺		☹☹
Truncations	-		+++	-		+
Restarts			+++	-		
Repetitions			++			++
Component 3 Pragmatic (dis)fluency	☺	☹☹		☺		
Discourse markers	+	-				
Filled pauses	-	+		-		
Component 4 Discourse cohesion			☹☹			
Conjunctions			+			
False starts			+++			
Component 5 Lexico-grammatical disfluency			☹☹			☺
Foreign words			++			-
False starts			+++			

Table 6-25: Crossing the results of the PCA and CA (LINDSEI-FR+)
Note: The 6 learner profiles are (tentatively) ranked in decreasing order of fluency

(Dis)fluency variables		Cluster E	Cluster D	Cluster C	Cluster B	Cluster A
Component 1 Temporal (dis)fluency		☺☺	☺		☹	☹☹
	Mean length of runs	+++			-	-
	Unfilled pauses	-	-		++	+
	Phonation-time ratio	++	+		-	-
	Speech rate	++			-	--
	Filled pauses					++++
Component 2 Repair (dis)fluency		☹	☺			☹
	Truncations					
	Restarts					
	Repetitions		-			+++
	Lengthenings	+				
Component 3 Pragmatic (dis)fluency						☹☹
	Discourse markers					--
	Filled pauses					++++
Component 4 Discourse cohesion			☹	☺	☹	
	Conjunctions		+			
	False starts		+	-	+	

Table 6-26: Crossing the results of the PCA and the CA (LOCNEC+)
Note: The 5 native profiles are (tentatively) ranked in decreasing order of fluency

6.2.4 Limitations

Before concluding the chapter, some **limitations** of the Cluster Analyses merit a brief discussion.

I firstly acknowledge the potential limitations related to the way I conducted this study. The outcome of a cluster analysis is contingent on at least two important settings, namely the (dis)similarity measure (here, squared Euclidean distance), and the clustering algorithm which determines how the clusters are amalgamated based on their level of (dis)similarity (here: Ward's method). As stated by Divjak & Gries (2006:37), "[t]here is no uniformly accepted combination of parameters that guarantees an optimal clustering solution" and, accordingly, a re-analysis of the same data with other clustering settings might generate somewhat different results. In addition, the choice of the adequate number of clusters partly depends on the interpretation of the researcher. In an attempt to limit the bias necessarily

involved in such a choice, I have deliberately opted to present two plausible solutions for LINDSEI-FR+ (2 and 6 clusters) and LOCNEC+ (2 and 5 clusters), all the while also acknowledging that the findings related to the 5- and 6-cluster solutions should be embraced conservatively as some of these clusters contain **few speakers** only.

It is also important to bear in mind that Cluster Analysis is a **squarely descriptive** technique and that the clusters (i.e. (dis)fluency profiles) resulting from a Cluster Analysis are necessarily **affected by background variables** such as the nature of the speaking task or the proficiency level of the learners as well as by the linguistic **variables and their measurement**. Performing the same analysis on a different speaking task, on speakers from another mother tongue background or at a different proficiency level would bring about different results. Performing the same analysis with another set of (dis)fluency variables or with another level of measurement (e.g. per minute) would also result in a different classification. For example, due to a slightly different operationalisation of (dis)fluency, Götz (2011; 2013a), who also used a Cluster Analysis on LOCNEC in order to identify (dis)fluency profiles, found three clusters of native speakers (*cf*. also Section 1.2.5 for more details on the results of her study), whereas I found two or five. Therefore, although the statistical analysis could identify and characterise several learner and native speaker (dis)fluency profiles, the findings strictly apply to the speakers of LINDSEI-FR+ and of LOCNEC+ based on an operationalisation of (dis)fluency into the 14 variables listed in Table 3-6. Findings should not be extrapolated to other learner or native speaker groups, and cross-study comparisons should be couched with all the caution warranted.

6.3 CONCLUSION

Much of the previous work on learner and native speaker (dis)fluency has concentrated on the analysis of separate (dis)fluency variables, but few studies have attempted to sort of the nature and extent of the relationship between these variables. The primary concern of this chapter was precisely the investigation of such relationships, first by looking at aggregate data, and second, by examining individual differences.

I first presented the results of two **Principal Components Analyses**, which aimed at identifying underlying dimensions of learner and native speaker (dis)fluency. In **LINDSEI-FR+, five dimensions** (or "components") were retained. The first component pertains to the temporal aspect of (dis)fluency. The second component was associated with repair (dis)fluency and the third with pragmatic (dis)fluency. The fourth and the fifth components were more difficult to interpret: the former was linked to discourse cohesion, and the latter to lexico-grammatical (dis)fluency. In **LOCNEC+, four components** emerged from the analysis. They are the native counterparts of the first four components identified in LINDSEI-FR+. The inner structure of these components is, however, slightly different across the learner and the native version of the components. For example, filled pauses were shown to contribute to the temporal (dis)fluency component in native speech, which was not the case in learner discourse.

Then, in the second part of the chapter, I submitted the data to two **Cluster Analyses** with a view to identifying groups of learners and groups of native speakers who perform similarly across the 14 (dis)fluency variables under investigation. In both the learner corpus and the native corpus, **two main (dis)fluency profiles** were identified which differed with respect to temporal (dis)fluency variables (and discourse markers in LOCNEC+). More detailed profiles were also presented, which showed that **(dis)fluency depends less on individual (dis)fluency features than on how these are combined**.

Statistics are a key feature of this chapter. These cannot be usefully employed apart from a theoretically-motivated research design. That is, before conducting statistical tests, the range of (dis)fluency features must be determined and the variables measured, and the tests are dependent on this foundation. In this respect, a key strength of this chapter is that it has demonstrated the power and usefulness of the (dis)fluency **annotation** and the **time alignment** of the corpora, despite some caveats; none of the multivariate analyses presented in this chapter would have been possible without them, or, at best, they would have been far reduced in scope. A great advantage of conducting **exploratory statistical techniques** such as principal components analyses and cluster analyses is that the data is not oriented (or skewed) towards an a priori hypothesis. However, the major difficulty of such multivariate

tests lies in the **interpretation of the results**, which is sometimes very tentative and in need of corroboration from other studies.

Taken together, the results of the two analyses statistically indicate that (dis)fluency is primarily a temporal phenomenon (*cf.* Fillmore 1979; Grosjean 1980c), but that it is certainly not restricted to it and also involves other aspects, such as the ability to cope with stumbles satisfactorily (e.g. Skehan 1999; 2014; Towell 2012), especially in learner language, or the ability to use pragmatic markers adequately (e.g. Denke 2009; Hasselgren 2002). Despite their exploratory nature, the analyses also contributed to our knowledge of (dis)fluency by suggesting that the way (dis)fluency features or components are combined may correspond to different types or **levels of fluency**, or even to different **types of learners**. It might, for example, be interesting to investigate the learner clusters from the perspective of the Monitor Hypothesis (Krashen 1982; Krashen 1983) or to relate them to perceived fluency levels (see Chapter 7 in this respect).

A major **implication** of the findings from this chapter is the calling into question of fluency **assessment grids and descriptors**. For example, (dis)fluency dimensions could be integrated into fluency descriptors. For assessors and teachers, it might also be easier and more efficient to focus on a limited number of (dis)fluency dimensions than on a wide range of (in fact, dependent) individual features during the assessment of a learner's fluency level. Besides, the role of **individual variation** should definitely not be underestimated: the cluster analyses showed that there exist different profiles, both among learners and native speakers. It should be recognised that (learner and native) fluency can be pluralised, and, finally, it should be accepted that the target of language learning should not be a perfect, pause- or repetition-free, discourse.

Chapter 7

LINKING UP LEARNERS' PRODUCTIVE (DIS)FLUENCY, THE CEFR FLUENCY SCALE AND ASSESSED CEFR FLUENCY RATINGS

Nobody trips over mountains.
It is the small pebble that causes you to stumble.
Pass all the pebbles in your path and you will
find that you have crossed the mountain.

Unknown

Whereas the previous chapter encompassed both learner and native speaker data, the present chapter examines the French-speaking learners' (dis)fluency from the perspective of the *Common European Framework of Reference for Languages* (CEFR, Council of Europe 2001). The chief aim of this chapter is to investigate the tripartite relationship between learners' productive (dis)fluency, the CEFR fluency scales and descriptors, and assessed CEFR fluency levels.

This chapter is made up of five main sections. Section 7.1 examines the scales and descriptors of the *Common European Framework of Reference for Languages* for language production skills. The second section (Section 7.2) describes the rating procedure that was followed to assess the level of fluency of the 50 French-speaking learners of LINDSEI-FR+. In the third section (Section 7.3), I discuss some methodological issues related to the use of CEFR ratings. Section 7.4 then sets out the results of the analysis of the relationship between CEFR fluency ratings and learner language. Finally, conclusions are drawn in Section 7.5.

7.1 THE CEFR FLUENCY SCALE UNDER SCRUTINY

7.1.1 The Common European Framework of Reference

The *Common European Framework of Reference for Languages: Learning, teaching and assessment* (CEFR) was developed by the Council of Europe and published in 2001[149] to provide a "common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe" (Council of Europe 2001:1). As the subtitle "*learning, teaching, assessment"* makes clear, the CEFR is not restricted to assessment. The three stated aims of the CEFR are (1) to promote and facilitate cooperation among educational institutions and language practitioners in different countries, (2) to provide a basis for mutual recognition of language skills, and (3) to assist teachers, learners and course designers to coordinate their efforts (Council of Europe 2001:xi; Council of Europe 2017:25–26).

From a historical perspective, the creation of the CEFR resulted from the conjunction of two factors. First, it was born from the transition from the traditional grammar-translation method to two innovative approaches to language learning where language is seen as a purposeful tool to achieve goals: the **functional-notional approach** and the **communicative action-oriented approach**. These two approaches are promoted in the framework and are probably best reflected in the level descriptions, where each level is described in terms of what the learners can do and how well they can do it (i.e. the famous "can-do statements"). The second factor that gave rise to the CEFR was the **need for an international framework** for language learning that would facilitate collaboration between European educational institutions, for example, by making language mastery levels comparable among countries (Little 2007; Broek & van den Enden 2013).

Given the fact that the *Common European Framework of Reference* was originally intended as a **general guide** rather than as a prescriptive tool or an international standard, it is **language- and context-neutral**. The CEFR does not, for example, provide a checklist of language features or learning points for learners at a specific level, and users are supposed to adapt the

[149] The "CEFR Companion Volume with New Descriptors" was issued online in September 2017 (https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/168074a4e2 last accessed 11/10/2017). This Companion is intended "as a complement to the CEFR" and "does not change the status of that 2001 publication" (Council of Europe 2017:23). At the time of the writing of this thesis, the only available version of the 2017 Companion is still a "provisional edition". Moreover, after close inspection, it appeared that the descriptor scales for speaking skills (and fluency in particular) did not undergo drastic change. In what follows, I will thus primarily refer to the 2001 publication, and underline differences with the 2017 Companion only when relevant.

CEFR to the language and learning context they are working with. To ensure that the CEFR can be adapted to national languages, the Council of Europe has encouraged the production of so-called **Reference Level Descriptions**[150] (RLDs) where the language-neutral CEFR levels and descriptors "have been mapped against the actual linguistic material" (ESOL Examinations 2011:4). In other words, RLDs provide language-specific descriptions of the language that learners know and use at each CEFR level. For example, the RLD for English, also known as the *English Profile*[151], resulted in two online resources, namely the *English Vocabulary Profile* and the *English Grammar Profile.* These tools offer information about which words or grammatical structures learners typically know at each CEFR level, and what vocabulary or grammar is suitable for teaching at each level.

The CEFR distinguishes **six levels of language mastery**: two "basic user" levels (*Breakthrough* and *Waystage*), two "independent user" levels (*Threshold* and *Vantage*), and two "proficient user" levels (*Effective operational proficiency* and *Mastery*). Each level is defined in the form of can-do statements in the *Global scale* (Table 7-1). This scale – like the other CEFR scales – can be read horizontally or vertically: the horizontal axis offers a description of the different aspects of linguistic competence per level, and the vertical axis represents progress in proficiency. As the name indicates, the *Global scale* aims to qualify the overall abilities typical of learners at each level. It thus includes, but is not restricted to, spoken skills (see bold font).

Proficient User	Mastery (C2)	- Can understand with ease virtually everything heard or read. - Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. - Can express him/herself spontaneously, **very fluently** and precisely, differentiating finer shades of meaning even in more complex situations.
	Effective operational proficiency (C1)	- Can understand a wide range of demanding, longer texts, and recognise implicit meaning. - Can express him/herself **fluently** and spontaneously without much obvious searching for expressions. - Can use language flexibly and effectively for social, academic, and professional purposes. - Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors, and cohesive devices.
Independent User	Vantage (B2)	- Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. - Can interact with **a degree of fluency** and spontaneity that makes regular interaction with native speakers quite possible without strain for either party.

[150] The list of language-specific descriptors is available at: https://www.coe.int/t/dg4/linguistic/DNR_EN.asp#P42_6429 (last accessed 15/03/2018).

[151] http://www.englishprofile.org/ (last accessed 11/10/2017).

		- Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	Threshold (B1)	- Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. - Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. - Can produce simple connected text on topics which are familiar or of personal interest. - Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	Waystage (A2)	- Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). - Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. - Can describe in simple terms aspects of his/her background, immediate environment, and matters in areas of immediate need.
	Breakthrough (A1)	- Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. - Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. - Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Table 7-1: The CEFR Global scale (from Council of Europe 2001:24)

Level **A1** (Breakthrough) is considered the lowest level in the framework, though it "is not the lowest imaginable level of proficiency" (Council of Europe 2017:35). At this level, emphasis is laid on simplicity of the linguistic output and the familiarity of the communicative tasks. Likewise, the descriptor for **A2** (Waystage) stresses the simplicity of output and the fact that the learner can engage in routine tasks. The next band, **B1**, reflects the Threshold level: it is characterised by the "ability to maintain interaction and get across what you want to, in a range of contexts" and "the ability to cope flexibly with problems in everyday life" (Council of Europe 2001:34). The subsequent level (**B2**) reflects a metaphorical vantage point in language acquisition: learners are now able to deal with complex and varied tasks. The next level (i.e. **C1**) is characterised by a good access to and flexible use of a broad range of language. The last – and highest – level in the CEFR is **C2**. Although it is also termed Mastery, the authors of the CEFR Companion emphasise that this top level "has no relation whatsoever with what is sometimes referred to as the performance of an idealised 'native-speaker', or a 'well-educated native speaker' or a 'near-native speaker'. Such concepts were not taken as a point of reference during the development of the levels or the descriptors" (Council of Europe 2017:35).

From the point of view of fluency, it is interesting to underline that, in the *Global Scale* displayed in Table 7-1, **learners at B2 level upwards are explicitly said to be fluent** (see bold

font). Fluency shows **three apparent degrees** at the three highest levels: while there is "a degree of fluency" at B2 level, C1 learners are said to be "fluent" and learners at the highest level can express themselves "very fluently". Spontaneity is also stressed at those levels. At the lower levels (A1 to B1), fluency (or the lack of it) does not seem to be a salient feature worth mentioning in the general descriptors.

7.1.2 Orality in the CEFR

As visually represented in Figure 7-1, the *Common European Framework* draws a distinction between "communicative language activities" (Council of Europe 2001:57–90) and "communicative language competences" (*ibid*.:108-130).

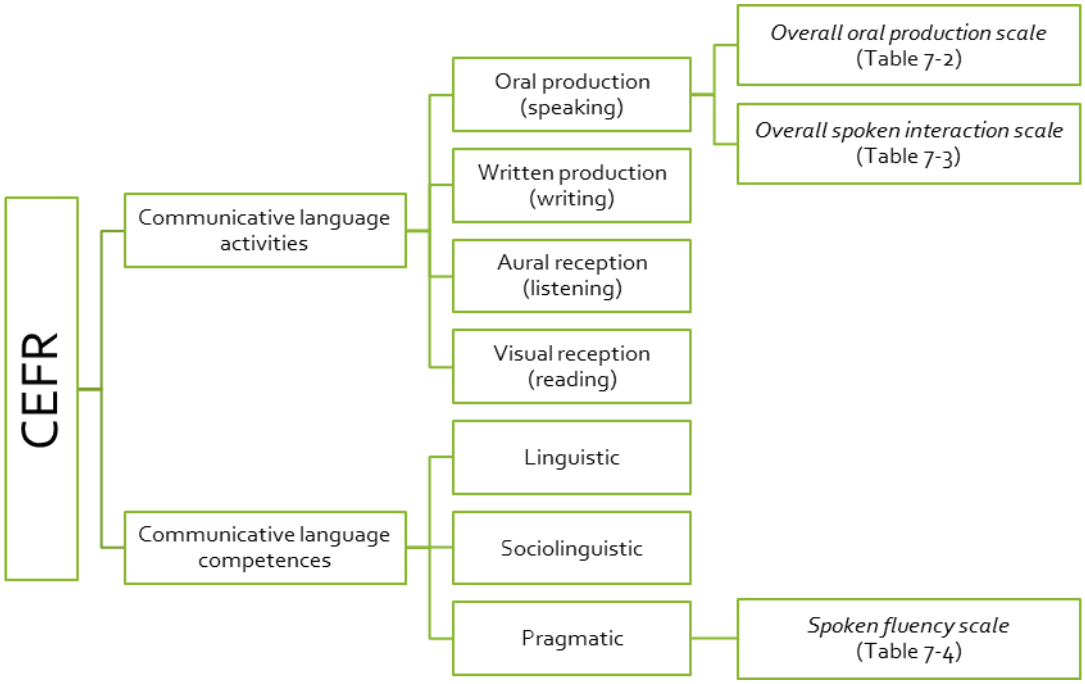


Figure 7-1: Overview of the CEFR

Communicative language activities subsume four language skills, namely listening, reading, speaking, and writing. While the first two are said to be receptive language activities, speaking and writing can be either productive or interactive activities (i.e. activities that combine reception and production). Several sub-skills are distinguished within each of the four skills and come with their corresponding "**illustrative scale**" for each of the six reference levels. For example, Table 7-2 displays the five illustrative scales provided for the "oral production" skill (i.e. monologic speech) and the nine illustrative scales provided to the "spoken interaction" skill (i.e. dialogic speech) (Council of Europe 2001:58–60, 73–82). In the illustrative scales descriptors, a distinction is often made between the "criterion levels" (e.g. A2, B1, B2) and the "plus levels" (e.g. A2+, B1+, B2+), which represent a stronger performance

within the band, but which does not yet reach the minimum standard for the following criterion level.

Oral production
Overall oral production
Sustained monologue: describing experience
Sustained monologue: putting a case (e.g. in debate)
Public announcements
Addressing audiences

Spoken interaction
Overall spoken interaction
Understanding a native speaker interlocutor
Conversation
Informal discussion
Formal discussion and meetings
Goal-oriented co-operation
Transactions to obtain goods and services
Information exchange
Interviewing and being interviewed

Table 7-2: Illustrative scales for oral production and spoken interaction

The *Overall oral production scale* and the *Overall spoken interaction scale*, which are the most relevant for the present study among the aforementioned illustrative scales, are shown in Table 7-3 and Table 7-4, respectively.

References to fluency are rather scarce in the ***Overall oral production scale*** (Table 7-3). The lower levels seem to emphasise that learners can only produce short and isolated utterances. B1 learners are able to sustain a description "reasonably fluently". The B2 and C1 descriptors do not refer to fluency or fluency phenomena at all: at those levels, the emphasis seems to shift on the clarity of exposition instead. At the highest level, speech is said to be "smoothly flowing", which might imply that very little to no disfluencies occur at all. It is worth underlining that, contrary to its name (the *overall* spoken production scale), the scale seems to be designed for one particular type of monologic task, namely descriptions, and nothing is said about other types of monologic tasks.

C2		Can produce clear, **smoothly flowing** well-structured speech with an effective logical structure which helps the recipient to notice and remember significant points.
C1		Can give clear, detailed descriptions and presentations on complex subjects, integrating sub-themes, developing particular points, and rounding off with an appropriate conclusion.
B2	+	Can give clear, systematically developed descriptions and presentations, with appropriate highlighting of significant points, and relevant supporting detail.

		Can give clear, detailed descriptions and presentations on a wide range of subjects related to his/her field of interest, expanding and supporting ideas with subsidiary points and relevant examples.
B1		Can **reasonably fluently** sustain a straightforward description of one of a variety of subjects within his/her field of interest, presenting it as a linear sequence of points.
A2		Can give a simple description or presentation of people, living or working conditions, daily routines, likes/dislikes, etc. as a short series of simple phrases and sentences linked into a list.
A1		Can produce simple mainly isolated phrases about people and places.

Table 7-3: CEFR Overall Oral Production scale (Council of Europe 2001:58)

C2		Has a good command of idiomatic expressions and colloquialisms with awareness of connotative levels of meaning. Can convey finer shades of meaning precisely by using, with reasonable accuracy, a wide range of modification devices. **Can backtrack** and restructure around a difficulty so smoothly the interlocutor is hardly aware of it.
C1		Can express him/herself **fluently** and spontaneously, almost effortlessly. Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions. There is **little obvious searching** for expressions or avoidance strategies; only a conceptually difficult subject can hinder a **natural, smooth flow** of language.
B2	+	Can use the language **fluently**, accurately and effectively on a wide range of general, academic, vocational or leisure topics, marking clearly the relationships between ideas. Can communicate spontaneously with good grammatical control without much sign of having to restrict what he/she wants to say, adopting a level of formality appropriate to the circumstances.
		Can interact with **a degree of fluency** and spontaneity that makes regular interaction, and sustained relationships with native speakers[152] quite possible without imposing strain on either party. Can highlight the personal significance of events and experiences, account for, and sustain views clearly by providing relevant explanations and arguments.
B1	+	Can communicate with some confidence on familiar routine and non-routine matters related to his/her interests and professional field. Can exchange, check and confirm information, deal with less routine situations and explain why something is a problem. Can express thoughts on more abstract, cultural topics such as films, books, music etc.
		Can exploit a wide range of simple language to deal with most situations likely to arise whilst travelling. Can enter unprepared into conversation on familiar topics, express personal opinions, and exchange information on topics that are familiar, of

[152] "With native speakers" was modified into "with speakers of the target language" in the 2017 Companion (Council of Europe 2017:217). This suggests that the new descriptors of the CEFR aim to take into account considerations from the domain of English as a Lingua Franca (ELF).

		personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events).
A2	+	Can interact with **reasonable ease** in structured situations and short conversations, provided the other person helps if necessary. Can manage simple, routine exchanges without undue effort; can ask and answer questions and exchange ideas and information on familiar topics in predictable everyday situations.
		Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters to do with work and free time. Can handle very short social exchanges but is rarely able to understand enough to keep conversation going of his/her own accord.
A1		Can interact in a simple way but **communication is totally dependent on repetition at a slower rate of speech, rephrasing and repair**. Can ask and answer simple questions, initiate and respond to simple statements in areas of immediate need or on very familiar topics.

Table 7-4: CEFR Overall spoken interaction scale (Council of Europe 2001:74)

In the ***Overall spoken interaction scale*** (Table 7-4), fluency seems to be slightly more salient. At the lowest level, fluency is clearly lacking in the interaction, as "communication is totally dependent on repetition at a slower rate of speech, rephrasing and repair". At A2 level, speakers are said to be able to interact with "reasonable ease", provided the interlocutor is willing to help in case of communicative breakdown. At the threshold level, learners show more confidence in their language skills, and are able to engage into unprepared or less routine dialogues. When their reach B2 level, speakers can interact "with a degree of fluency" (and "fluently" at B2+). Likewise, C1 level speakers are characterised by their fluent, natural, and smooth speech. The descriptor also specifies that, although C1 speakers still sometimes hesitate, there is "little obvious searching for expressions". Lastly, the descriptor for the highest level implies that, despite the fact that learners may still encounter some difficulties, they are nonetheless perceived as fluent because they are able to "backtrack and restructure around a difficulty so smoothly the interlocutor is hardly aware of it".

Besides the four main skills (and their sub-skills), the *Common European Framework* also develops the idea of "**communicative language competences**" (see Figure 7-1). These competences may be either linguistic, sociolinguistic or pragmatic. Pragmatic skills subsume discourse and functional competences. The former refers to "the ability of a user/learner to arrange sentences in sequence so as to produce coherent stretches of language" (Council of Europe 2001:123) and "functional competence" is concerned with "the use of spoken discourse and written texts in communication for particular functional purposes" (*ibid*.:125). According to the CEFR, the **two qualitative factors** that **determine the functional success of a language learner** are propositional precision, that is, the "ability to formulate thoughts and propositions so as to make one's meaning clear" (*ibid*.:128) and fluency, defined as the "ability to articulate, to keep going, and to cope when one lands in a dead end" (*ibidem*).

C2		Can express him/herself at length with a natural, effortless, unhesitating flow. Pauses only to reflect on precisely the right words to express his/her thoughts or to find an appropriate example or explanation.
C1		Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.
B2	+	Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech.
		Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions, there are few noticeably long pauses. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers[153] quite possible without imposing strain on either party.
B1	+	Can express him/herself with relative ease. Despite some problems with formulation resulting in pauses and 'cul-de-sacs', he/she is able to keep going effectively without help.
		Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.
A2	+	Can make him/herself understood in short contributions, even though pauses, false starts, and reformulation are very evident.
		Can construct phrases on familiar topics with sufficient ease to handle short exchanges, despite very noticeable hesitation and false starts.
A1		Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.

Table 7-5: The CEFR Spoken Fluency scale (Council of Europe 2001:129)

The scale for fluency, i.e. one of the two qualitative factors of functional success, is shown in Table 7-5. The lowest level in the ***Spoken fluency scale*** seems to suggest the absence of fluency because, not only do learners produce "very short, isolated, mainly pre-packaged utterances", but they also (need to) devote much time to pausing, articulating, and repairing. At the next level (A2), learners are able to handle short exchanges with "sufficient ease", albeit with "very noticeable hesitations and false starts". At the plus level (A2+), pauses, false starts, and reformulations are "very evident". What seems to characterise learners at B1 level (compared to the lower levels) is that they can speak in longer speech runs ("can keep going", "longer stretches"). Pausing, however, remains very evident, and is used for lexico-grammatical planning as well as for repair communication. B1+ speakers can express themselves "with relative ease", but still encounter problems and produce pauses and "cul-de sacs". At B2 level, although hesitations might occur, learners can adopt a "fairly even

[153] "Native speakers" was changed into "speakers of the target language" in the 2017 Companion (Council of Europe 2017:217).

tempo" (e.g. they produce few long pauses), can communicate "with a degree of fluency and spontaneity" or even with "remarkable fluency" (B2+). At the two highest levels, effortlessness of speech production seems to be key. Whereas at C1 level, hesitations might occur as a result of a "conceptually difficult subject", C2 learners are able to use pauses strategically, for example to reflect on their word choice.

7.1.3 The qualitative aspects of spoken language use scale

The vast majority of the 54 scales provided in the CEFR are descriptors of specific communicative activities (*cf.* e.g. Table 7-3 and Table 7-4). The framework also offers an alternative to assess a performance on the basis of the aspects of communicative language competence. The *Qualitative aspects of language use* scale (Council of Europe 2001:28–29) was for example designed to include **five qualitative aspects of language use**, namely:

- **Range**, which corresponds to lexical and grammatical diversity and complexity;
- **Accuracy**, or grammatical correctness;
- **Fluency**, i.e. the smoothness of speech;
- **Interaction**, or the ability to engage in (and manage) the conversation;
- **Coherence**, or the ability to use "organisational patterns", "connectors", and "connecting devices" in the discourse.

It is interesting to underline that the *Qualitative aspects of language use* scale was expanded with additional descriptors for **phonology** in the 2017 companion of the CEFR (the other descriptors have not been modified). In this new edition, the scale is called the *Qualitative features of spoken language* (Table 7-6). The *Qualitative aspects of spoken language use* scale (i.e. the 2001 scale, or Table 7-6 excluding the descriptors for phonology) is of particular interest to this study as an adaptation of this scale was **used to assess the proficiency level of a sample of each component of LINDSEI** (Gilquin, De Cock & Granger 2010:10–11) **and of each French learner of English within the French component** (see Section 7.2).

In the *Qualitative features of spoken language scale* (Table 7-6), fluency at the two A levels is mainly characterised by very short utterances. Learner speech at those levels also features a great number of "very evident" pauses, false starts and reformulations, as well as issues with articulation. At the intermediate stages, the length of speech runs is significantly longer. Pausing (for grammatical and lexical planning) and repairs are typical of B1 fluency, while "hesitations" are emblematic of B2. As described in Table 7-5, B2 learners can speak at a "fairly even tempo", and produce "few noticeably long pauses". As the learners move to C1 level, they are said to be able to speak "fluently and spontaneously, almost effortlessly". Difficulties only arise with conceptually difficult topics. C2 learners are characterised by their ability to cope with difficulties. In fact, they can cope so well that "the interlocutor is hardly aware of it".

	Range	Accuracy	Fluency	Interaction	Coherence	Phonology
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously **at length** with a natural colloquial flow, avoiding or **backtracking** around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.	Can employ the full range of phonological features in the target language with a high level of control – including prosodic features such as word and sentence stress, rhythm and intonation – so that the finer points of his/her message are clear and precise. Intelligibility is not affected in any way by features of accent that may be retained from other language(s).
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/ herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, **smooth flow** of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skilfully to those of other speakers.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors, and cohesive devices.	Can employ the full range of phonological features in the TL with sufficient control to ensure intelligibility throughout. Can articulate virtually all the sounds of the TL; some features of accent retained from other language(s) may be noticeable, but they do not affect intelligibility at all.
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a **fairly even tempo**; although he/she **can be hesitant** as he or she searches for patterns and expressions, there are	Can initiate discourse, take his/her turn when appropriate and end conversation when he / she needs to, though he /she may not always do this elegantly. Can help the discussion along on familiar ground confirming	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.	Can generally use appropriate intonation, place stress correctly and articulate individual sounds clearly; accent tends to be influenced by other language(s) he/she speaks, but has little or no effect on intelligibility.

	complex sentence forms to do so.		**few noticeably long pauses**.	comprehension, inviting others in, etc.		
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Can **keep going** comprehensibly, even though **pausing** for grammatical and lexical planning and **repair** is very evident, especially in longer stretches of free production.	Can initiate, maintain, and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.	Pronunciation is generally intelligible; can approximate intonation and stress at both utterance and word levels. However, accent is usually influenced by other language(s) he/she speaks.
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in **very short utterances**, even though **pauses, false starts and reformulation** are very evident.	Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.	Can link groups of words with simple connectors like "and, "but" and "because".	Pronunciation is generally clear enough to be understood, but conversational partners will need to ask for repetition from time to time. A strong influence from other language(s) he/she speaks on stress, rhythm and intonation may affect intelligibility, requiring collaboration from interlocutors. Nevertheless, pronunciation of familiar words is clear.
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire.	Can manage **very short**, isolated, mainly prepackaged utterances, with much **pausing** to search for expressions, to articulate less familiar words, and to **repair** communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear Connectors like "and" or "then".	Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by interlocutors used to dealing with speakers of the language group concerned. Can reproduce correctly a limited range of sounds as well as the stress on simple, familiar words and phrases.

Table 7-6: CEFR Common Reference Levels: Qualitative Features of Spoken Language Use (Council of Europe 2017:156)

7.1.4 A critical perspective on the CEFR scales

Several elements contributing to fluency can be identified from Table 7-5 and Table 7-6. These can be classified into two broad categories: **quantitative criteria** – how many things the language learner can do in the target language – and **qualitative criteria** – how well he/she can do them. Table 7-6, for example, describes learners' fluency with the following quantifiable (i.e. measurable) criteria: length of runs ("stretches of language"), speech rate ("tempo"), number of filled and unfilled pauses ("hesitations"), number of false starts (and "cul-de sacs"), and number of reformulations and repairs. The descriptor scale also describes learner fluency in qualitatively terms, with adjectives such as "spontaneously", "effortlessly", "natural", and "smooth".

The interpretation and **application of the CEFR descriptors**, however, pose several problems. Many of these are already well-discussed in the literature (see e.g. Alderson 2007; Davidson & Fulcher 2007; Fulcher 1996; Hulstijn 2007; Isaacs & Thomson 2013; Iwashita *et al.* 2008; Osborne 2011a), particularly the observation that the CEFR scales are barely "**scaled teacher perceptions**" (North 2014:23; my emphasis) that were neither derived from a model of L2 competence nor matched onto actual learner data. It is, moreover, far from clear how much attention has been paid to recent **empirical findings from research into second language acquisition and learner corpus research**, especially in the new 2017 edition of the CEFR (Council of Europe 2017). Due to space limitations, I will not discuss this aspect further, but I will summarise some of the issues most relevant for the present research.

Firstly, the descriptors identify a number of **(dis)fluency features**, but these **lack precise definitions**, which often leads to ambiguity over the meaning or scope of some terms. It is, for example, unclear whether the term "hesitations" exclusively refers to filled pauses, or whether it also encompasses other phenomena. Similarly, it is not evident whether "cul-de-sacs" and "false starts", or "reformulations" and "repairs" are synonyms referring to exactly the same concepts.

Moreover, although (dis)fluency features are quantifiable, they are actually **not quantified**. Some features are said to be more or less "evident" up to B1 level. Not only is the interpretation of "evident" subjective, but it is also not explicitly said in the higher-level descriptors whether the features are altogether absent, or simply not noticeable. In addition, it is left to the expertise and experience of the evaluator whether a given L2 learner can produce "few(er)" instances of, for example, reformulations, "noticeable" pauses or "long(er)" stretches of language.

A related issue that was pointed out by Osborne (2011a:182) is that **the qualitative aspects of the scales involve a certain degree of subjectivity.** Whether a learner is able to express

him/herself "spontaneously", "with a natural colloquial flow", "relative ease", or "almost effortlessly" is dependent on the evaluator's own expertise.

It is widely recognised that descriptors and rating scales tend to **oversimplify** the processes involved in L2 production (Isaacs & Thomson 2013; Lumley 2005). The CEFR scale descriptors for fluency conform to this tendency as they **fall short of reflecting the plurality of (dis)fluency features**. Only five (dis)fluency features are explicitly mentioned in the descriptors, and no account is made of the **underlying dimensions of (dis)fluency** nor of the interactions between them[154]. In addition, no provision is made for learners who may score high on one (dis)fluency feature or (dis)fluency component but low on another.

Another limitation of the CEFR descriptors is that the **distinction between neighbouring levels** often relies on "downtoners and semantic niceties" (Osborne 2011a:182), such as "effortless" vs. "almost effortlessly"; "very evident" vs. "very noticeable", making it difficult for raters or teachers to apply them systematically (see also Jin, Mak & Zhou 2012).

The CEFR is "action-oriented" (Little 2007), but in the *Spoken fluency scale* as well as in the *Qualitative aspects scale*, no reference is made to different types of communicative situations. It seems to be assumed that the linguistic production of a learner at a given level does not vary across **communicative tasks**. More specifically, there appears to be a **bias towards monologic fluency** (*cf.* also the "monologic bias" in McCarthy 2010): there is no suggestion that fluency also involves the ability to create flow and smoothness across turn-boundaries in interactive settings. Although it is difficult to elaborate at this stage on the extent to which tasks should be embedded in rating scale descriptors, it is clear that a reappraisal of the descriptors should take account of previous research into task effect.

A last criticism concerns the **C2 level**. Despite the claim that the C2 level has "no relation" with native-speaker performance (Council of Europe 2017:35), the examination of the C2 descriptors suggests that this level might correspond to an idealised native speaker performance. In this respect, North (2007) acknowledges that many of the C2 descriptors were written up after the descriptor scales were elaborated and included "for the sake of completeness" (*ibid.*:657). The way in which some of these descriptors were added to the scales may be partly responsible for this criticism. Moreover, it is debatable whether native speakers could maintain a "natural, effortless, unhesitating flow" (C2 level, *Spoken fluency scale*) or could backtrack around any difficulty "so smoothly that the interlocutor is hardly aware of it" (C2 level, *Qualitative features* scale), particularly in the context of an oral examination.

[154] However, "the practical consideration of needing to provide raters with a user-friendly instrument with a manageable number of assessment criteria appears to be at odds with representing the construct comprehensively in descriptors" (Isaacs 2010:10).

For evaluators and teachers, the aforementioned issues create considerable leeway to rely on other factors in the decision-making process.

For researchers, an additional potential drawback of the CEFR is that it is not a numerical scale (like Likert scales, see Section 2.4). For many research purposes, CEFR levels thus first need to be converted into **numerical values**. For example, to analyse the correlation between (dis)fluency measures and CEFR fluency ratings [155], CEFR levels have to be translated numerically (e.g. Préfontaine, Kormos & Johnson 2015).

Despite those gaps and flaws, there is no doubt that the *Common European Framework* has had considerable impact on foreign language learning, teaching, and assessment (e.g. Little 2007) and it can be considered a potentially powerful tool in (dis)fluency research as well.

The next section sets out the methodology that was followed to assess the CEFR level of the 50 French-speaking learners of LINDSEI-FR+, including the inter-rater reliability analysis of the rating.

[155] For a review of the literature on the relationship between assessed fluency levels and objective (dis)fluency measures, see Section 2.4.2.

7.2 RATING LEANERS' CEFR (DIS)FLUENCY

7.2.1 Rating procedure

The rating procedure followed in this study is adapted from the procedure described in Gilquin *et al.* (2010). For the **release of LINDSEI in 2010**, excerpts from five learners of each component of the corpus were evaluated in order to provide researchers with a glimpse into the proficiency level(s) represented in each component. The extracts consisted in c. 5 minutes from the free discussion task, and the rater was asked to use a slightly adapted version of the CEFR *Qualitative Aspects of Spoken Language Use* (termed the "CEFR descriptor scale for linguistic competence").

In the frame of **this thesis**, CEFR fluency scores of the 50[156] learners of LINDSEI-FR+ were obtained from professionally-trained raters based on c. 5-minute recording extracts. The raters are **professionally-trained raters** who are native speakers of British English and have experience in rating both spoken and learner data. Originally, two raters (R1 and R2) were solicited for this rating task, and, following Thewissen (2012), a third rater (R3) was called upon to evaluate the interview excerpts on which R1 and R2 did not agree with regard to the global assessment score (*cf.* below). This procedure is known as the 2+1 principle, where a third judge is called in case of disagreement, and has been recommended by testing experts such as Alderson *et al.* (2001). However, in view of the large number of disagreements between R1 and R2 (28 out of 50 for global assessment and 27 out of 50 for fluency), it was decided to also have R3 assess all the learners and not only those on which R1 and R2 disagreed. Each learner in LINDSEI-FR+ has thus been assessed by **three raters**.

The raters were required to work with the **CEFR descriptor scale for linguistic competence** (Table 7-7). This scale targets **five distinct grades**, ranging from A2 to C2. The lowest level, A1, was not included, the reason being that learners in LINDSEI-FR+ are second or third-year university students majoring in English and whose proficiency is not at a beginner level. Gilquin *et al.* (2010) did not include this level either.

[156] The five French-speaking learners of LINDSEI-FR rated in the frame of Gilquin *et al.* (2010) were rated a second time, so as to follow exactly the same procedure as for the other 45 learners, especially so that the scores come from the same raters (i.e. fully-crossed design).

Linguistic competence	A2	B1	B2	C1	C2
Range	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.
Accuracy	Uses some simple structures correctly, but still systematically makes basic mistakes.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).
Fluency	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.
Phonological control	Pronunciation is generally clear enough to be understood despite a noticeable foreign accent, but conversational partners will need to ask for	Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.	Has a clear, natural, pronunciation and intonation.	Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.	Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.

	repetition from time to time.				
Coherence	Can link groups of words with simple connectors like "and", "but" and "because".	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors, and cohesive devices.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.
Global assessment	***Relates basic information on, e.g. work, family, free time etc.*** Can communicate in a simple and direct exchange of information on familiar matters. Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can describe in simple terms family, living conditions, educational background, present or most recent job. Uses some simple structures correctly, but may systematically make basic mistakes.	***Relates comprehensibly the main points he/she wants to make.*** Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair may be very evident. Can link discrete, simple elements into a connected, sequence to give straightforward descriptions on a variety of familiar subjects within his/her field of interest. Reasonably accurate use of main repertoire associated with more predictable situations.	***Expresses points of view without noticeable strain.*** Can interact on a wide range of topics and produce stretches of language with a fairly even tempo. Can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest. Does not make errors which cause misunderstanding.	***Shows fluent, spontaneous expression in clear, well-structured speech.*** Can express him/herself fluently and spontaneously, almost effortlessly, with a smooth flow of language. Can give clear, detailed descriptions of complex subjects. High degree of accuracy; errors are rare.	***Conveys finer shades of meaning precisely and naturally.*** Can express him/herself spontaneously and very fluently, interacting with ease and skill, and differentiating finer shades of meaning precisely. Can produce clear, smoothly-flowing, well-structured descriptions.

Table 7-7: The CEFR descriptor scales for linguistic competence used for the rating of LINDSEI-FR+

Along with the descriptor scale (Table 7-7), the raters were provided with the guidelines for rating and a rating grid in Excel format, in which they were required to record their scoring decisions. As set out in the **rating guidelines** provided to the raters, the rating procedure consisted in three successive steps:

1. The raters were invited to provide an analytic assessment of **five competences**, namely range, accuracy, fluency, phonological control[157], and coherence, and to assign a CEFR grade (A2, B1, B2, C1, or C2) to each of these five competences. Sublevels could also be distinguished by using + or - increments: "B2+" thus represents a stronger performance within B2, and "C1-" a weaker performance within C1.

2. Complementarily to the analytic assessment of five competences, the raters were invited to provide a holistic CEFR grade for **global assessment**. This global score is a composite score based on the overall impression of the proficiency displayed in each extract and on all the descriptors taken overall. Plus and minus increments could also be added to describe stronger or weaker performance within a band.

3. Additionally, the raters also had the opportunity to mark down personal **comments** on each learner's performance. These comments can pertain to any aspect of the learner's performance (i.e. not only fluency). Whilst one rater (R2) did not provide any such comments in his rating, R1 and R3 briefly commented on each learner.

The CEFR rating is based on **5-minute audio excerpts** from each learner interview. The corresponding transcripts were not provided to the raters. The audio excerpts were in the same order as in LINDSEI-FR, that is, not classified according to any feature. The 5-minute excerpts used for the rating were selected from the **free discussion task** (i.e. the second speaking task of the LINDSEI interviews) for two main reasons. Firstly, the first task (the set topic) was intended as a warm-up activity so that the learner would feel at ease with the interviewer and his/her environment. The set topic task was thus set aside for the CEFR rating to avoid start-up effects. In addition, not all set topic tasks in LINDSEI-FR+ last 5 minutes. Secondly, the last task (the picture description) prompts control over content and was consequently considered too constrained to adequately reflect the learner's fluency competence with a view to evaluating it according to the CEFR grids and descriptors. Furthermore, the third task is most generally far shorter than 5 minutes. The free discussion task thus appeared to lend itself well to our purposes from the point of view of type of speech (extemporaneous, dialogic, unconstrained) and duration. In Gilquin *et al*. (2010), the 5-minute excerpts were also taken from the free discussion task.

[157] The *Qualitative Aspects* scale contains descriptors for "interaction", which was substituted by "phonological control" in the adapted *Linguistic Competence* rating grid. With the benefit of hindsight, "interaction" could have provided valuable insights into dialogic aspects of fluency and is definitely worth re-integrating for the potential assessment of other components of LINDSEI.

I manually extracted the first 5 minutes of the free discussion task for each learner, with the exception of the five learners who had been evaluated in Gilquin *et al*. (2010). For these five learners, I used the same audio excerpt as in 2010. During the extraction, I paid particular attention to:

- the beginning and end of each excerpt: start and ending should occur at turn-taking places and not in the middle of an utterance, be it of the interviewer or the interviewee. The beginning of the excerpts generally coincides with the beginning of the free discussion task;
- the discourse uttered during the c. 5-minute extracts: the excerpt was aimed to stand on its own so that it could be understood by the raters without its surrounding context. For example, it can begin with a question by the interviewer and end after the interviewee's answer;
- bearing in mind that the selected speaking task is a free discussion, it may happen that the interviewer takes the foreground and holds the floor for very long turns. Particular attention was paid to avoid such cases as much as possible so that the learner remains the main speaker in the excerpt.

The following section presents the results of the rating by the three expert raters, examines the degree of agreement between the raters, and the reliability of the CEFR fluency ratings.

7.2.2 Fluency rating results and inter-rater reliability

The rating results were set side-by-side in an Excel file and the **CEFR fluency grades** were compared. The plus and minus increments were disregarded at this stage. I did not analyse the grades for the other four skills nor for global assessment (but see Section 7.4.3). Table 7-8 to Table 7-10 summarise the number of excerpts the three raters agreed or disagreed on with respect to fluency.

As can be observed in Table 7-8, **R1 and R2** reached an agreement on 23 excerpts (46%). The grade that led to the most cases of agreement is C1 (14 excerpts), followed by B2 (6 excerpts). The data reveal that the vast majority of disagreements concern the B2/C1 band scores (15 excerpts in total; 30%). Table 7-8 seems to indicate that R2 might be somewhat more generous in his grades than R1 because, for the majority of the disagreements, R2 provided a higher score than R1. It is also noteworthy that the raters disagreed by two band scores (B2 for R1, C2 for R2) for four learners, namely FR015, FR018, FR027 and FR036. In three of these cases, R1 explains his lower rating by commenting that the learner searches for words or has problems expressing finer shades of meaning.

R1 and R3 agreed on the CEFR fluency score of 26 learners (52% - see Table 7-9). Unlike for R1 and R2, where most agreements pertained to the C1 level, most agreements between R1 and R3 are on the B2 level (16 agreements). There were only 8 agreements on the C1 level. Five disagreements by two band scores fell on the distinction between B1 and C1 and one on the distinction between B2 and C2. The comments by the two raters provide some clues as to the reason for these differences: FR001, for instance, was marked down by the third rater because she used "*some L1 lexical items*" while she was given a C1 by R1 for expressing herself "*fluently and spontaneously with little obvious effort*". It is also interesting to see that, although raters may disagree by two band scores, their comments on the same performance may be very similar. In his comment on FR050, R1 (who gave a C1) wrote the following: "*Some searching for lexical and grammatical resource to express intended meaning but overall, clear and reasonably fluent*". R3 provided a very similar comment ("*Reasonably fluent but with some hesitation around unknown lexis on a common topic*"), although he evaluated the learner down at B1 level for fluency.

As for **R2 and R3**, given the fact that R2 seems to be the most "generous" of the three raters and R3 the "strictest", it is not surprising to see that they only agreed on 13 excerpts (26%) and disagreed on the remaining 37 (74%). An equal number of agreements was reached for levels B2 and C1 (6 agreements each). There is one case in which the two raters disagreed by three band scores: whereas R2 rated FR025 at C2 level for fluency, for R3, it was only B1. The rater justified this poor evaluation in his comment by explaining that he found the learner's speech to be "*a bit slow*" and "*unnatural*". Note that R1 and R3 also disagreed by two band scores on the same learner – R1 commented that the learner, although he expresses himself fluently, does not appear to always do so "*effortlessly*".

Taken together, the examination of the data from the three raters shows that there are **10 perfect agreements** (5%): 4 on B2, 5 on C1, and 1 on C2.

An excerpt and short comment on FR002, for whom there was disagreement between the three raters, is provided below (Example 7-1).

		R2				*Totals*
		B1	B2	C1	C2	
R1	B1	**0**	1	0	0	
	B2	0	**6**	11	4	
	C1	0	4	**14**	6	
	C2	0	0	1	**3**	
Totals						***23***

Table 7-8: R1 and R2 assessed CEFR fluency scores

		R3				*Totals*
		B1	B2	C1	C2	
R1	B1	**1**	0	0	0	
	B2	5	**16**	0	0	
	C1	5	11	**8**	0	
	C2	0	1	2	**1**	
Totals						***26***

Table 7-9: R1 and R3 assessed CEFR fluency scores

		R3				Totals
		B1	B2	C1	C2	
R2	B1	**0**	0	0	0	
	B2	4	**6**	1	0	
	C1	6	14	**6**	0	
	C2	1	8	3	**1**	
Totals						***13***

Table 7-10: R2 and R3 assessed CEFR fluency scores

7-1: FR002-F – disagreement between the three raters with respect to CEFR fluency level

<B> because er (0.260) the carnival of Binche is quite eh (0.850) odd for eh foreigners er </B>

<A> mm yes I've seen some photos extraordinary <overlap /> er (0.110) things </A>

<B> <overlap /> **with eh their (0.170) with their hats** </B>

<A> yeah </A>

<B> and their er (0.900) **hea= heav=** (0.220) er **feathers** </B>

<A> yeah </A>

<B> **their feathers** (0.440) er (0.680) **os= ostrich** it's er **ostrich** (0.210) feathers (0.640) and er it's very beautiful to see (0.450) and er also their costumes (0.450) eh they are dressed in er special costumes (0.280) and they er (0.760) they have em (1.420) special stuff within it (0.800) </B>

<A> (0.800) in wha= inside the costume </A>

<B> inside the costume (0.320) and er they burn it (0.220) er on the last day of eh the carnival </B>

<A> ah </A>

<B> and it's also a: (0.420) a great ritual if you: (0.780) if you understand what I mean (0.360) it's er (0.650) they are (0.250) they form circle (0.300) and er they d= (0.310) they dance around the fire </B>

<A> are these the: the gilles de Binche or <overlap /> or whatever </A>

<B> <overlap /> yes yes it's the gilles de Binche (0.470) and er there's also er (0.100) fireworks (0.450) and er (0.370) it's an occasion to: (0.460) to eat (0.380) er <laughs> chips </B>

[…]

<A> and what sort of music is it at these occasions </A>

<B> is it er it's traditional music with er (0.940) drums and er with a: (1.990) **grosse caisse** (1.310) er </B>

<A> grosse caisse (0.850) what **what is a grosse caisse** (1.000) </A>

<B> (1.000) it's a: (0.230) a round instrument with er (1.200) and you er bat= (0.250) beat against it </B>

<A> it's not a drum though (0.500) it's a sort of drum (0.280) </A>

<B> (0.280) it's a sort of drum <overlap /> but er </B>

<A> <overlap /> is is it a big drum </A>

<B> it's big= it's bigger and it <overlap /> and the sound is </B>

<A> <overlap /> oh it's what they call a bass drum I think I think it's called a (0.150) like they play in military bands the person in front has a big drum and (0.340) is it like that I think it's a bass drum </A>

The transcript shows several types of struggles the learner had to cope with (see bold font). The learner is talking about the carnival of Binche and tries to explain how the *gilles*[158] are dressed. She first seems to hesitate over the word *hat*, and then stumbles on the words *heavy*, *feathers* and *ostrich* (which she also mispronounces). She goes on explaining the way the gilles are dressed, but the interviewer seems to have problems understanding her explanations. The topic moves on to the music played during the carnival and FR004 uses the French word *grosse caisse*. The interviewer and the learner spend some time to identify its English equivalent (*bass drum*).

In addition to lexical issues, the learner's speech is quite tainted by a French accent and her rate of speech is rather slow.

FR002 was given a B2 by R1, a C1 by R2, and a B1 by R3. It may have been that, while some raters graded the performance down because of the numerous lexical issues, truncations or reformulations, R2 might have taken other aspects into account, such as the fact that the learner tries to cope with her lexical issues by using a French equivalent and defining it.

To say the least, these first results are not very encouraging: agreement between raters on the CEFR fluency level (without taking the + and - increments into account) is reached in, at most, just over half the cases. However, they confirm the results from previous studies where a high lack of agreement among native speaker raters has also been reported (e.g. Chambers 1997; Schmitt 2000). The **evaluation of learners' fluency**, it seems, is **not an easy task, even for professional raters**[159]. Slightly more reassuring perhaps is that, even if there is quite a high number of disagreements, few of them are disagreements by two band scores or more.

[158] The gilles are a group of Binche's inhabitants clad in traditional costumes for the carnival. They are characterised by their colourful dress, wax mask, wooden footwear and hats adorned with large ostrich feathers.

[159] While this thesis focuses on expert rating, Gilquin *et al.* (2016) conducted a rating experiment with 27 seasoned Belgian French-speaking secondary school teachers of English as a foreign language. For this experiment, they adopted the same rating procedure as was described above, but used only ten of the 5-minute LINDSEI-FR excerpts used in this thesis. Gilquin *et al.* (*ibid.*) also made use of the CEFR ratings of R1 and R2. Their study revealed large discrepancies in the grades assigned by the non-native teachers, "sometimes covering the whole spectrum from A2 to C2 for one and the same sample", which seems to corroborate Isaacs & Thomson's (2013) finding that expert raters (R1 and R2) can achieve greater consensus than more novice raters (i.e. the teachers). Gilquin *et al.* (*ibid.*) also confirmed the well-documented tendency for non-native speakers to be generally less tolerant than native speakers (Fayer & Krasinski 1987; Jo 2015; Koster & Koet 1993; see also Winke, Gass & Myford 2013). The teachers, indeed, almost systematically assigned lower grades than the native experts R1 and R2.

This might suggest that it is perhaps the **boundary between adjacent levels** that raises issues rather than an intrinsic mis-interpretation of the descriptors.

In an attempt to get deeper insights into the rating results, the **inter-rater reliability** was also assessed. Inter-rater reliability aims to quantify the **degree of covariance** between individuals. Two (or more) variables are said to covary when "changes in one variable [are] met with similar changes in the other variable[s]" (Field 2013:264). In the assessment literature, **Cronbach's alpha** is typically used to assess the reliability of the data from more than two raters and the degree of agreement between pairs of individuals is usually expressed by means of a correlation score (**Pearson's product-moment** *r* score).

In preparation for the reliability tests, the CEFR fluency grades were converted into **numerical values** using a ten-point numerical scale, as presented in Table 7-11 (adapted from Thewissen 2012). The scale ranges from B1 (= 1) to C2 (= 10), and each increment corresponds to one additional point (B1+ = 2; B2- = 3 etc.). Learner FR011, for example (Table 7-12), was attributed a B2 grade by R1, a C1 by R2 and a B2- by R3. Following the numerical scale, these grades were converted into a 4, 7 and 3, respectively.

CEFR grade	**B1**	**B1+**	**B2-**	**B2**	**B2+**	**C1-**	**C1**	**C1+**	**C2-**	**C2**
Value	1	2	3	4	5	6	7	8	9	10

Table 7-11: The 10-point numerical scale used to calculate the CEFR fluency score

Learner	**Rater**	**CEFR fluency grade**	**CEFR fluency score**[160]
FR011	R1	B2	4
	R2	C1	7
	R3	B2-	3

Table 7-12: Using the 10-point numerical scale - example

For learner FR036, R1 attributed a double grade [161] ("B2/C1"), presumably because he hesitated between the two grades (note that, for some reason, the rater did not choose to use + and - increments). The two grades were converted into numerical values (4 and 7, respectively), and a mean score was calculated (5.5).

[160] "CEFR fluency grades" refer to the CEFR fluency levels, i.e. B2, C1, C2 etc. "CEFR fluency scores" correspond to the numerical equivalent for the grades.

[161] Note that while R1 used only one double grade in the fluency scale, he used them repeatedly with the other skills and the global proficiency grades.

The first step of the **inter-rater reliability analysis** aims to assess the reliability of the data from the three professionally trained raters as a group. Values of **Cronbach's alpha** were computed at **0.69**, which represents an **acceptable level** of inter-rater reliability (Tavakol & Dennick 2011). Although α value for the three raters is slightly lower than values reported elsewhere for read speech (e.g. Cucchiarini, Strik & Boves 2002; Derwing *et al.* 2004), it is **very similar to values reported for analogous tasks of spontaneous speech** (around 0.68) in Riggenbach (1991) or Freed (1995), for instance.

Given these results, the CEFR fluency scores from the three raters are thus deemed reliable enough for use in further analyses.

Going further into the relationship between the scores of the three raters, the **degree of agreement between pairs of raters** can be assessed by means of a **Pearson's product-moment** correlation score *r* that ranges from 0 to 1. A coefficient +1 indicates a perfect positive relationship – as one variable increases, so does the other by a proportionate amount – and a coefficient of -1 indicates a perfect negative relationship – as one variable increases, the other decreases by a proportionate amount. A coefficient of 0 indicates that there is no linear relationship between the variables. Alderson *et al.* (2001:132) advise researchers to aim for a $r = 0.8$ score, while Jarvis (2002) points out that there are three levels of reliability, namely "moderate" ($0.5 \leq r < 0.7$), "substantial" (from 0.7 to 0.9), and "complete" agreement (from 0.9 onwards).

Figure 7-2, Figure 7-3 and Figure 7-4 neatly illustrate the correlations between the three pairs of raters for the fluency scores. For example (see Figure 7-3), the excerpts evaluated at C1 level (i.e. a score of 7) by R1 may correspond to a score 1 (B1), 2 (B1+), 3 (B2-), 4 (B2), 5 (B2+), 6 (C1-) or 7 (C1) according to R3's standards. Likewise, a score of 4 (B2) by R2 may correspond to a 1 (B1), 2 (B1+) or 4 (B2) by R3's standards (see Figure 7-4).

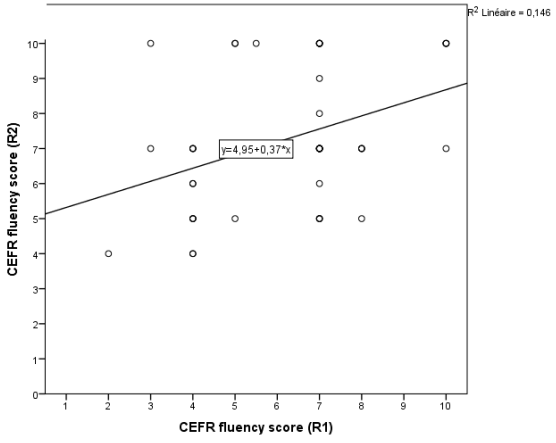


Figure 7-2: Correlation between R1 and R2 CEFR fluency scores

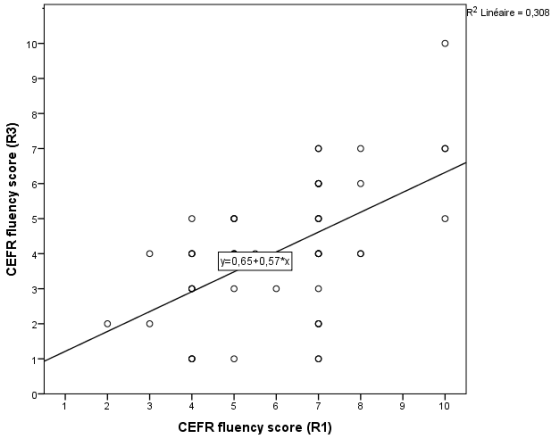


Figure 7-3: Correlation between R1 and R3 CEFR fluency scores

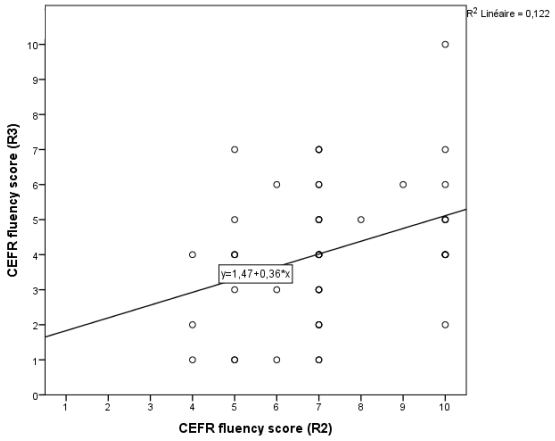


Figure 7-4: Correlation between R2 and R3 CEFR fluency scores

A two-tailed bivariate Pearson's product-moment **correlation test** revealed that there are **significant** positive correlations between the scores attributed by pairs of raters (Table 7-13). The highest correlation is found between R1 and R3, with $r = .56$ ($p < .000$), which represents a **moderate** relationship (Jarvis 2002). The correlations between R1 and R2 and between R2 and R3 are slightly lower, with $r = .38$ ($p < .01$) and $r = .35$ ($p < .02$), respectively.

	R1	**R2**	**R3**
R1		.382	.555
R2			.349

Table 7-13: Pearson's correlation coefficients between pairs of raters

In addition to the widely recognised intrinsic difficulty of evaluating volatile spoken data, a number of factors may partly account for the acceptable but moderate values of α and *r*. These factors are outlined below.

A. <u>The professional experience of the raters</u>

Although the three raters have some experience in using the CEFR, they were originally trained to use the Cambridge descriptor scales. As argued in the literature (e.g. Chalhoub-Deville 1995; Shaw 2004), raters **may not always confine to the descriptor scales provided** but rather rely on their knowledge of and experience in using other scales. It is thus difficult to assess the extent to which they successfully distanced themselves from the Cambridge descriptors and actually applied those of the CEFR.

In a personal communication, however, R1 insisted on the power and "liberating effect [of the CEFR scales] on the way [raters] assess users across levels". He added: "I recall the days when, as examiners, we couldn't recognise what teachers have recognised for years: most users of an interlanguage have a profile which ranges across levels and a B2 or C1 is very likely to have elements of levels below and above if the descriptors really capture different aspects of

performance". R2, for his part, is a specialist of speech evaluation, but has little experience with the CEFR descriptors: he may have had more difficulties in distancing himself from his better knowledge of the Cambridge descriptor scales. R3, by contrast, has a solid experience in speech assessment according to both the CEFR and the Cambridge descriptor scales. The different professional experience of the raters might thus have affected their rating to some extent.

B. The methodology adopted for the rating

The raters were asked to assess **audio-recorded speech**. It was thus impossible for them to engage in the conversation or even to watch the learner they had to evaluate[162]. These two impediments may have hindered their evaluation to some extent. Additionally, the guidelines provided to the raters did not specify the maximum number of times they could listen to the excerpts, nor did they mention that the excerpts had to be listened in full. These considerations were left to the raters' choice. Although it is unlikely that the raters listened several times to all 50 excerpts (which already amount to more than 4 hours of recording), it is not impossible that they felt that the evaluation of some learners was easier and did not listen until the end of the excerpt, hence running the risk of missing some precious elements.

The rating guidelines did not specify that the raters had to **assess the five skills and global competence in a specific order**. It may well be that one rater first gave grades for the five competences and then, based on these five grades, attributed a global proficiency grade, while another might have started with the more general picture of the learner's proficiency, before differentiating subtler aspects.

It is also worth underlining that the raters were not given the opportunity to have a **training session**, nor to discuss disagreement cases prior to the full rating. These factors might also account for the moderate correlations between the raters.

C. The nature of the rated excerpts

In spite of the fact that the raters were familiar with the assessment of spoken data, they may have more or less experience in assessing different types of speaking tasks. For the assessment of the LINDSEI-FR+ learners, the excerpts contained spontaneous, unplanned dialogues and may have thus included some features the raters were not entirely familiar with, such as overlapping speech or noise. Such **free and unconstrained dialogues** are likely to raise more difficulties in rating than read or monologic speech. In addition, although precise criteria were adopted to select the rated samples, excerpts greatly vary, for example, with respect to **topic** (holidays vs. plans for the future), **formality** (some learners laughed and joked with the interviewer), or speech rate (the number of words uttered by the learner was

[162] Body language, gestures, or mimics undoubtedly play a role in communication as well.

not controlled). Undoubtedly, those factors might have exercised some influence on the perception and evaluation of the learners' productions.

D. The nature and properties of the rating scale

As explained in the introductory section of this chapter, criticism has been levelled at the CEFR descriptors (e.g. Alderson 2007; Hulstijn 2007). The monologic bias of the fluency scale, for example, might have affected the raters to some extent.

E. The number of raters

While some studies (e.g. Bosker *et al.* 2013; Cucchiarini, Strik & Boves 2000; Cucchiarini, Strik & Boves 2002; Derwing *et al.* 2004; Kormos & Dénes 2004; Préfontaine & Kormos 2015; Rossiter 2009) reported using between 6 and 30 raters (sometimes even more), the present study has only three. From a statistical point of view, the degree of agreement between raters tends to increase as the number of observers goes up.

F. (Dis)fluency profiles and rater profiles

Götz (2013a) has shown that several fluency profiles can be distinguished among learners, i.e. they use different combinations of (dis)fluency features. It may be that different raters do not assess the same profile in a similar way because they themselves might be more tolerant or, on the contrary, more critical, of one or another feature typical of the profile.

All these reasons might explain why inter-rater reliability is somewhat surprisingly low. They might be worth bearing in mind if further fluency ratings should be performed.

The following section presents the methodology adopted for the measurement of the final CEFR fluency score of each learner.

7.2.3 Assigning a final CEFR score

Given that the inter-rater reliability analysis indicated that the CEFR fluency scores are sufficiently reliable and that all the correlations between pairs of raters are significant, the three raters' CEFR fluency scores were pooled and **mean CEFR fluency scores** were computed per learner.

A concrete example is provided for excerpt FR007 and FR030 in Table 7-14. R1 rated FR007 at C1 level (i.e. 7 on the scale, which is reprinted in Table 7-15 for clarity), R2 gave C1+ (which corresponds to an 8), and R3 a B2+ (i.e. 5). A mean score of 6.67 was calculated for this learner. FR009, who obtained a C1-, a C1 and a B2- (6, 7 and 3, respectively), receives a mean score of 5.33. The mean fluency scores of the learners in LINDSEI-FR+ range between 2.67 and 10, with a **mean of 5.8**.

Learner	Rater	Fluency grade	Fluency score	Mean fluency score
FR007	R1	C1	7	6.67
	R2	C1+	8	
	R3	B2+	5	
FR030	R1	C1-	6	5.33
	R2	C1	7	
	R3	B2-	3	

Table 7-14: Assigning a mean fluency score (examples)

CEFR grade	B1	B1+	B2-	B2	B2+	C1-	C1	C1+	C2-	C2
Value	1	2	3	4	5	6	7	8	9	10

Table 7-15: The 10-point numerical scale used to calculate the CEFR fluency score

It is also possible to convert the mean scores back to CEFR grades. The **interpretation of the scores** (adapted from Thewissen 2012), is presented in Table 7-16. Scores from 1 to 2.4 fall into the B1 band; from 2.5 to 5.4 they fall into the B2 band; from 5.5 to 8.4 they fall into the C1 band; and scores over 8.5 fall into the C2 band.

Mean CEFR fluency score	CEFR fluency grade	CEFR fluency level
1 to 1.4	B1	B1
1.5 to 2.4	B1+	
2.5 to 3.4	B2-	B2
3.5 to 4.4	B2	
4.5 to 5.4	B2+	
5.5 to 6.4	C1-	C1
6.5 to 7.4	C1	
7.5 to 8.4	C1+	
8.5 to 9.4	C2-	C2
9.5 to 10	C2	

Table 7-16: The interpretation of the final fluency score

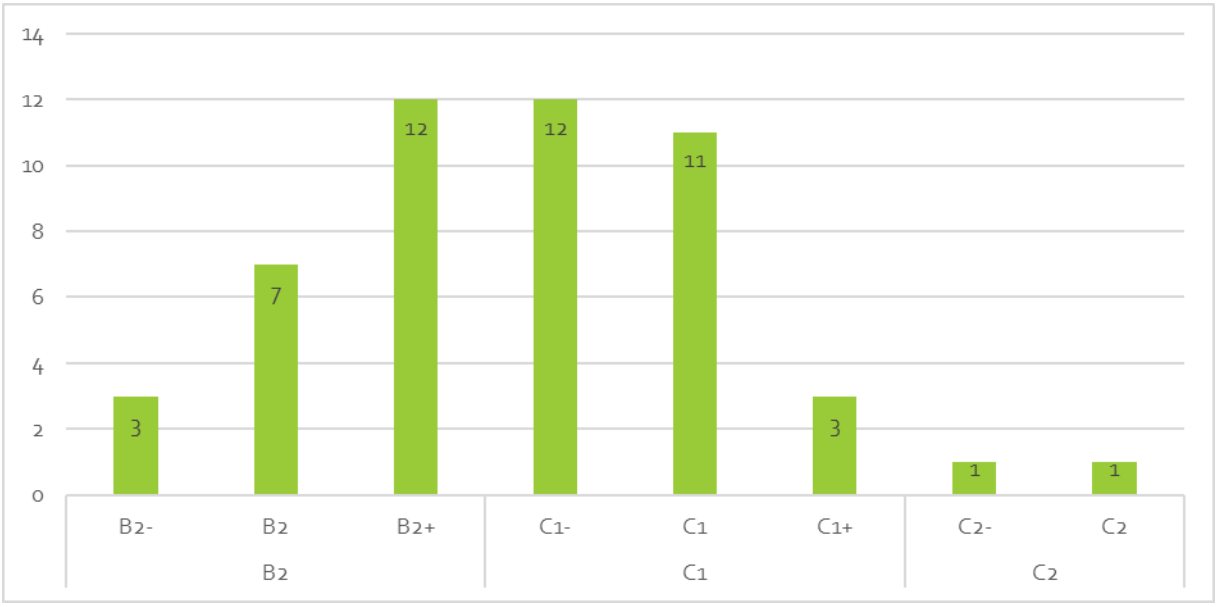


Figure 7-5: CEFR fluency levels in LINDSEI-FR+

Figure 7-5 displays the CEFR fluency grade (with and without increments) of the learners in LINDSEI-FR+. The learners mainly belong to the **B2** (*n* = 22) **and C1** (*n* = 26) levels for fluency, and two learners are evaluated at **C2** level. While the learners' proficiency was labelled as "advanced" on the basis of an external criterion (they were all third- or fourth-year students of English at university level) (*cf.* Gilquin, De Cock & Granger 2010), it appears that, from the point of view of fluency, their CEFR level is rather **upper intermediate or lower advanced**.

The following excerpt (Example 7-2) comes from one of the learners (FR004) whose mean fluency score was very high (9; i.e. C2). It is interesting to see that, although her fluency level is very high, this learner does use quite a lot of restarts (*he said he told me*; *it w= there was*; *the next Sep= the following September* etc.). She also seems to repeat the personal pronoun *I* quite regularly and the interviewer even helps her find the word "places", which is arguably not a very complex word. However, it is also worth noticing that she uses the discourse marker *well* appropriately, uses a phrasal verb (*put me off*) as well as multiword units such as *quite little chance for me.* The two raters seem to have marked up the use of colloquialisms and idiomatic expressions in FR004's speech. R1, for example, wrote: "Not always wide range of academic language but very colloquial, very fluent, pronunciation enhances intended message"; likewise, R3 indicated: "Excellent control of language with idiom[a]tic expressions and use of colloquialisms".

7-2: FR004-F

<B> yeah **I I** went to see (erm) somebody at the: . Polytechnic in Nottingham and (er) **he said he told** me that **it w= there was** quite little chance for me to get in because there were only three hundred . </B>

<A> places </A>

<B> yeah places for next it was in September so it was for next Sep= **the** <overlap /> **following September** </B>

<A> <overlap /> for one year yeah </A>

<B> and it was already full and he said he would contact me again and **he never he never** did so **I I** don't know **I I .. I** thought it was quite so I said **well** . I'll go back to Belgium and then I'll see </B>

[…]

<A> and in England . <overlap /> there's a </A>

<B> <overlap /> he really **put me off** </B>

<A> a fixed number of places and (er) </A>

<B> (mm) </B>

Excerpt 7-3 comes from the learner who has the lowest fluency score, FR006 (2.67; B2). As shown in bold front, the speech of FR006 is interspersed with quite a lot of repetitions, as well as filled and (generally long) unfilled pauses. Contrary to FR004, she does not use many colloquialisms (R1 comments that she has "sufficient" language on "straightforward topics"). R3 seems to have marked down this learner because she is "reluctant" to speak, which, he writes, "hinders communication and pronunciation".

7-3: FR006-F

<B> yes I think (0.190) er (0.960) **I don't** (0.530) **I I don't** come here er (1.010) for parties or something like that I prefer **to go to go** and visit friends and (0.670) I don't like **to: to to stay** (1.390) **to stay** up all night or something like that </B>

<A> yeah </A>

<B> er (0.840) I have a social life but not a night life if you (0.360) can </B>

[…]

<B> yes **I I I** would like **to: to** try but (0.260) erm (0.370) my parents told me **y= y= you** can choose (0.600) er either a student room or (0.460) **a a** car and I took the car and now it's finished (0.310) I've the car **for** (1.920) **for** (3.830) er the length of **my stud= my studies** </B>

In what follows, I will refer to B2, C1, and C2 as the *CEFR fluency*[163] *grades* or *CEFR fluency levels*. The numerical values associated with those levels will be referred to as the *CEFR fluency scores*. *CEFR fluency rating* and *rating results* are used as more generic terms.

[163] Following the official name of the CEFR scale used for the rating, I use the term *fluency* (and not *(dis)fluency*) when referring to the CEFR grades and scores.

7.3 THE RATED EXCERPT, THE FREE DISCUSSION TASK, AND THE INTERVIEW

As has been made clear above, only a 5-minute excerpt from the free discussion of each learner was evaluated according to the CEFR. While the practice of using short speech samples for rating purposes is common in L2 research (Isaacs & Thomson 2013), it is less clear from the literature **whether the rating results can reliably be generalised to a larger portion of the learner's speech**. In other words, can the rating results be reliably generalised to the complete free discussion task, or even to the whole interview? To my knowledge, the question of the reliability of the generalisation of rating results has never been thoroughly addressed before. While acknowledging that this issue is complex and multifaceted, I would like to make a first contribution in this direction.

As a preliminary step in the analysis of the relationship between (dis)fluency ratings and productive (dis)fluency measures, this section therefore seeks to determine whether the data in the rated excerpt is representative of the data in the complete free discussion task and in the whole interview. The underlying assumption is that, if there are no significant differences between (dis)fluency measures calculated from the rated excerpt, the free discussion, and the interview, then the rating results can quite reliably be assumed to be generalisable.

It is important to underline that the results of this analysis may have major consequences on the analyses described in Section 7.4. **Two case scenarios** may happen. In the worst case, (dis)fluency measures will significantly differ between the rated excerpt and the complete free discussion and/or the whole interview. Following the aforementioned hypothesis, the **rating results should thus not be generalised** either to the complete free discussion or the whole interview. In addition, the analyses examining the relationship between CEFR fluency ratings and (dis)fluency measures should use the measures calculated on the basis of the data in the rated excerpt only. If, however, it appears that the (dis)fluency measures calculated from the rated excerpt are *not* significantly different from the same measures calculated on the basis of the data contained in the complete free discussion task or the whole interview, then the CEFR fluency ratings could reliably be **generalised to a larger part** of the learner production. In the best-case scenario, CEFR fluency ratings could be generalised to the interview, and links could also be made between the results of the Principal Components Analysis or the Cluster analysis (*cf.* Chapter 6) and CEFR grades and scores.

To determine whether (dis)fluency measures differ significantly between the rated excerpt, the complete free discussion task, and the whole interview, a series of **repeated-measures one-way ANOVAs** were conducted. Repeated-measures are used to account for the fact that the data from the three datasets (the rated excerpt, the free discussion and the interview) come from the same speakers. A summary of the fourteen ANOVAs (i.e. one ANOVA per (dis)fluency measure) can be found in Table 7-17.

(Dis)fluency variables	Sphericity	*F* / Greenhouse-Geisser
Conjunctions	0.00	F = 5.899 (*p* = .010; η^2 = .107)
Discourse markers	> .05	*F = 3.930 (p = .023; η^2 = .074)*
False starts	0.00	n.s.
Filled pauses	0.03	n.s.
Foreign words	0.04	n.s.
Lengthenings	>.05	*n.s.*
Mean length of runs	0.01	F = 7.207 (*p* = .002; η^2 = .128)
Mean length of unfilled pauses	0.00	F = 3.684 (*p* = .040; η^2 = .070)
Phonation-time ratio	0.00	F = 4.870 (*p* = .020; η^2 = .090)
Restarts	0.00	n.s.
Repetitions	>.05	*n.s.*
Speech rate	0.00	F = 30.937 (*p* = .000; η^2 = .387)
Truncations	>.05	*n.s.*
Unfilled pauses	0.00	F = 12.591 (*p* = .000; η^2 = .204)

Table 7-17: Results of repeated-measures ANOVAs (rated excerpt, free discussion, and interview) in LINDSEI-FR+
Note: The condition of sphericity was met for DM, L, Rep, and T; the F values for those variables are shown in italics. The condition of sphericity was not met for the other variables; the Greenhouse-Geisser correction was applied in those cases.

Table 7-17 reveals that, while the frequency of false starts, filled pauses, foreign words, lengthenings, restarts, repetitions and truncations is not significantly different in the rated excerpt, the free discussion task and the interview, the other **seven (dis)fluency measures do differ significantly**. Among these seven variables are the measures of temporal (dis)fluency, conjunctions, and discourse markers.

Although these results partly match the "worst-case scenario" described above, two other elements need to be taken into consideration. First, the examination of **effect sizes** reveals that the size of the differences is generally **(very) small**, except for speech rate, where the coefficient (η^2 = .387) represents a medium effect. Second, ANOVA is an omnibus test, which means that it tests whether the means in the three conditions are equal or not, but it does not provide specific information about where the differences might lie. It is therefore necessary, after conducting an ANOVA, to carry out **post-hoc tests** to find out which conditions differ.

Table 7-18 summarises the results of **pairwise comparisons** (a more detailed overview of the post-hoc results can be found in Appendix 9.9). As can be seen, the rated excerpt and the complete free discussion differ by two (dis)fluency variables, namely mean length of runs and unfilled pauses (the runs are shorter and the pauses less frequent in the free discussion task). It is perhaps more (positively) surprising to see that the rated excerpt and the whole interview

significantly differ by only one (dis)fluency variable, namely speech rate. The mean speech rate in the interview is, in fact, lower than in the rated excerpt (162.6 vs. 168.6 words per minute). This is probably due to the fact that the interview also contains monologic speech (especially the picture description), which can be associated with lower speech rates (Ejzenberg 2000; Riggenbach 1989; Tavakoli 2016). The rated excerpt and the whole interview, however, do not differ significantly when it comes to the other 13 (dis)fluency variables. Overall, the results suggest that, with one exception, **the data in the rated excerpts are representative of the data in the whole interviews**. Consequently, following our hypothesis, the **CEFR fluency ratings can reliably be generalised** to the whole learner production, i.e. the whole interview.

	Free discussion	**Interview**
Excerpt	MLR UP	SR
Free discussion	/	C MLR PTR SR UP

Table 7-18: ANOVA post-hoc tests with Bonferroni correction[164]

In the following section, CEFR fluency ratings will thus be related to (dis)fluency variables as measured in the whole interview (i.e. the same measures as in Chapter 6). Although it has been shown that the mean speech rate significantly differs in the rated excerpt and in the whole interview, the gains from using (dis)fluency variables as measured in the interview are greater. First, it increases the overall homogeneity of the analyses throughout this dissertation. Second, it makes it possible to further exploit the results described in Chapter 6, more particularly by examining the nature of the relationship between underlying dimensions of (dis)fluency and CEFR fluency ratings (what is the nature of the relationship between (dis)fluency components and CEFR fluency ratings?) as well as between (dis)fluency profiles (clusters) and CEFR fluency ratings (can some profiles be associated with higher ratings?). Results for speech rate will, of course, need to be interpreted somewhat more cautiously.

[164] Although the ANOVA returned significant results for discourse markers and mean length of unfilled pauses, the differences in pairwise comparisons are very small and do not reach significance.

7.4 RELATING THE CEFR FLUENCY RATINGS AND LEARNER LANGUAGE

This section seeks to investigate the nature and extent of the relationship between French-speaking learners' CEFR fluency ratings (*cf*. Section 7.2) and their productive (dis)fluency (as measured in the whole interview, see Section 7.3). I first focus on the relationship between CEFR fluency ratings and individual (dis)fluency measures and (dis)fluency components (as identified in Section 6.1.1). Then, the relationship between CEFR fluency ratings and (dis)fluency clusters (*cf*. Section 6.2) is examined. Finally, the relationship between CEFR fluency ratings and other CEFR ratings (e.g. pronunciation and accuracy) is analysed.

7.4.1 CEFR fluency ratings, (dis)fluency variables and (dis)fluency components

7.4.1.1 Correlations between CEFR fluency scores and (dis)fluency measures

A Pearson's bivariate **correlation analysis** was run to examine the nature of the relationship between the CEFR fluency ratings and the 14 (dis)fluency measures on the one hand and the 5 (dis)fluency components on the other. As a reminder, the correlation coefficient *r* ranges between 0 and 1 (the closer to 1, the stronger the correlation), and may be positive or negative (*cf*. Section 7.2.2).

The results of the correlational analysis between the CEFR fluency ratings and the 14 (dis)fluency measures are displayed in Table 7-19. The results of the correlational analysis between CEFR fluency ratings and (dis)fluency component scores are shown in Table 7-20.

	Pearson's *r*	***p***
Conjunctions	-.136	.348
Discourse markers	**.332**	**.018**
False starts	-.062	.670
Filled pauses	-.234	.102
Foreign words	-.173	.230
Lengthenings	-.008	.959
Mean length of runs	.262	.066
Mean UP length[165]	-.272	.056

[165] The correlations between CEFR fluency ratings and *mean UP length* and *mean length of runs* is, however, significant in the rated excerpt (r = -.330; p = .019 and r = .295; p = .038, respectively). Although there was no significant difference in the post-hoc tests in Section 7.3 for those variables, the slight difference seems to have affected the strength of the correlations nonetheless.

Phonation-time ratio	**.320**	**.024**
Repetitions	-.106	.465
Restarts	**-.317**	**.025**
Speech rate[166]	**,433**	**.002**
Truncations	-.262	.066
Unfilled pauses	**-,358**	**.011**

Table 7-19: Pearson's correlations between (dis)fluency measures and CEFR fluency ratings

	Pearson's *r*	***p***
Comp. 1 – temporal (dis)fluency	**,368**	**.009**
Comp. 2 – repair (dis)fluency	-.256	.072
Comp. 3 – pragmatic (dis)fluency	**,351**	**.012**
Comp. 4 – cohesion	-.114	.430
Comp. 5 – lexico-grammatical (dis)fluency	-.160	.268

Table 7-20: Pearson's correlations between (dis)fluency component scores and CEFR fluency ratings

It is striking from Table 7-19 that only **five (dis)fluency variables** out of the 14 are significantly correlated with CEFR fluency ratings. The strongest correlation pertains to speech rate, followed by the frequency of unfilled pauses, of discourse markers, the phonation-time ratio, and the frequency of restarts. All significant correlations represent **medium-sized effects** (*r* > .3). Unsurprisingly, whilst two correlations are negative (the fewer the unfilled pauses or the restarts, the higher the rating), the other three correlations are positive (the higher the speech rate, the phonation-time ratio or the discourse markers, the higher the fluency rating). A visual representation of the five significant correlations is provided in Figure 7-6 to Figure 7-10.

Given that three variables contributing to the **temporal dimension of (dis)fluency** are significant, it could be predicted that Component 1 significantly correlates with CEFR fluency ratings too. The temporal component and CEFR ratings do correlate, as shown in Table 7-20. Again, the size of the correlation represents a medium-size effect (*r* = .368). **Component 3** also significantly correlates with CEFR fluency ratings, which can be explained by the fact that two of its constituent variables (discourse markers and speech rate) are also significantly correlated with the ratings. Although restarts are correlated with (dis)fluency ratings, the repair component (Component 2) is not. This is probably due to the fact that this component

[166] The correlation between the speech rate in the rated excerpt and the CEFR fluency ratings is slightly stronger (*r* = .511; *p* = .000).

also contains other variables besides restarts, namely repetitions and truncations, which are not significantly correlated with CEFR fluency ratings.

These results **corroborate previous findings** that **temporal features** of learner speech such as speech rate and unfilled pauses are indeed correlated with perceived (dis)fluency level (Derwing *et al.* 2004; 2009; Riggenbach 1991; Rossiter 2009). However, my results diverge from those reported by Gilquin *et al*. (2016). In their analysis of EFL teachers' ratings, Gilquin *et al*. did not find any relationship between fluency ratings and the frequency of unfilled pauses. This might be due to the fact that the authors used the rate of transcribed unfilled pauses (i.e. unfilled pauses in the non-aligned version of LINDSEI-FR), or to the different types of raters used (EFL teachers vs. native speaker raters). Alternatively, the diverging results might partly be due to the more limited number of learner samples used (5 learners in Gilquin *et al*. (2016) vs. 50 in this thesis).

Moreover, my results do not confirm that the **length of runs** is a primary factor correlating with (dis)fluency ratings, as was found in Kormos and Dénes (2004) and Préfonaine *et al.* (2015) (see, however, footnote 165), and they do not support the predominant status given to length of runs in the **CEFR descriptor scales**.

With respect to **repair (dis)fluency**, the statistical analysis confirms the weak relationship with perceived (dis)fluency reported in the literature: restarts are moderately correlated with CEFR ratings, but the other variables contributing to the repair (dis)fluency component, as well as the repair component itself (Component 2), do not. Incidentally, this might explain the absence of relationship found by Cucchiarini *et al.* (2002) between perceived (dis)fluency and *number of disfluencies* (which also included repetitions and corrections).

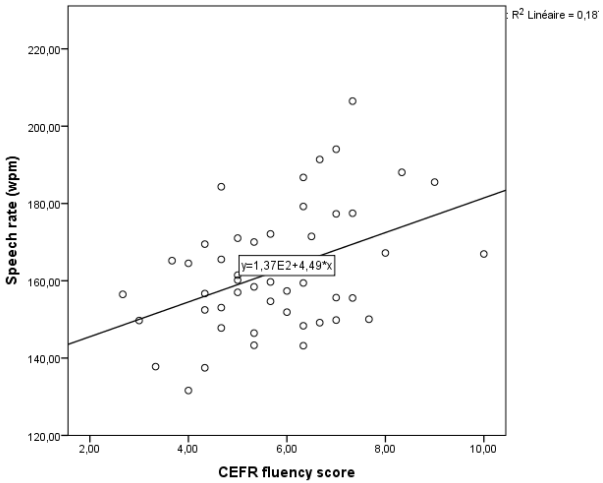


Figure 7-6: The relationship between speech rate and CEFR fluency score

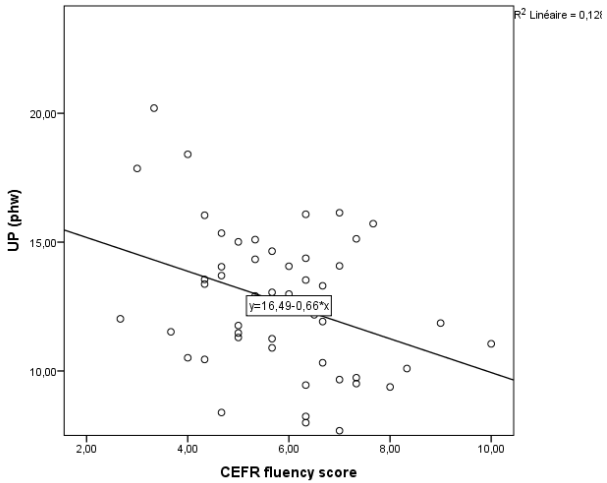


Figure 7-7: The relationship between unfilled pauses and CEFR fluency score

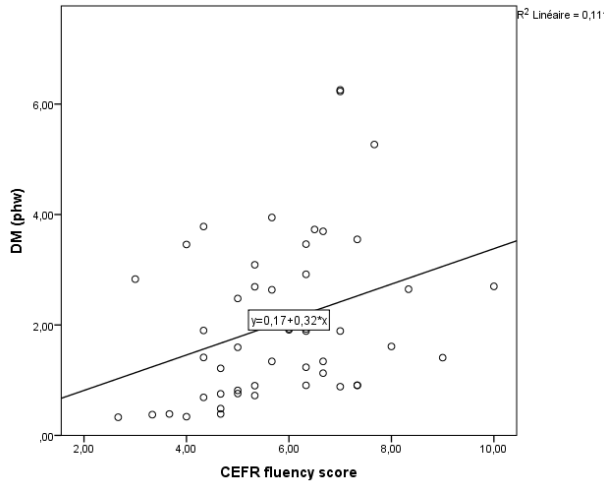


Figure 7-8: The relationship between discourse markers and CEFR fluency score

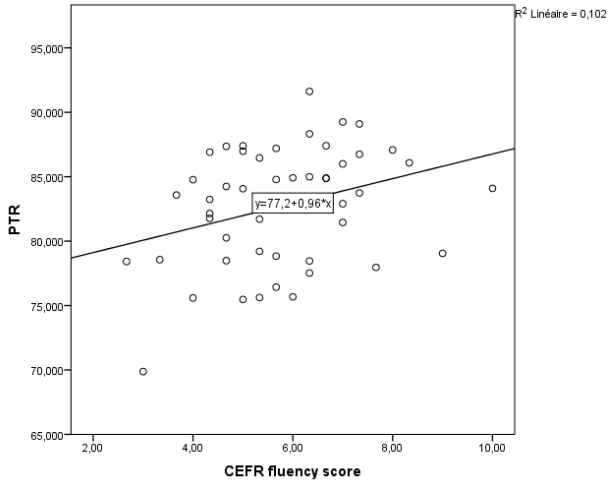


Figure 7-9: The relationship between phonation-time ratio and CEFR fluency score

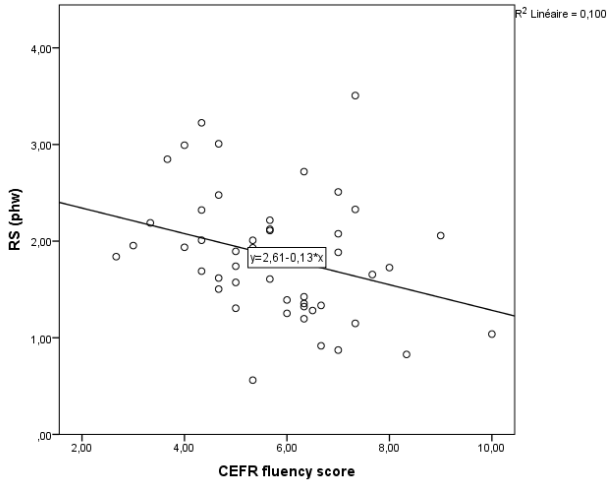


Figure 7-10: The relationship between restarts and CEFR fluency score

7.4.1.2 Contrasting B2 and C1 learners

While the previous section sought to establish how the (dis)fluency measures relate with CEFR fluency ratings in the previous section, this section **zooms in on B2 and C1 learners**[167] and aims to identify whether the two groups differ significantly in their use of (dis)fluency features.

The means of B2 (*n* = 22) and C1 (*n* = 26) learners for each of the 14 variables as well as for the 5 (dis)fluency components are displayed in Table 7-21 and Table 7-22, respectively. As can be observed, compared to C1 learners, B2 learners generally produce more and longer unfilled pauses, produce shorter runs of speech, have a lower phonation-time ratio as well as a lower speech rate. In other words, their mean **temporal (dis)fluency** score is generally lower than that of C1 learners. B2 learners are also associated with a higher **repair (dis)fluency** score (i.e. a higher rate of truncations, restarts, and repetitions). Discourse markers occur on average less frequently and filled pauses more frequently in the B2 group, which corresponds to a lower **pragmatic (dis)fluency** score at B2 level. Although false starts are nearly as frequent in the two CEFR groups, conjunctions are less frequent at the C1 level (i.e. the mean Component 4 score is lower at C1), and foreign words are more frequent at the B2 level (i.e. Component 5 score is higher at B2).

(Dis)fluency measures	B2 (*n* = 22)	C1 (*n* = 26)
Conjunctions (phw)	5.37 (1.53)	4.82 (0.91)
Discourse markers (phw)	1.43 (1.12)	2.55 (1.57)
False starts (phw)	0.68 (0.37)	0.70 (0.28)
Filled pauses (phw)	8.25 (2.57)	7.64 (2.96)
Foreign words (phw)	0.55 (0.48)	0.44 (0.58)
Lengthenings (phw)	2.93 (0.89)	3.36 (1.07)
Mean length of runs	5.40 (0.91)	5.79 (1.11)
Mean UP length (sec)	0.53 (0.09)	0.49 (0.09)
Phonation time ratio	81.46 (4.68)	83.93 (4.26)
Repetitions (phw)	4.02 (1.21)	3.87 (1.60)
Restarts (phw)	2.02 (0.61)	1.71 (0.62)
Speech rate (wpm)	156.37 (12.64)	166.84 (16.52)
Truncations (phw)	1.76 (0.73)	1.57 (0.66)

[167] The number of C2 learners is too small (there are only two) for them to be reliably included in the analyses of this section.

Unfilled pauses (phw)	13.64 (2.82)	11.98 (2.62)

Table 7-21: Means (sd) of B2 and C1 learners in LINDSEI-FR+ for the 14 (dis)fluency variables

(Dis)fluency components	**B2** (***n*** **= 22)**	**C1** (***n*** **= 26)**
Comp. 1 – temporal (dis)fluency	-0.34 (0.94)	0.26 (1.02)
Comp. 2 – repair (dis)fluency	0.16 (0.86)	-0.10 (1.13)
Comp. 3 – pragmatic (dis)fluency	-0.38 (0.89)	0.25 (1.02)
Comp. 4 – cohesion	0.15 (1.27)	-0.20 (0.65)
Comp. 5 – lexico-grammatical (dis)fluency	0.11 (0.93)	-0.09 (1.08)

Table 7-22: Means (sd) of B2 and C1 learners in LINDSEI-FR+ for the 5 (dis)fluency component scores

Independent-samples t-tests were run to compare the means of B2 and C1 learners for each of the 14 variables (Table 7-23) as well as for the 5 (dis)fluency components (Table 7-24). With the exception of Component 4, Levene's test[168] of equality of variances is non-significant for all of the variables under investigation: homogeneity of variances in the two groups can thus be assumed for those variables and components. For Component 4, however, Levene's test is significant, indicating that homogeneity of variances cannot be assumed; the reported values for the *t*-test for this component are adapted correspondingly. In the two tables, the significant mean differences are shown in bold font.

Although five (dis)fluency measures were significantly correlated with CEFR ratings, it is striking to see that **B2 and C1 learners** only significantly differ with respect to three of these (dis)fluency variables: **discourse markers, speech rate and unfilled pauses** (a visual representation is provided in Figure 7-11 to Figure 7-13). C1 speakers use significantly more discourse markers per hundred words, speak faster on average, and produce fewer unfilled pauses than B2 speakers. Restarts tend to occur less frequently in C1 speech and phonation-time ratio tends to be higher in the advanced group, but these tendencies do not reach significance. In terms of (dis)fluency components, the two components that were correlated with CEFR ratings also significantly distinguish B2 from C1 learners: C1's **temporal and pragmatic (dis)fluency** are generally higher than that of B2 speakers. Note also that the effect sizes (Cohen's *d*) of the significant differences range between 0.6 and 0.8, which correspond to medium to large effects. This indicates that the two groups do not only differ with respect to those (dis)fluency variables or components, but that they differ a lot.

[168] Levene's test results can be found in Appendix 9.10.

(Dis)fluency variables	Independent-samples *t*-test
Conjunctions (phw)	*t* = 1.533; *p* = .132
Discourse markers (phw)	***t* = -2.792; *p* = .008; *d* = 0.820**
False starts (phw)	*t* = -.111; *p* = .912
Filled pauses (phw)	*t* = .754; *p* = .455
Foreign words (phw)	*t* = .683; *p* = .498
Lengthenings (phw)	*t* = -1.489; *p* = .143
Mean length of runs	*t* = -1.341; *p* = .187
Mean UP length (sec)	*t* = 1.505; *p* = .139
Phonation-time ratio	*t* = -1.918; *p* = .061
Repetitions (phw)	*t* = .365; *p* = .716
Restarts (phw)	*t* = 1.745; *p* = .088
Speech rate[169] (wpm)	***t* = -2.432; *p* = .019; *d* = 0.712**
Truncations (phw)	t = .924; *p* = .360
Unfilled pauses (phw)	**t = 2.117; *p* = .040; *d* = 0.611**

Table 7-23: Independent-samples t-test results for the 14 (dis)fluency variables in B2 and C1 learner speech
Note: Levene's test of equality of variances was non-significant for all variables.

(Dis)fluency variables	Independent-samples *t*-test
Comp. 1 – temporal (dis)fluency	***t* = -2.082; *p* = .043; *d* = 0.605**
Comp. 2 – repair (dis)fluency	*t* = .853; *p* = .398
Comp. 3 – pragmatic (dis)fluency	***t* = -2.265; *p* = .028; *d* = 0.660**
Comp. 4 – cohesion	*t* = 1.164; *p* = .254
Comp. 5 – lexico-grammatical (dis)fluency	*t* = .678; *p* = .501

Table 7-24: Independent-samples t-test results for the 5 (dis)fluency components in B2 and C1 learner speech
Note: Levene's test of equality of variances was significant for Component 4 (F = 4.751; p = .034), indicating that the assumption of equality of variances was not met. T-test results are adapted for the equality of variances not assumed.

[169] In the rated excerpt: *t* = -2.509; *p* = .016 (non-significant Levene's test).

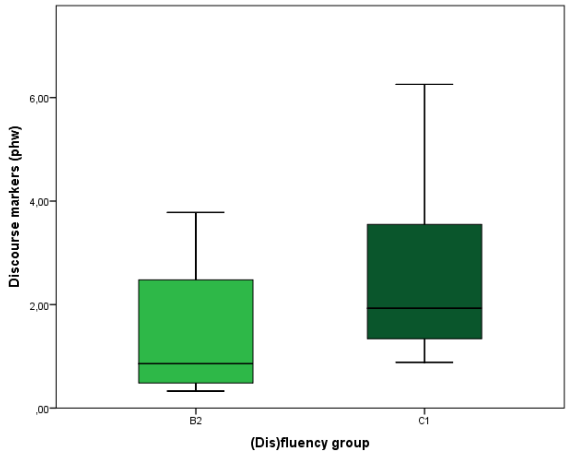


Figure 7-11: Boxplots of discourse markers (phw) at B2 and C1 level

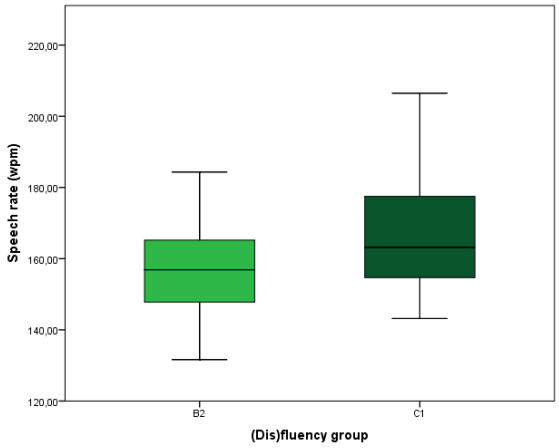


Figure 7-12: Boxplots of speech rate (in wpm) at B2 and C1 level

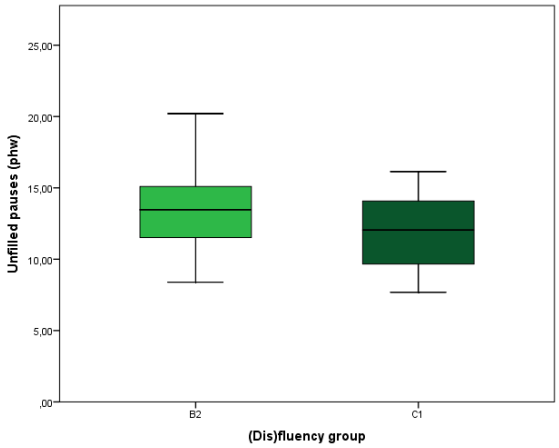


Figure 7-13: Boxplots of unfilled pauses (phw) at B2 and C1 level

Before concluding this section contrasting B2 and C1 learners, I would like to illustrate a **typical B2 fluency level and a typical C1 fluency level** by means of two examples.

FR005 (Example 7-4) is a typical B2 learner. He speaks with a speech rate of 153.06 words per minute, and produces 0.75 discourse marker as well as 13.70 unfilled pauses per hundred words. In the excerpt, the learner also produces seven filled pauses, four truncations, and sometimes repeats or restarts words or utterances. FR031 (Example 7-5) is a typical C1 learner. This learner speaks significantly faster than FR005 (a rate of 177.32 wpm) and produces 7.68 unfilled pauses and 6.23 discourse markers (mainly the DM *well*) per hundred words. In the excerpt, the UPs of FR031 are generally slightly shorter than those of FR005. Although the C1 learner also sometimes produces truncations or restarts, overall, FR031 seems to produce less interrupted, more flowing speech.

7-4: A typical B2 learner (FR005-F)

<A> have you have you travelled much

<B> em not not a lot I went er twice to Ireland (0.470) er (0.320) twice to London (0.210) I thi= yes

<A> and what did you think of London

<B> I I don= don 't like it very much (1.080) er <sighs/> I I f= I fou= (0.420) I found it er (0.400) I found that it was very polluted (0.690) erm I don 't know <sighs/> (0.490) no <sighs/> I don 't like er (2.240) […]

7-5: A typical C1 learner (FR031-F)

<B> well of cour= well at licence level we we don't have any eh conversation classes any more anyway

<A> yeah

<B> so it doesn't make a grea= (0.220) great change

<A> yeah (1.470) yes I suppose

<B> well of course for the (0.160) the ability of speaking (0.300) well (0.230) <laughing/> a small= smaller groups were better of course

7.4.1.3 Predicting (dis)fluency ratings (multiple linear regression analysis)

The two previous sections looked at the relationship between CEFR fluency ratings and grades and (dis)fluency measure considered independently. This section goes a step further and examines the extent to which CEFR fluency ratings can be predicted based on a combination of (dis)fluency measures using multiple regression analysis.

Multiple regression analysis is a statistical method involving the prediction of an outcome variable (here CEFR fluency ratings) from several predictor variables. "Predictor variables" is actually a cover term for both individual independent variables (here the (dis)fluency measures) and their interactions (for example, the interaction between discourse markers and speech rate). Mathematically speaking, the combined effect of two variables (i.e. an interaction) is the effect of the two variables multiplied together (Field 2013:400). Gries (2013:249) further states that two (or more) variables are said to interact when their joint effect on the outcome variable is *not predictable* from their individual effects on the same dependent variable. For example, Table 7-25 displays the significant correlations between individual (dis)fluency variables as well as their interactions, and CEFR fluency ratings. Five (dis)fluency variables significantly correlate with CEFR fluency ratings: speech rate, unfilled pauses, discourse markers, phonation-time ratio and restarts (*cf*. Section 7.4.1.1). The interaction between unfilled pauses and speech rate is, however, not significant, and, conversely, the interaction between conjunctions and filled pauses (both of which are not significantly correlated with CEFR fluency ratings) is actually significantly correlated with CEFR fluency ratings (r = -.295; p = .019).

	Pearson's r	***p***
DM	.332	.009
PTR	.320	.012

	Pearson's r	***p***
FP*T	-.289	.021
FP*UP	-.395	.002

RS	-.317	.013	**FP*W**	-.239	.047
SR	.433	.001	**MLR*PTR**	.280	.024
UP	-.358	.005	**MLUP*RS**	-.421	.001
C*DM	.284	.023	**MLR*SR**	.354	.006
C*FP	-.295	.019	**MLUP*T**	-.341	.008
C*MLUP	-.254	.037	**MLUP*UP**	-.392	.002
C*RS	-.352	.006	**PTR*RS**	-.255	.037
C*T	-.331	.009	**PTR*SR**	.444	.001
C*UP	-.328	.010	**PTR*UP**	-.327	.010
DM*L	.291	.020	**Rep*UP**	-.300	.017
DM*MLR	.364	.005	**RS*T**	-.283	.023
DM*MLUP	.265	.032	**RS*UP**	-.478	.000
DM*PTR	.354	.006	**T*UP**	-.416	.001
DM*SR	.361	.005	**UP*W**	-.235	.050
FP*MLUP	-.319	.012	***Other variables and interactions***		*n.s.*
FP*RS	-.342	.008			

Table 7-25: Pearson correlations between predictor variables and CEFR fluency ratings in the linear regression analysis
Notes: (1) only significant correlations are shown; (2) C = conjunctions; DM = discourse markers; FP = filled pauses; FS = false starts; L = lengthenings; MLR = mean length of runs; MLUP = mean length of unfilled pauses; PTR = phonation-time ratio; Rep = repetitions; RS = restarts; SR = speech rate; T = truncations; UP = unfilled pauses; W = foreign words

Although interactions are not easy to interpret, they are essential in regression models because they help account for the data, often much better than by using individual variables only: leaving out interactions from a regression model actually runs the risk of decreasing both the explanatory and predictive aspects of the modelling process (Gries 2013:255).

It is sometimes advised to only include interactions (and, more generally, predictors) for which there is a sound *a priori* theoretical rationale (Field 2013:321). To the best of my knowledge, no study in L2 (dis)fluency research has demonstrated the importance of any interaction between (dis)fluency variables in predicting CEFR fluency ratings. In addition, given that this analysis is exploratory in nature, it seems more appropriate to cast the net wide and not to exclude any interaction.

In addition to the issue of the selection of predictors, another concern in multiple regression analysis is **multicollinearity**, that is, the strong correlation between two or more predictors. High levels of multicollinearity pose threat to the model estimates (Field 2013:324–326; Hee Jeon 2015:137–140). The examination of the correlation matrix of predictor variables as well as of two collinearity diagnostics (the variance inflation factor (VIF) and the tolerance statistic – see Table 7-26) showed no cause for concern for this analysis.

Finally, in multiple regression analysis, the **method (or direction) in which the predictors are entered into the model** can also influence the final model. There are three main approaches (Field 2013:322–323; Gries 2013:260):

- the backward selection, which starts with a maximal model containing all predictors; the predictors that do not contribute (or contribute very little) to the model are successively discarded;
- the forward selection, which starts with a minimal model and successively adds predictors until no addition of a predictor improves the model (or when all available predictors are already in the model);
- the bidirectional selection, which is a combination of the backward and forward selection.

Although backward selection is more generally used in learner corpus research, in this study, the number of predictors is too high[170] compared to the number of learners to use this method accurately. Considering the large number of predictors, the **forward selection** was selected instead[171].

The forward selection procedure starts with an initial model containing only a constant, and includes, from the 105 predictors available, the predictor that best predicts CEFR fluency ratings. In fact, this first predictor corresponds to the variable that has the highest simple correlation with CEFR fluency ratings (i.e. RS*UP, r = -.478; p = .000; *cf*. Table 7-25). Then, a second predictor is included that has "the largest semi-partial correlation with the outcome" (Field 2013:322). In other words, the algorithm looks for the predictor that can account for the highest part of the remaining variation and make a significant contribution to the predictive power of the model. The procedure stopped after this second model containing two predictor variables as the inclusion of other predictors did not make significant contributions to the model.

The **final regression model** thus consists of two predictor variables, or, more specifically, two interactions: first, the **interaction between restarts and unfilled pauses** (RS*UP), and, second, the **interaction between discourse markers and speech rate** (DM*SR). It is particularly important to underline at this stage that the combination of predictors that can best account for CEFR fluency ratings is a combination of two interactions, and not individual variables (such as speech rate or unfilled pauses). This stresses the importance of taking interactions into account in such models. Second, it seems noteworthy that the four original

[170] In total, there are 105 predictor variables: 14 main effects and 91 interactions.

[171] I also tested the bi-directional method and it resulted in the same final model as the one presented here, i.e. with two predictor variables.

variables in the interactions (restarts, unfilled pauses, discourse markers, and speech rate) significantly correlated with CEFR fluency ratings (see Table 7-25).

Table 7-26 indicates that the **amount of variance** explained in LINDSEI-FR+ by the final model (model 2) is R^2 = .333, i.e. **33.3%**. With only RS*UP as predictor (model 1), R^2 already amounted to .228 (22.8%), and the addition of the second predictor (DM*SR) caused R^2 to improve by 10.5%. This change in the amount of variance explained gave rise to an F-ratio of 7.374, which is significant with a probability of p = .009. It can thus be safely concluded that the addition of the second predictor significantly contributes to the model.

The **adjusted R^2** indicates how well the model generalises, and, ideally, should be the same as, or very close to, the value of R^2. The difference for the final model is rather small (.333-.305 = .028, or 2.8%), which means that if the model was derived from the population, it would account for about 2.8% less variance in CEFR fluency ratings than in LINDSEI-FR+.

As can be seen in Table 7-26, the **Durbin Watson** statistic is very close to 2, which indicates that the assumption of independent errors has almost certainly been met (Field 2013:337).

Model		R	R^2	Adjusted R^2	R^2 change	F change (sig.)	Durbin-Watson	Tolerance	VIF
1	RS*UP	.478	.228	.212	.228	14.197 (p = .000)	/	1.000	1.00
2	RS*UP	.577	.333	.305	.105	7.374 (p = .009)	2.253	.993	1.01
	DM*SR							.993	1.01

Table 7-26: Linear model summary

Model		B (95% CI)	Std. Error	Beta	*t*	Sig.
1	(Constant)	7.471 (6.503, 8.439)	0.482	/	12.515	.000
	RS*UP	-0.071 (-0.109, -0.033)	0.019	-0.478	-3.768	.000
2	(Constant)	6.739 (5.680, 7.799)	0.527	/	12.799	.000
	RS*UP	-0.067 (-0.103, -0.031)	0.018	-0.451	-3.776	.000
	DM*SR	0.002 (0.000, 0.003)	0.001	0.325	2.716	.009

Table 7-27: Linear model of predictors of CEFR fluency ratings

Table 7-27 displays the **model parameters** of the first and second (i.e. final) model. The first column ("B") provides estimates for *b*-values, which indicate the individual contribution of each predictor to the model and the degree to which each predictor affects the outcome

variable if the effects of all the other predictors are held constant. A positive value indicates that there is a positive relationship between the predictor and the outcome. A negative value represents a negative relationship (Field 2013:338). In Table 7-27, values for RS*UP are negative (as the value of the interaction increases, the CEFR fluency rating decreases) and the value for DM*SR is positive (as the value for the interaction increases, so does the CEFR fluency rating). The positive and negative values for each interaction can be traced back to the direction of the correlation of their component variables (for example, both RS and UP correlate negatively with CEFR fluency ratings, which explains the negative value for their interaction).

The **standardised values of *b*** (labelled "Beta") provide better insights into the importance of predictors because they are measured in standard deviation units and are directly comparable. In model 2, the absolute value of the first interaction is higher than that of DM*SR, indicating that **RS*UP contributes more to the model** than the second interaction.

The last two columns in Table 7-27 show the significance of the *t*-test associated with the *b*-value ("B"). A significant result indicates that the predictor contributes significantly to the model. For model 1, the interaction RS*UP is a significant predictor. For model 2, the two interactions are significant predictors of CEFR fluency ratings.

The last step in the validation of the model consists in the **examination of residuals** to find evidence of bias (Field 2013:345–348). In the interest of space, the data for the analysis of residuals as well as a report of their examination are included in Appendix 9.12. Based on the analysis of residuals, the model appears to be fairly reliable, and not unduly influenced by any case. The assumptions of linearity and homoscedasticity [172] have been met, and the distribution of residuals is roughly normal.

All in all, the model appears to be accurate for the sample. It includes two significant predictor variables: a first interaction between restarts and unfilled pauses, and a second interaction between discourse markers and speech rate, which, together, account for about a third of the variance in LINDSEI-FR+. A third of the variance might not seem very significant, but it is in fact a very promising starting point. Also, as Jarvis *et al.* (2003:377) rightly point out, "[e]ven when variables such as proficiency, language background, topic, and audience have been controlled, straightforward predictive relationships between linguistic variables and quality ratings have remained elusive, and perhaps they always will".

Further analyses are needed to explain why the two interactions were retained in the final model (and why only interactions were retained), but it is important to stress that the four variables that make up the two interactions are all significantly correlated with CEFR fluency ratings. This seems to suggest that those variables and their interactions could be

[172] Homoscedasticity is "an assumption in regression analysis that the residuals at each level of the predictor variable(s) have similar variances" (Field 2013:876).

emphasised more in rating descriptors of high intermediate and advanced learner (dis)fluency, at least for French-speaking learners of English. Similar analyses should be carried out to examine whether the same predictor variables emerge in other learner populations. It should also be stressed that the model presented here was trained on the whole of LINDSEI-FR+, but this runs the risk of yielding overoptimistic estimates if the results are generalised. For more reliable predictive results, cross-validation procedures should be used, where part of the data is used to train the model, and the remaining data are used to evaluate its performance (see e.g. Arlot & Celisse (2010) for a review of cross-validation procedures).

7.4.2 CEFR fluency levels and (dis)fluency profiles

In the previous sections, the nature of the relationship between (dis)fluency variables (and (dis)fluency components) and CEFR fluency ratings has been investigated by means of correlational analyses, *t*-tests, and multiple linear modelling. This section takes a slightly different perspective: rather than assuming a linear relationship between (dis)fluency features (and components) and CEFR fluency scores, this section explores the link between the **(dis)fluency profiles** established in Chapter 6 (i.e. the clusters) **and the CEFR fluency *levels*** (i.e. B2, C1 and C2). In other words, this section seeks to determine whether the clusters can reliably be associated with a specific CEFR fluency level. For example, are C1 learners more likely to belong to the (dis)fluency profile B?

As a reminder, **two cluster solutions** were identified in Chapter 6: a 2- and a 6-cluster solution. Cluster 1 in the 2-cluster solution was associated with a higher temporal (dis)fluency, and Cluster 2 with a lower temporal (dis)fluency. It could thus be expected that Cluster 1 is more strongly related with the C1 (and C2) level, and Cluster 2 with the B2 level. In the 6-cluster solution, Cluster 1 and Cluster 2 were subdivided into three sub-clusters each: Cluster 1 includes Clusters 1 to C and Cluster 2 includes Clusters D to F. As for the 2-cluster solution, it could be expected that Clusters A to C are more strongly associated with C1 (and C2) level and Clusters D to F with B2 level. Or, alternatively, following the suggested (albeit tentative) ranking of the (dis)fluency profiles along the (dis)fluency continuum in Chapter 6, Cluster A could be associated with a higher perceived (dis)fluency level, and Cluster E with a lower level. If there is indeed some correspondence between the clusters and the human ratings, this could open up new avenues for automated ratings of (dis)fluency levels. If the correspondence is limited or inexistent, however, this implies that (dis)fluency profiles are independent from CEFR fluency level and that the raters may have been influenced by other factors that are not captured by the profiles, such as the topic, the degree of interactivity, or the pronunciation.

Table 7-28 shows the cross-tabulated data between the clusters from the **2-cluster solution** and the CEFR fluency levels. For Cluster 1, about two thirds (14; 61%) of the learners were

assessed at C1 level for (dis)fluency. Moreover, 1 learner has a C2 level. About 35% (8) of the learners have a B2 level for fluency. In the second cluster, 52% (14) of the learners have a B2 level, and 44% (12) have a C1 level. One learner has C2 for fluency. Although Cluster 1 seems to be more closely related to the C1 level, the number of C1 learners is, in fact, spread nearly evenly across the two clusters (14 in Cluster 1 and 12 in Cluster 2), which does not point to a strong association between cluster membership and CEFR level. A Chi-squared test confirmed that the association between (dis)fluency profile and CEFR level is **not significant** ($X^2 = 1.479$; $p = 0.477$).

	Cluster 1	**Cluster 2**	**Total**
B2	8	14	22
C1	14	12	26
C2	1	1	2
Totals	**23**	**27**	**50**

Table 7-28: Contingency table for the 2-cluster solution

	Cluster 1			**Cluster 2**			**Total**
	Cluster A	**Cluster B**	**Cluster C**	**Cluster D**	**Cluster E**	**Cluster F**	
B2	2	2	4	10	3	1	22
C1	7	2	5	4	3	5	26
C2	1	0	0	1	0	0	2
Totals	**10**	**4**	**9**	**15**	**6**	**6**	**50**

Table 7-29: Contingency table for the 6-cluster solution

Figure 7-14: Proportion of B2, C1, and C2 learners per cluster in the 6-cluster solution

Table 7-29 displays the contingency table for the **6-cluster solution** (a more reader-friendly representation of the data is shown in Figure 7-14). While Clusters A and F both mainly include C1 learners (about 80%), three clusters (B, C, E) include an even, or a nearly even, proportion of B2 and C1 learners each. Lastly, although Cluster D contains the highest proportion of B2 learners (66%), it also includes a C2 learner (Fr004). This is very intriguing and deserves closer examination.

FR004 has a low phonation-time ratio (79.05%, which is actually lower than the mean for B2 learners [81.46 %]) and produces quite a lot of restarts (2.06 phw), but few discourse markers (1.41 phw[173]). All of these values can be associated with a lower perceived (dis)fluency (as these three variables are significantly correlated with CEFR fluency ratings). However, FR004 also speaks very fast – in fact, faster than the mean speech rate of C1 learners (185.5 wpm vs. 166.8 wpm for a mean C1 learner), and produces slightly fewer unfilled pauses than the mean C1 learner (11.86 phw vs. 11.98 phw). These two measurements can be associated with higher perceived CEFR fluency level. In terms of individual (dis)fluency measures, FR004 thus seems to offer a mixed picture, between B2 and C1. Furthermore, it has been shown that the values obtained for two interactions, namely RS*UP and DM*SR, can be reliably related to CEFR fluency level. In this case, the values obtained by FR004 for those two interactions both predict a lower perceived (dis)fluency level. This might also explain the statistical association between FR004 and B2 learners. It does not, however, explain why this learner was evaluated at C2 level for fluency by the raters. The regression model presented in the previous section accounted for c. 33% of the variation in CEFR fluency ratings. It may well be that the raters also took other elements into account in their evaluation of the learner's (dis)fluency that

[173] As a reminder, the mean number of restarts for B1 learners is 2.02 phw and 1.71 phw for C1 learners; and the mean number of discourse markers is 1.43 phw (B2) and 2.55 phw (C1), see Table 7-21.

were not represented by the (dis)fluency measures in the present analysis. For example, the topic might have exerted some influence on the raters (in the CEFR rated excerpt, FR004 talks about her ambition to enter a UK university), or the use of colloquialisms (see my comment on Example 7-2 in Section 7.2.3), or other aspects of the learner's spoken competence such as pronunciation, vocabulary range or grammatical accuracy (see Section 7.4.3 in this respect).

Moving away from the special case of FR004, Table 7-29 above seems to point to some correspondence between Clusters and CEFR fluency level, especially between Clusters A and F (with higher CEFR fluency ratings) and – to some extent – Cluster D (with lower ratings). However, a Chi-squared test on the data from Table 7-29 indicates that the association between cluster membership in the 6-cluster solution and CEFR fluency level is **not statistically significant** (X^2 = 9.96, p = 0.443). Nonetheless, the examination of the standardised residuals reveals that Cluster D does include more B2 speakers than expected, and fewer C1 speakers than expected, which reveals a tendency for this cluster to be more typical of B2 learners.

In **conclusion**, it emerged from the analysis that the correspondence between CEFR fluency levels and statistically-based clusters of learners is nearly non-existent, although some tendencies could be delineated. The lack of correspondence could be explained by three main factors.

First, it could be hypothesised that (dis)fluency profiles and fluency levels are two distinct concepts that provide **different types of information**. The (dis)fluency profiles reveal the overall **idiosyncratic speaking style** of a speaker, without this being indicative of a higher or a lower perceived (dis)fluency level. A slower speaker is for example not disfluent (or less fluent) by default, and a fast speaker might be perceived as not very fluent if his/her use of the other (dis)fluency features is not adequate. Within each profile, however, the specific weighting of the (dis)fluency variables might be different, which therefore results in a different perceived fluency level. Alternatively, it might be that the identification of the fluency level of the learners remains largely dependent on aspects of performance that are only partially represented by the (dis)fluency measures and profiles.

A second possible explanation lies in the methodology used to obtain the **CEFR fluency level** (for example, the conversion of the CEFR grades into scores, which are averaged across the three raters, and converted back into CEFR fluency levels). The methodology applied here might perhaps not ideally reflect the learners' (dis)fluency level and further consideration could be given to the rating procedure. It might, for example, be important to integrate a greater number of raters in the analysis, or to adapt the averaging/conversion system used here.

A last factor that could account for the lack of correspondence between the (dis)fluency profiles and the CEFR fluency ratings is that the clusters are based on the combination of **14**

(dis)fluency variables, including measures that are not explicitly mentioned in the CEFR descriptors, while the raters were required to assess the learners based on the CEFR descriptors. It might be that the inclusion of a larger panel of (dis)fluency measures in the cluster analysis, to some extent, skews the results. It is reasonable to think that clusters based only on a subset of those variables (that is, those that are mentioned in the CEFR fluency scale) could reveal clearer tendencies. Furthermore, as discussed in Chapter 6, the way the (dis)fluency profiles were computed depends on many factors. It would be interesting to test whether different clustering procedures affect the nature of the relationship between (dis)fluency clusters and CEFR fluency level.

7.4.3 The relationship between CEFR ratings

Before concluding this chapter, I would like to come back to a comment on FR004, who, despite being a C2, was included in the cluster that contained the most B2 learners (see Section 7.4.2). I argued that, although empirical measurements of the (dis)fluency of this learner could explain this association, other factors might also have influenced the raters' judgement. In this section, I will briefly examine the extent to which ratings of vocabulary range, grammar, pronunciation, coherence, and global proficiency are related to CEFR fluency ratings.

As a reminder, for the purpose of this thesis, three raters were asked to provide CEFR grades not only for (dis)fluency, but also for four other skills as well as for the global proficiency of the learner (*cf*. Section 7.2) based on a scale from the *Common European Framework of Reference for Languages*. Each learner was thus attributed a grade for his or her vocabulary range, grammatical accuracy, phonological control, coherence, and global proficiency. Prior to examining the nature of the relationship between those scales and the fluency scale, the grades were converted into **numerical values** following the same methodology as described in Section 7.2.2 (see esp. Table 7-11) for the fluency grades.

Pearson's product-moment correlations were run to examine the nature of the relationship between CEFR fluency scores and the scores obtained for the other skills. I ran the analysis on each rater separately to have a better understanding of potential differences between judges. Table 7-30 displays the correlation coefficients per rater (all *p* values = .000).

	Rater	Range	Accuracy	Phonological control	Coherence	Global proficiency
CEFR fluency score	**R1**	*r* = .78	*r* = .73	*r* = .83	*r* = .77	*r* = .88
	R2	*r* = .71	*r* = .53	*r* = .67	*r* = .79	*r* = .72
	R3	*r* = .89	*r* = .86	*r* = .75	*r* = .84	*r* = .92

Table 7-30: Correlations between CEFR fluency scores and the other CEFR skills per rater

Table 7-30 reveals that **for all three raters, fluency scores are highly significantly correlated with the other skills as well as with global proficiency**. The correlation coefficients exceed 0.70, indicating **substantial levels of agreement** (Jarvis 2002), except for two correlations (both for the second rater) which are moderate (though one is nearly substantial too). The **global proficiency** scores are particularly highly correlated with fluency scores for the three raters, with *r*s of .72 (R2), .88 (R1) and up to .92 (R3). It is also interesting to see that, for R1 and R2, the lowest correlation is with accuracy, and for R1 and R3, the highest correlation is with global proficiency. Phonological control, which is often thought to be closely related with the perception of (dis)fluency (Anderson-Hsieh & Koehler 1988; Munro & Derwing 2001; Derwing *et al.* 2004; Pinget *et al.* 2014), is highly correlated with fluency for R1 and R3, but moderately related with fluency for the second rater. This difference might be due to slightly different rating styles (see e.g. Upshur & Turner 1999).

The results of the correlational analysis tend to confirm previous findings from the literature. Baker-Smemoe *et al.* (2014), for example, investigated the relationship between fluency and **proficiency** and showed that some utterance fluency measures correlate with learner proficiency across L2s. Previous studies also suggest that the stronger the perceived foreign **accent**, the lower the fluency ratings (Anderson-Hsieh & Koehler 1988; Derwing *et al.* 2004; Munro & Derwing 2001; Pinget *et al.* 2014; Rose 2011). Similarly, in this thesis, perceived fluency and perceived phonological control are significantly correlated, and perceived fluency and perceived proficiency are correlated as well. With respect to **accuracy**, although empirical measures of accuracy and fluency were found not to be correlated (Brand & Götz 2011), it appears that the *perception* of accuracy and the perception of fluency are. To date, few studies have addressed the contribution of **vocabulary** knowledge (i.e. range) to L2 fluency. Findings tend to indicate a positive relationship between productive vocabulary knowledge and learner fluency (Hilton 2008; Uchihara & Saito 2016). The results from Table 7-30 likewise suggest that there is a strong positive correlation between the perceptions of vocabulary range and fluency – the strength of the correlation even reaches 0.89 for R3. Lastly, according to the CEFR descriptor scale, **coherence** implies the use of "connectors and other cohesive devices" (see Table 7-7) such as conjunctions and discourse markers. Previous literature suggests that the use of such cohesive devices is positively related to perceived fluency (Dore 2016; Neary-Sundquist 2014; House 2013). In LINDSEI-FR+, the perception of coherence is also correlated with perceived fluency and the strength of the correlation is substantial for the three raters.

Overall, thus, it seems that the perception of all the CEFR rated skills are closely and positively correlated. It seems plausible that, in the case of FR004 (but also for all other learners), a high performance on one (or several) of those skills might have coated a somewhat poorer performance on some (dis)fluency measures. In other words, and although there is obviously need for further research in this domain, it is not impossible that a learner's performance on other spoken skills might, in fact, also account for some part of the variability in CEFR fluency ratings.

7.5 DISCUSSION: TOWARDS A NEW PERSPECTIVE ON B2 AND C1 FLUENCY DESCRIPTORS?

This chapter has sought to investigate the links between perceived (dis)fluency level and corpus-based measurements of the (dis)fluency of French-speaking learners.

I have first presented the *Common European Framework of Reference for Languages* and the various **CEFR grids and descriptors** pertaining to speech and (dis)fluency. I have underlined that, despite its widespread use, the CEFR grids for spoken competence and fluency suffer from major weaknesses including the following: the lack of consistency across the levels, the overreliance on downtoners ("*fluent*" vs. "*very fluent*"), the vagueness around the use of some terms (e.g. *hesitations*), the idealised C2 level descriptors, the bias towards monologic speech, or the lack of awareness of task influence on (dis)fluency.

In the second section of this chapter, the **procedure for the CEFR rating** of the 50 learners of LINDSEI-FR+ has been described in detail. Three native-speaker and professionally trained raters graded a five-minute excerpt from the free discussion part of each learner interview in LINDSEI-FR+ according to the CEFR grid and descriptors for linguistic competence (*cf*. Table 7-7). An inter-rater agreement analysis on the grades attributed for fluency showed that these grades were reliable for further analysis. The question was then raised whether the CEFR fluency grades could reliably be applied to the whole learner interview, or whether the learner performance differed in terms of (dis)fluency measures in the 5-minute CEFR rated excerpt and in the whole interview. The analysis revealed that the CEFR fluency grades could quite safely be extended to the whole interview.

Section 7.4, which includes the main analyses, was organised into three subsections. I first examined the relationship between **CEFR fluency ratings and corpus measurements of L2 (dis)fluency** by means of correlational analyses, *t*-tests and multiple linear regression. The analyses highlighted that, among the panel of 14 (dis)fluency measures, only five measures significantly correlate with (dis)fluency ratings, namely discourse markers, phonation-time ratio, restarts, speech rate, and unfilled pauses. Moreover, among these five variables, only three (speech rate, discourse markers and unfilled pauses) significantly discriminate the B2 from the C1 learners in LINDSEI-FR+. Furthermore, the interactions between restarts and unfilled pauses, as well as between discourse markers and speech rate, were shown to have the potential to predict the CEFR fluency level of LINDSEI-FR+ learners.

A second subsection addressed the question whether the CEFR fluency grades could be related, to some extent, to the **(dis)fluency profiles** identified in Chapter 6. No clear association was found, as each profile included both B2 and C1 learners, which indicates that profiles and levels provide two different perspectives on a speaker's (dis)fluency. It was suggested that, while the profiles reveal the general speaking style of a speaker, it is the

specific weighting of the (dis)fluency variables within each style that might be indicative of the perceived (dis)fluency level.

Finally, in a last analysis, I examined the extent to which the grades provided for perceived CEFR fluency level by the three raters correlated with the perceived CEFR level for **vocabulary range, grammatical accuracy, phonological control, coherence, and global proficiency**. All CEFR scales, and particularly the global proficiency scale, highly correlate with the CEFR fluency scale. It fell out of the scope of this thesis to investigate further the extent to which range, accuracy, pronunciation, coherence and proficiency might affect CEFR fluency ratings, but future research could further probe into the interrelationships between CEFR scales and speaking subskills, e.g. via the Complexity-Accuracy-Fluency (CAF) framework.

This study focused on higher **intermediate and advanced French-speaking learners**, i.e. mainly B2 to C1 learners. The CEFR descriptors from the *Spoken Interaction scale* (Table 7-4) and from the *Spoken Fluency scale* (Table 7-5) claim that B2 learners are able to interact *with a degree of fluency* despite *some hesitations*. Higher performers within the B2 band are able to use the language *fluently, accurately and effectively*. At the advanced level, in addition to being able to overcome any lexical gap, learners are said to be able to express themselves *fluently and spontaneously, almost effortlessly* and with *a natural, smooth flow*. It is, however, a particularly difficult endeavour to try to relate these qualitative appreciations with group means and other quantitative figures. What is, for example, a *degree* of fluency? How much is *some* hesitations?

The statistical analyses in this chapter support the CEFR descriptors in that learners, even at a high-intermediate level (B2), still use *some* filled and unfilled pauses, reformulations etc. However, the analyses stressed that the *fluent, natural,* and *smooth flow* of C1 learners is far from being free from pauses, reformulations and other *hesitations*. Although they tend to decrease as the CEFR fluency level increases, **a fluent C1 or C2 French-speaking learner of English still produces a non-negligible number of (dis)fluency features**. In fact, nearly 40%[174] of the speech of C1 learners in LINDSEI-FR+ is related to (dis)fluency: on average, every five words, there is a pause (filled or unfilled) and some other (dis)fluency feature (e.g. a repetition). C1 and C2 levels thus cannot be equated with the absence of (dis)fluency features. Furthermore, B2 and C1 learners actually only significantly differ in terms of speech rate, unfilled pauses, and discourse markers. Future analyses could investigate whether this is also the case in learners from other mother tongue backgrounds. If it is so, this aspect could be amended in a revised version of the CEFR fluency scale.

It was out of the scope of this thesis to investigate the qualitative use of (dis)fluency features at B2 and C1 levels. Although the frequency of some (dis)fluency features does not

[174] *Cf*. Table 7-21. I added the frequencies per hundred words of all (dis)fluency features, which amounts to a total of 38.64 (dis)fluency features per hundred words. The total for B2 learners is 40.65.

significantly differ from the intermediate to the advanced level, it is not excluded that their qualitative use has evolved. For example, the range of discourse markers might have broadened, the placement of filled and unfilled pauses might be less disruptive, or the combinations of features might be different, which might all contribute to a lower perception of (dis)fluency features.

Another important contribution of this chapter is the finding that **B2 and C1 learners can share the same (dis)fluency profile**. In particular, a C2 learner was shown to be quantitatively close to the mean B2 learner, with long unfilled pauses, many restarts, and few discourse markers. This questions the traditional monolithic view of (dis)fluency: my results seem to indicate that **there might be several paths towards higher CEFR fluency level**.

Lastly, a few words on CEFR **assessment** are in order. In this study, three raters were asked to rate the learner samples. Although they were carefully selected based on their expertise in rating learner spoken data, their degree of agreement showed that the assessment of speech is all but an easy task, even for professionally-trained raters. With the benefit of hindsight, considering the number of raters and their level of agreement, a **weighted averaging method** could have been used instead of a simple (unweighted) average. A weighted average takes into account the strength of the correlations between raters and this might have better reflected the actual CEFR fluency level of the learners.

If the assessment of fluency and speech is not easy for native speaker raters, it is *a fortiori* an even more complex, and time-consuming, endeavour for **non-native teachers** (*cf.* Gilquin, Bestgen & Granger 2016). Investigations of assessment practices in school settings and initiatives to share experiences and good practices should definitely be encouraged. Promising avenues in terms of assessment might come from the domain of natural language processing and the new technologies: tools and **apps** to practise speaking, and algorithms for the **automated scoring** of learner speech are currently being developed. To give but one example, Rose (2015) recently developed an application where a learner is immediately provided with some statistics about the fluency of his or her speech. Coupled with quantitative data on (dis)fluency features at various CEFR fluency levels, this app might offer new perspectives for teachers and the assessment of speech and (dis)fluency.

GENERAL CONCLUSION

I haven't been everywhere, but it's on my list.

(Susan Sontag)

The main quest of this thesis has been to bring to the fore corpus insights into the construct of (dis)fluency in learner and native speech. This general conclusion takes stock of the main findings yielded by this study and puts forward its contributions to three research fields. The discussion finishes off by pointing to worthy avenues to further our current knowledge on fluency and disfluency in learner and native speech.

SUMMARY OF THE MAIN FINDINGS

The summary of the main findings of this study is subdivided into four sub-sections. Each section answers one of the four main research questions that have guided this thesis.

Learner vs. native speaker (dis)fluency

The first research question has to do with the characterization of the (dis)fluency of French-speaking learners of English as compared to British English native speakers. This question was mainly tackled in Chapter 5, which offered a bird's-eye view into the use of (dis)fluency features by French-speaking learners and native speakers, focussing first on the four temporal (dis)fluency measures and then on the ten annotated (dis)fluency features.

LINDSEI-FR+ learners, despite being at a high intermediate to advanced level, prove to have a **significantly lower temporal fluency** than their native counterparts: they speak on average more slowly and produce more unfilled pauses, and their phonation-time ratio and mean length of runs are, consequently, also lower. Surprisingly, however, L2 unfilled pauses, although more numerous, are slightly shorter on average. More analyses need to be carried out to explain this finding. With respect to the **overall frequency of (dis)fluency features**, a mean of 39 (dis)fluency features per hundred words (phw) was found in the learner corpus, which is considerably higher than in the native corpus (22 phw), and in previously reported frequency counts (6 phw in Fox Tree 1995; 5 disfluencies per minute in Kormos & Dénes 2004). This higher incidence is, however, largely due to the wider panel of (dis)fluency features considered. Zooming in on **individual (dis)fluency features**, statistical analyses reveal significant differences in mean frequency between the learner and the native speaker data

for all annotated (dis)fluency features, except for conjunctions and discourse markers. Closer examination of each feature further revealed more subtle differences between L1 and L2 speech.

Three important findings related to this research question ought to be highlighted.

First, this study has shed light on the fact that (dis)fluency features rarely come up alone in speech: on the contrary, they are more often **used in "chunks"** together with one, or more, other (dis)fluency feature. A typical example is filled pauses: only a quarter of FPs occurs in isolation (i.e. with no other (dis)fluency feature in adjacent position), and both learners and native speakers generally use them in chunks together with, e.g., an unfilled pause or a conjunction. This finding is very important as (dis)fluency features have generally been examined from the point of view of the "lexical" context, and only exceptionally from the point of view of the "(dis)fluency" context. Further insights could definitely be gained by a more systematic examination of (dis)fluency features in their "(dis)fluency" context and by bringing to the fore how exactly (dis)fluency occurs in a linear perspective.

Also, despite the fact that, on average, there is, indeed, a "fluency gap" between learners and native speakers, the data actually reveals a more intricate picture: not only is there considerable variability within each group of speakers, but there is also a non-negligible degree of overlap between L1 and L2 distributions for all fourteen (dis)fluency measures.

With respect to the former aspect, dispersion indices reveal that there is **considerable variability** within learner and native performances for each of the fourteen (dis)fluency measures. More specifically, the largest variations pertain to the frequency of foreign words, discourse markers, false starts, truncations, filled pauses and repetitions; the smallest variations lie at the level of temporal (dis)fluency measures (in particular speech rate and phonation-time ratio). Furthermore, while some features appear to display larger variations in learner speech (e.g. discourse markers, and, to a smaller extent, phonation-time ratio), the **majority of (dis)fluency measures display a more considerable variation in native speech** (especially the measures of filled pauses, truncations, and repetitions). Although part of this variation can be attributed to differences in proficiency level or to speaker idiosyncrasies, these findings call for further probing into the under-researched domain of inter-speaker variability.

Related to the issue of variability is the observation that there is a non-negligible degree of **overlap between the learner and native speaker** distributions for each of the fourteen (dis)fluency measures. For example, while learners and native speakers differ significantly in terms of mean speech rate, some learners in LINDSEI-FR+ speak faster than some native speakers from LOCNEC+. Likewise, while learners produce, on average, more filled pauses than L1 speakers, some learners actually produce fewer of them than native speakers. Future research could focus on those "better performing" learners to get a better understanding of the factors contributing to good utterance fluency, as well as on those "poorer performing"

native speakers to get a better understanding of the modulations inherent in native speaker fluency and thereby challenge the long-held assumption that L1 speakers are all, and equally, fluent by default.

All in all, the findings reveal a far more intricate picture of learner and native (dis)fluency than what might at first have been conceived, with speakers performing very differently, and with learners sometimes performing "better" or "more fluently" than native speakers on some (dis)fluency measure.

(Dis)fluency profiles

The second research question aimed to further gauge the importance of idiolects in learner and native speech, and, more particularly, to examine whether (dis)fluency profiles may be identified among French-speaking learners of English and native speakers.

To determine whether learners and native speakers might fall into multiple and significantly different groups based on each speaker's use of (dis)fluency variables, two hierarchical Cluster Analyses were carried out, one on each speaker group. For both learners and native speakers, two cluster solutions were examined.

For the French-speaking **learners**, two large clusters (or "profiles") were identified that significantly differ with respect to five (dis)fluency variables, namely unfilled pauses, speech rate, phonation-time ratio, mean length of runs and mean length of unfilled pauses (i.e. the variables contributing to temporal (dis)fluency, aka learner Component 1). Each cluster contains three sub-clusters that may be characterised by a specific (high or low) use of (dis)fluency features (an overview of the six learner (dis)fluency profiles is provided in Figure 8-2).

Likewise, in the **native** data, two large clusters were uncovered that differ along the six (dis)fluency variables that make up the native Component 1 (i.e. filled and unfilled pauses, mean length of runs, mean length of unfilled pauses, phonation-time ratio and speech rate). Moreover, the two main clusters also with respect to discourse markers. This latter finding is particularly interesting as it sheds new light on the differential use of discourse markers in L1 speech and opens the way to future investigations into the variability of discourse markers in the speech of native speakers. The two large native clusters can further be broken down into five sub-clusters that, like the native profiles, may be characterised by a high or low use of specific (dis)fluency features (a summary of these five native (dis)fluency profiles is shown in Figure 8-3).

The findings from the Cluster Analyses bring support to previous studies that stress the central role of **temporal (dis)fluency variables**. However, they also stress that these measures are equally central when native speech is considered. Another interesting finding

is that the six learner profiles differ with respect to restarts, repetitions, and truncations, i.e. the three repair (dis)fluency variables of Component 2, whereas LOCNEC+ speakers only differ in their use of repetitions: this highlights the importance of the **repair dimension** in learner (dis)fluency, as compared to its relatively less salient role in native (dis)fluency.

Furthermore, it becomes apparent from the characterization of the NS and NNS clusters that learner and native (dis)fluency depends less on the use of individual (dis)fluency variables than on how these are used in combination with one another. Three different ways of combining (dis)fluency variables and components were distinguished and tentatively ranked along a (dis)fluency scale. For the learners, the following ranking was suggested, from the (arguably) most fluent to the least fluent:

Profile A > Profiles C, B, and F > Profile D > Profile E

Likewise, the native profiles have tentatively been ranked in the following order:

Profile E and D > Profile C > Profiles B and A

Such profiles obviously need further corroboration from other studies. It would also be particularly interesting to examine whether the different L1 and L2 profiles are perceived differently by listeners. In this respect, a first step has been made by probing into the **relationship between learner profiles and the learners' assessed CEFR fluency level**. It appears that neither the two nor the six (dis)fluency profiles are related with the learners' assessed CEFR fluency level. Several hypotheses have been offered to account for the absence of clear relationship between learner profiles and CEFR fluency levels, including the fact that (dis)fluency profiles are revealing of learners' idiolects, and thus independent of their perceived fluency level. All in all, my results suggest that the same CEFR fluency level might cover very different types of performances and that the reality is far more complex than a linear relationship between fluency level and independent utterance (dis)fluency measures.

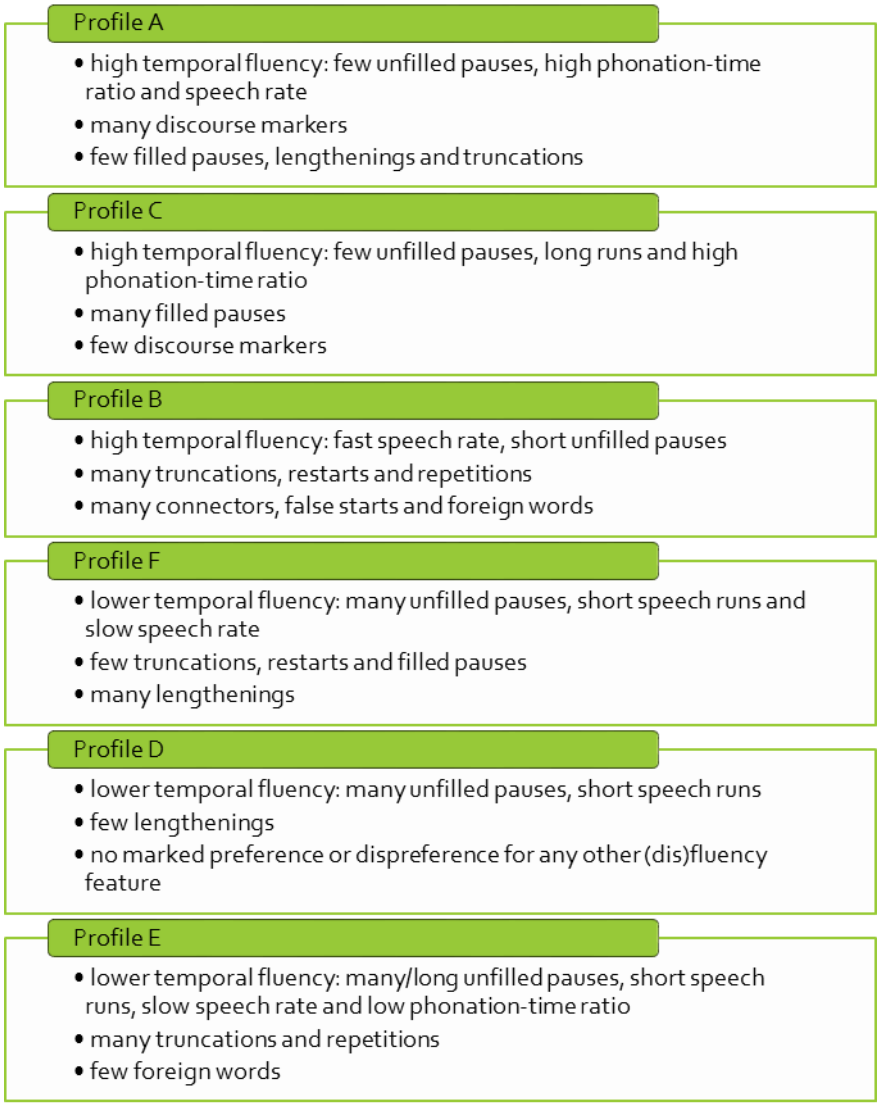


Figure 0-1: Overview of the 6 learner (dis)fluency profiles (from the most to the least fluent)

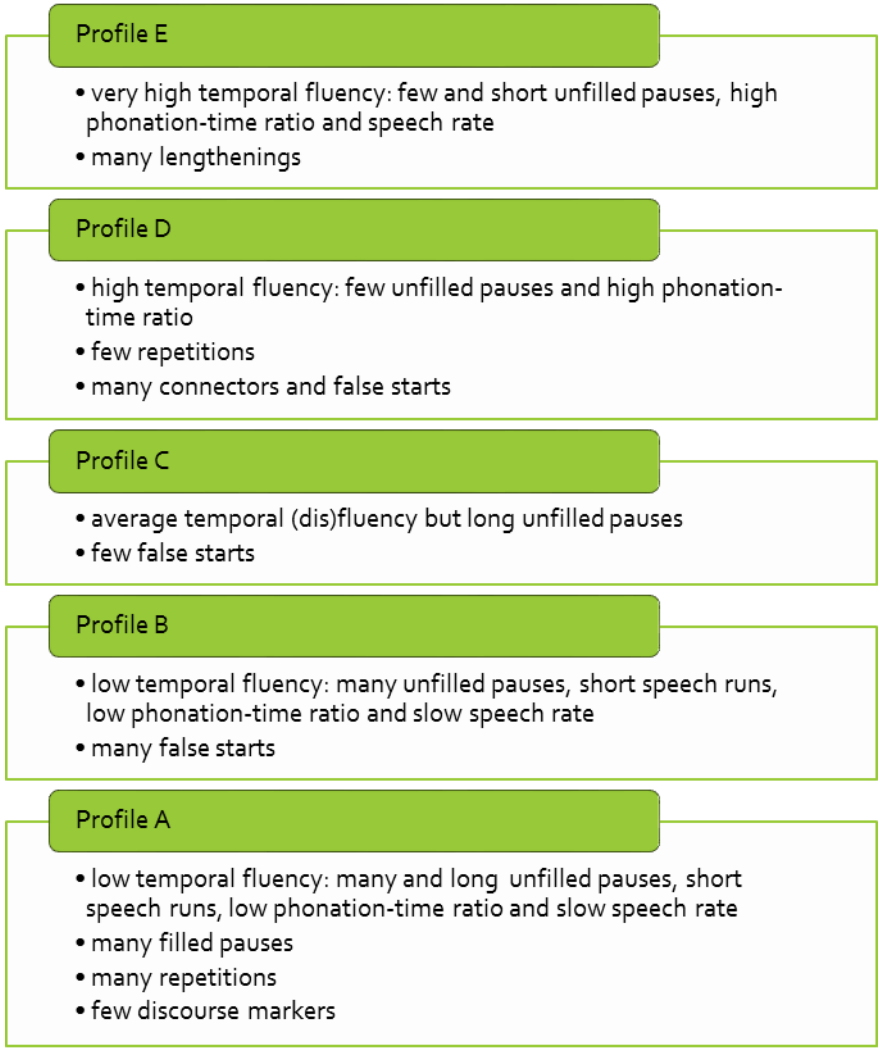


Figure 0-2: Overview of the 5 native (dis)fluency profiles (from the most to the least fluent)

(Dis)fluency dimensions

Any analysis of (dis)fluency has at its heart the question of how the variables contributing to it are interrelated. This issue was at the core of the third research question of this thesis, and was addressed by carrying out two Principal Component Analyses[175], one on the learner data, and one on the native data.

In the **learner corpus, five underlying dimensions** of (dis)fluency were delineated. The first component, termed *temporal (dis)fluency*, includes unfilled pauses, mean length of runs, phonation time ratio and speech rate. The second component, *repair (dis)fluency*, loads on measures of truncations, restarts and repetitions. The third component, *pragmatic (dis)fluency*, includes discourse markers and filled pauses (and, to a smaller extent, speech rate). *Discourse cohesion*, the fourth component, includes conjunctions and false starts. Lastly, the fifth component, tentatively termed *lexico-grammatical (dis)fluency*, loads on foreign words and false starts.

As for native speakers, **four dimensions of L1 (dis)fluency** were uncovered. The first temporal dimension includes the same variables as in the learners' temporal dimension plus filled pauses. The second dimension corresponds to *repair (dis)fluency*, which, like its L2 counterpart, includes truncations, restarts and repetitions, but also lengthenings. The third dimension (*pragmatic (dis)fluency*) includes discourse markers, filled pauses and false starts. Finally, the last dimension, *discourse coherence*, includes conjunctions and false starts[176].

The results of the two Principal Component Analyses show that the underlying structure of L1 and L2 (dis)fluency is largely similar, with a prevalent temporal (dis)fluency dimension, and (at least) three other dimensions. The major difference between the learner and native

---

[175] As a reminder, the results of a Factor Analysis were almost identical.

[176] A small parenthesis is in order here. In the introduction to this thesis, I briefly mentioned the (dis)fluency profiles of some American politicians. As convincingly demonstrated by Tian (2016) and Liberman (2015a; 2015b; 2017), Donald Trump is characterised by a high use of false starts and a very low use of filled pauses. These researchers did not consider the use of the *and, so* and *but*, but it would seem, at first sight, that Donald Trump could also be characterised by a high use of these words (and more particularly *and*), which would arguably make him a "good" representative of the (disfluent end of the) "discourse coherence" dimension. Consider, for instance, the following example (taken from Liberman 2015b):

> Now normally, they want to make their fifteen points or their twenty points **and** then it has to go back **and** it has to be re-voted **and** everything else. I said just pass it along, **and** they said think we're going to do that. **And** let's see what happens. **But** I think they're going to do that. **And so** look, we have actually in the Republican Party, in a true sense, we have great unity. Look at the Democrats with Bernie Sanders who got absolutely taken advantage of by the DNC. **And** now see it, you know, all the stuff coming out. […] That's a massive– this will be the biggest tax cut in history. In the history of our country. **And** that's great. **And** we need it. Because right now, our country's about the highest taxed or certainly one of the highest taxed in the world. **And** we can't have that. **So** we're going to have a country that's toward the lower end.

(dis)fluency structure lies in filled pauses: learner filled pauses are, in fact, strongly associated with discourse markers in Component 3, and not associated with unfilled pauses in the temporal (dis)fluency component as is the case for native speakers. This suggests that, contrarily to a regular practice, **filled and unfilled pauses should not be conflated into one category in learner language** because they represent different dimensions of L2 (dis)fluency. The analysis also brought to light new underlying dimensions of (dis)fluency such as the pragmatic and discourse coherence dimensions that definitely merit further scrutiny in future research.

Assessed CEFR fluency levels

The last research question focused on how the learners' assessed CEFR fluency level relates with empirical measurements of (dis)fluency features.

Overall, the analyses highlighted the importance of the first dimension of L2 (dis)fluency (**temporal (dis)fluency**), and of speech rate and the frequency of unfilled pauses in particular. Two other features come into prominence as key measures of learner (dis)fluency, namely **discourse markers** and **restarts**, the other variables and dimensions being related to CEFR fluency level only to a marginal extent.

More specifically, a correlational analysis showed that the learners' CEFR fluency scores significantly correlate with three temporal (dis)fluency measures (phonation-time ratio, speech rate and number of unfilled pauses) as well as with learner's temporal (dis)fluency component. They also positively correlate with the frequency of discourse markers and the pragmatic (dis)fluency component, and negatively correlate with the frequency of restarts. Probing deeper into learners who were assessed at **B1 and C1** for fluency, it was quite striking to see that these two groups of learners only significantly differ with respect to three (dis)fluency variables: **discourse markers are more frequent at C1, speech rate is higher at C1 and unfilled pauses are fewer at C1**. The other (dis)fluency measures do not differentiate between high-intermediate B2 and advanced C1 learners. Lastly, a multiple regression analysis revealed that CEFR fluency scores can best be predicted by two predictor variables, namely the interaction between restarts and unfilled pauses, and the interaction between discourse markers and speech rate.

A word of caution

Before concluding this summary of the main findings, a word of caution is in order. Although there has been an implicit tendency to broaden the perspective of the phenomenon of (dis)fluency beyond the speakers analysed in this work, the findings reported here are based

on two corpora of learner and native speech and should not be extrapolated to all learners or to all native speakers. In particular, (dis)fluency dimensions and (dis)fluency profiles are highly dependent on the characteristics of the data they are based on. It is only after corroborative studies that such extrapolations might be envisaged. Moreover, I have tried to emphasise at several points that, despite their appeal, statistics depend on the researcher's informed decision about all kinds of options, especially when multifactorial analyses are concerned. Different choices could have led to slightly different results.

GENERAL DISCUSSION

This study is situated at the crossroads between four domains, namely (dis)fluency research, spoken corpus research, learner corpus research, and (dis)fluency testing and assessment, with special emphasis on the *Common European Framework of Reference* descriptors for spoken competence and fluency. In the following, the major contributions of this thesis to each of these areas will be discussed.

Contributions to (dis)fluency research

Disfluencies such as pauses or restarts have traditionally been seen not only as particularly pervasive in speech, but also as detrimental to fluency. Even today, fluency is still often equated with flawless (i.e. disfluency-free) performance. An important contribution of my work has thus been to **reconsider the field of (dis)fluency research in a more positive perspective**. This has firstly been done by analysing (dis)fluency as a complex interplay between a set of measures that both contribute to, and are a window on, processing and monitoring: they may either be indices of high cognitive load or functional, fluency-enhancing clues to the listener. Re-addressing the field in a more positive light has also been done by contrasting learner with native (dis)fluency. Examining the speech of learners in the light of empirical (i.e. non-idealistic) data by speakers who, although generally considered fluent, also produce all kinds of disfluencies has challenged long-held assumptions about L2 and L1 (dis)fluency.

Crucially, one of the main contributions of my work has been to address (dis)fluency in a **wide componential perspective**. This research has taken fourteen (dis)fluency measures into its scope; these have been analysed in two corpora totalling 30 hours of recorded speech produced by a hundred speakers (fifty learners and fifty native speakers). In total, about 70,000 annotations have been analysed. These figures, which by far exceed most current research, have made it possible to pen a precise quantitative picture of L1 and L2 (dis)fluency and to highlight the considerable variability between speakers.

Furthermore, an analysis of (dis)fluency in such a wide componential perspective has at its heart the question of how the variables contributing to it are interrelated, and my work has raised awareness of the **multifaceted character of (dis)fluency** by outlining underlying dimensions of (dis)fluency as well as (dis)fluency profiles. With respect to the former, several **dimensions** of learner and native (dis)fluency have been uncovered. My analyses have confirmed that a temporal component is at the core of learner and native (dis)fluency. They have, however, also demonstrated that the temporal component interacts with a repair, a pragmatic, and a discourse coherence component to create a complex network of underlying dimensions. Each of these dimensions has both a more fluent and a less fluent end, and one

of the key findings of my work is that a "fluent" performance in one dimension does not necessarily imply a "fluent" performance in the others. More specifically, in proposing **learner and native (dis)fluency profiles**, my work has also shown that individual speakers are better characterised in terms of *associations* of (dis)fluency features or dimensions than in terms of individual (dis)fluency measures.

Another important methodological outcome of my work related to the componential perspective is the **(dis)fluency annotation scheme**. Contrary to many existing annotation systems, this scheme is applicable to large datasets of both L1 and L2 data. It is language-independent and enables the annotation of a wide panel of (dis)fluency features. This annotation scheme, I believe, offers many possibilities for the analysis of the contextual occurrence of (dis)fluency features not only in LINDSEI-FR+ and LOCNEC+, but also potentially in other spoken corpora.

Finally, reviewing previous (dis)fluency studies has led to an important general **word of caution** in relation to the use of different terminologies and subtly different measures. The use of (dis)fluency terminology is sometimes treacherous and the same term should not necessarily be interpreted as covering the same phenomena. Alternatively, different terms may be used to refer to the same reality. Two cases in point are the category of restarts (sometimes also called reformulations, repairs, recasts, or false starts) and of unfilled (silent) pauses. Similarly, the issue of the measurement of frequencies and of temporal variables has emerged as one of the key aspects to be addressed in the future.

Contributions to spoken corpus research

The increasing availability of spoken corpora has marked an important turning point in (dis)fluency research. Although they have led to important improvements such as the use of larger speaker groups and speech samples or the analysis of (dis)fluency features in their context of use, spoken corpora are not necessarily the panacea for researchers wishing to embark on the (dis)fluency journey. Three aspects in particular are considered in my work.

One of the key issues in spoken corpus research relates to speaking **task characterization**. Many corpora make use of vague denominations, but, given the considerable impact of task on spoken production, the domain would benefit from in-depth reflection on how to best characterise speaking tasks. A characterization of speaking tasks that refers to concrete aspects of the communicative situation such as the number of interlocutors, the degree of naturalness, or the extent of planning time. would pave the way for more reliable comparative analyses.

A key contribution of my work is the reflection carried out on the **transcription of speech** and the effects of different displays, the availability of the original audio recordings and the **time alignment** of the transcriptions with the recordings. With respect to the latter aspect, my

study has documented a specific procedure for the time alignment of LINDSEI-FR and LOCNEC that involves both automatic and manual steps. While acknowledging the limitations of this procedure, especially in terms of time, I have been keen to stress its crucial importance to enrich the amount of information available in spoken corpora and, therefore, to offer new research perspectives. My time alignment procedure could be applied to time align other components of LINDSEI, and, more generally, hindsight gained from this experience could benefit future data collections and future undertakings to time aligning existing corpora.

Identifying (dis)fluency features accurately is one of the key concerns in spoken corpus research. To date, most (dis)fluency annotations are performed exclusively manually, but, given the increasing size of spoken corpora and the number of (dis)fluency phenomena, the question of the extent to which it is feasible to **annotate (dis)fluency features using automatic means** has become a hotbed of discussion. One of the key issues is that while some (dis)fluency features can be directly identified in the transcriptions (e.g. pauses), many escape formal identification (e.g. repairs or false starts). In my work, a happy medium between automatic and manual annotation has been found. Despite its limitations, semi-automatic annotation, I believe, offers the most flexibility for (dis)fluency annotation.

Contributions to learner corpus research

The analysis of the list of *Learner Corpora around the World*[177] and of previous corpus studies on (dis)fluency has yielded a number of observations which are of general relevance for learner corpus researchers wishing to navigate the meanders of the river (dis)fluency. One of the **dominant trends** in learner corpus (dis)fluency research is to study (dis)fluency cross-sectionally, rather than longitudinally, generally by contrasting two adjacent levels (e.g. beginner and intermediate learners) with a view to making claims about the evolution of (dis)fluency across levels. Another dominant trend consists in comparing learners with native speakers of the target language to examine the extent of the "gap" between the two.

Such contrastive interlanguage analyses have definitely led to great advances in my knowledge of L1 and L2 (dis)fluency, but they suffer from some drawbacks. One of the main issues that stood out was the **lack of data on the learners speaking in their mother tongue**. The learners' L1 and L1 speech patterns are claimed to influence learner language, but this aspect is only rarely really taken into account in LCR studies. Another key issue relates to the **proficiency level assignation**. Learner proficiency level assignation generally relies on a global assessment of the corpus or the learner group as a whole and is based on external criteria such as institutional status. Moreover, it mostly makes use of the untrustworthy

[177] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (last accessed 08/03/2018).

beginner/intermediate/advanced triad. My work rather argues in favour of the individual assessment of the language production of each individual learner.

This study has made use of learner corpus data to **characterise the (dis)fluency of French-speaking learners of English as compared to British English native speakers**, and to characterise the (dis)fluency of **B2 vs. C1** learners in particular. The study has emphasised not only the areas where L1 and L2, and B2 and C1 (dis)fluency diverge, but also where they are similar (e.g. the overall underlying structure of (dis)fluency). The comparison of L2 with L1 data (and of B2 and C1 data) has brought light on the fact that there is no simple and direct linear relationship between, on the one hand, the number of (dis)fluency features or temporal measures, and, on the other hand, fluency or disfluency. In fact, in learner as in native speech, (dis)fluency results from an interplay of factors and variables, including personal preferences.

Finally, my work has also **contributed to learner corpus methodology** in three major ways. First, the study was intent on showcasing the potential of a componential approach to L2 (dis)fluency. Contrary to many studies that are restricted to the examination of a limited number of features, fourteen (dis)fluency measures were explored here, as well as their interactions. The two highlights of this approach are the multifaceted picture of (dis)fluency (i.e. the (dis)fluency components) and the individual (dis)fluency profiles. Secondly, the learner corpus data was combined with a fluency assessment procedure of each individual learner. This additional procedure has substantially increased the value of the information contained in the learner corpus and demonstrated the shaky ground behind assessments solely based on external criteria. Lastly, whilst the field tends to over-rely on aggregate analyses of speaker groups, my work has demonstrated how it is possible to associate aggregate analyses with analyses of individual variation (especially in the Cluster Analyses).

Contributions to (dis)fluency testing and assessment

My study has made use of CEFR fluency ratings by professionally-trained raters based on excerpts of each learner's interview. It has contributed to the domain of testing and assessment by pinpointing several weaknesses involved in CEFR fluency rating.

One of the central questions when it comes to the assessment of spoken data is whether it is necessary for the raters to listen to the **complete recording**, or whether the rating can reliably be based on a **shorter excerpt**. Practical imperatives often tip the scales in favour of the latter, but research remains to be carried out to fully gauge the extent to which this practice affects subsequent ratings. A first step in this direction has been made in my work, by comparing the rated excerpt and the interview it was extracted from. Only negligible differences were found in my study, but, more generally, the issue of the generalisability of the rating is still largely an uncharted territory.

Another weakness relating to (dis)fluency rating relates to the **raters.** Raters of different "types" have previously been called upon to rate learner speech: they may be native speakers, non-native speakers, foreign language teachers, professionally-trained raters, naïve raters etc. In the hope of achieving a high level of agreement between raters, this work took the side of professionally-trained native speaker raters, but they only achieved a moderate level of agreement. This obviously led to many questions about the need for detailed rating guidelines and pre-rating training sessions, and the factors that might affect (dis)fluency rating. Several such factors have been highlighted, including the fact that raters might unconsciously be sensitive to different characteristics of learner's (dis)fluency profiles or deficiencies at the level of the CEFR fluency scales.

Contributions to the *Common European Framework of Reference*

This thesis had as one of its initial objectives to contribute to improved **CEFR descriptors for fluency**. It soon became clear that this was perhaps excessively ambitious, mainly because of the limited nature of the data, which only contains B2 and C1 learners with French as their mother tongue. Moreover, more fine-grained distinctions should have been made with respect to task. Specifically, the more constrained and monologic picture description task should probably have been looked at separately. Nonetheless, my work has contributed to the CEFR in several ways.

The thesis has offered a critical view on the CEFR descriptors for fluency and put forward several **inconsistencies** in the wording of the descriptors. One such main issue is the lack of coherence throughout the descriptor scale, with some features being mentioned at one level, but not at the preceding or following level. Another important issue is the implicit assumption that there is a linear relationship between perceived fluency and utterance fluency, with, as end-point, a C2 level which is doubtfully even achievable by native speakers.

A first important outcome of this work is the characterization of native (dis)fluency. Better insights into native (dis)fluency should allow for a better, more realistic, characterization of the top level of the fluency scale. A second important outcome of this work is the (lack of) discriminatory power of some (dis)fluency features between B2 and C1 learners. My work has shown that only a limited set of features can discriminate between these two bands, which inevitably leads to the question of whether it makes sense to distinguish six levels or whether it would be more appropriate to use another breakdown. Future analyses based on larger amounts of data and including a larger panel of proficiency levels could pave the way to gaining a clearer picture of the number of levels that should be subsumed under "fluency". Furthermore, while the CEFR fluency descriptors mention specific (dis)fluency features, the concrete use of the descriptor scales might be facilitated if (dis)fluency dimensions were exploited instead.

HIC SUNT DRACONES - AVENUES FOR FUTURE RESEARCH

Many aspects of (dis)fluency have not been discussed in this thesis, and the findings presented in the previous chapters pave the way to manifold interesting avenues for future research.

I would like to start by briefly listing three research avenues that I originally intended to include in this PhD. One of my original objectives after having time aligned LINDSEI-FR+ and LOCNEC+ was to study **unfilled pauses**, and, more precisely, to examine the extent to which they were perceived and transcribed, the relationship between their perceived and actual length depending on their context etc. Also, empirical measurements in LOCNEC+ show that there are important differences between native speakers. I originally intended to have not only the learners', but also the **native speakers' (dis)fluency rated** according to a (dis)fluency scale. Furthermore, I also planned to relate (dis)fluency measurements and profiles to speakers' **metadata**, such as the time spent abroad. As is often the case in a PhD project, objectives have to be reconsidered, and these aspects fell in the scope of worthwhile avenues for future research instead.

Besides those research avenues, future work may wish to carry out more rigorous analyses of how the **measurement** of (dis)fluency variables affects research findings. As became apparent through this thesis, there is no clear agreement on the most accurate way to measure the frequency of (dis)fluency features: is it per hundred words? Per minute? Using pruned words/time? Or unpruned words/time? Additionally, the issue of the lower threshold for unfilled pauses has far-reaching effects because it affects all measures of temporal (dis)fluency, which are, however, at the heart of L1 and L2 (dis)fluency. My work so far (Dumont 2017a) has only attempted to broach the subject superficially, but has nonetheless shown that different measurements do lead to different results. Further studies in this direction are more than needed.

Another promising avenue for future research is the contrastive analysis of the **three speaking tasks** that make up LINDSEI-FR and LOCNEC interviews. Previous research has indicated that speaking task affects learner and native speaker (dis)fluency, and highly relevant insights could be gained from probing further into the differential use of (dis)fluency features in monologic vs. dialogic speech, and in more vs. less constrained speaking tasks. My work so far has examined the extent to which learners and native speakers are affected similarly by task in LINDSEI-FR+ and LOCNEC+ (Dumont 2017b). This study revealed that speaking task affects the frequency of unfilled pauses, filled pauses, false starts, and speech rate, and that it affects learner and native speech in a similar way. By contrast, task does not affect the frequency of restarts in either LINDSEI-FR+ or LOCNEC+. As for the mean length of runs, it differs depending on the task in the native corpus, but does not vary in learner speech.

The issue of the **influence of the learners' mother tongue** on L2 (dis)fluency remains largely unexplored. The data did not allow for comparisons between the learners' speech patterns in their mother tongue and in L2 English, but future studies, turning impending data collections to good advantage, could explore the nature of the relationship between a speaker's (dis)fluency and (dis)fluency profile in their mother tongue and in a foreign language. Detailed comparative analyses of how L1 speech patterns shape L2 performance are necessary to help distinguish which performance features are idiosyncratic and which are due to a speaker talking in a foreign language. Additionally, there are still few **developmental analyses** of L2 (dis)fluency across proficiency levels. Using longitudinal corpus data to empirically trace how (dis)fluency changes, increases, or potentially stabilises across the proficiency continuum constitutes a truly worthwhile enterprise.

Yet another avenue for future research consists in the exploration of **chunks of disfluencies**, i.e. sequences of adjacent (dis)fluency features such as the chunk "*and er* + unfilled pause" or "*well er*". Preliminary attempts in this direction have already been conducted in the frame of Dumont (2016) and Crible *et al.* (2017) (see also Crible 2017a; Crible, Degand & Gilquin 2017). Closer examinations of (dis)fluency features used in chunks might not only contribute to our knowledge of their uses and functions, but also bring further insights into some (dis)fluency components (e.g. the association of filled pauses and discourse markers in the learner pragmatic (dis)fluency component).

Crucially, I also believe that the approach taken in this study should be complemented by a study of L1 and L2 **accuracy and complexity**. Considering (dis)fluency on its own may be misleading to some extent if other areas of proficiency are totally disregarded. In particular, potential trade-off effects between fluency, accuracy, and complexity could be examined in both learner and native speech. Relevant complementary information for the interpretation of (dis)fluency profiles may thus be found in other areas of language performance.

Another avenue worth pursuing involves trying out alternative **statistical tests**. My study has made use of a battery of multivariate statistical techniques, such as Principal Component Analysis, Cluster Analysis, or Multiple Regression Analysis. For each of these tests, alternative options could have been justified. For example, other distance measures in the cluster analysis could prove more efficient in distinguishing (dis)fluency profiles. Moreover, relationships between variables have been assumed to be linear, but there is some evidence that this relationship might in fact be cubic or quadratic.

The question **whether fluency can be taught** is complex and has not been addressed in this thesis. In the light of my findings, future work could critically examine what is (or is not) done in terms of fluency teaching, the types of fluency-oriented exercises that are generally advised, the type of audio material that is used (are they authentic recordings containing disfluencies?) etc. More generally, the issue of the "teachability" of fluency is related to the possibility of practising speaking skills in and out of the classrooms, and interesting ideas might, for example, be found in the use of the new technologies (Sweetlove *et al.* 2015). Also,

research indicates that benefits in terms of fluency might be gained by practising skills that, at first sight, do not seem to be related at all to spoken fluency, namely online messaging and written chats (*cf.* e.g. Bataineh 2014; Blake 2006; Sykes 2013).

As underlined at the beginning of this thesis, (dis)fluency is situated at the crossroads between many research fields. Although I did not take on board research findings from other perspectives, important advances could be made in our understanding of L1 and L2 (dis)fluency by adopting a more **multidisciplinary approach**, and by including insights from (or collaborating with researchers from) neurolinguistics, speech pathology research, computational linguistics, or written fluency research.

REFERENCES

Aas, Hege Larsson & Susan Nacey. 2017. Investigating fluency variables in learner language. Methodological concerns. Bolzano. https://susannacey.hihm.no/wp-content/uploads/2017/09/Presentation-LCR2-2017-Aas-and-Nacey.pdf (7 October, 2017).

Abdel Latif, Muhammad M. Mahmoud. 2013. What Do We Mean by Writing Fluency and How Can It Be Validly Measured? *Applied Linguistics* 34(1). 99–105. doi:10.1093/applin/ams073.

Adolphs, Svenja & Ronald Carter. 2013. *Spoken Corpus Linguistics: From Monomodal to Multimodal*. (Ed.) Tony McEnery & Michael Hoey. (Routledge Advances in Corpus Linguistics). Routledge. (11 October, 2013).

Ahmadian, Mohammad Javad & Mansoor Tavakoli. 2011. The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research* 15(1). 35–59. doi:10.1177/1362168810383329.

Aijmer, Karin. 1997. "I think" - an English Modal Particle. In Toril Swan & Olaf J. Westvik (eds.), *Modality in Germanic Languages. Historical and Comparative Perspectives.*, 1–47. Berlin: Mouton de Gruyter.

Aijmer, Karin. 2004. Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies* 3(1). 173–190.

Aijmer, Karin. 2011. Well I'm not sure I think... The use of well by non-native speakers. *International Journal of Corpus Linguistics* 16(2). 231–254. doi:10.1075/ijcl.16.2.04aij.

Aijmer, Karin & Anne-Marie Simon-Vandenbergen. 2003. The discourse particle well and its equivalents in Swedish and Dutch. *Linguistics* 41(6). 1123–1161.

Aitchison, Jean. 1989. *The Articulate Mammal: An Introduction to Psycholinguistics*. 3rd. ed. London: Unwin Hyman.

Aldenderfer, Mark S. & Roger K. Blashfield. 1984. *Cluster analysis*. (Sage University Papers Series 44). Newbury Park (Calif.): Sage.

Alderson, J. Charles. 1996. Do corpora have a role in language assessment? In J.A. Thomas & M.H. Short (eds.), *Using Corpora for Language Research*, 248–259. London: Longman.

Alderson, J. Charles. 2007. The CEFR and the need for more research. *The Modern Language Journal* 91(4). 659–663.

Alderson, J. Charles, Caroline Clapham & Dianne Wall. 2001. *Language test construction and evaluation*. 5th print. (Cambridge Language Teaching Library). Cambridge: Cambridge university press.

Allwood, Jens, Joakim Nivre & Elisabeth Ahlsén. 1990. Speech management - On the non-written life of speech. *Nordic Journal of Linguistics* 13(01). 3–48.

Altenberg, Bengt. 1987. Causal ordering strategies in English conversation. In James Monaghan (ed.), *Grammar in the construction of texts*, 50–64. London: Frances Pinter.

Andersen, Gisle. 2016. Semi-lexical features in corpus transcription: Consistency, comparability, standardisation. *International Journal of Corpus Linguistics* 21(3). 323–347. doi:10.1075/ijcl.21.3.02and.

Anderson-Hsieh, Janet & Kenneth Koehler. 1988. The effect of foreign accent and speaking rate on native speaker comprehension. *Language learning* 38(4). 561–613.

André, Virginie & Henry Tyne. 2010. Compétence sociolinguistique et dysfluence en L2. *Recherches récentes en FLE*. Bern: Peter Lang. https://hal.archives-ouvertes.fr/hal-00521112v1/document (22 January, 2015).

Arlington, Jody, Sebastian M. Brenninkmeyer, Danielle Arn, Rita Grundhauser & Daniel C. O'Connell. 1992. A usual extreme case: Pause reports of informal spontaneous dialogue. *Bulletin of the Psychonomic Society* 30(2). 161–163.

Arlot, Sylvain & Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* 4. 40–79. doi:10.1214/09-SS054.

Arnold, Jennifer E., Maria Fagnano & Michael K. Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of psycholinguistic research* 32(1). 25–36.

Arnold, Jennifer E., Michael K. Tanenhaus, R. J. Altman & Maria Fagnano. 2004. The old and thee, uh, new: disfluency and reference resolution. *Psychological Science* 15. 578–582.

Arnold, Jennifer, Carla L. Hudson Kam, Michael Tanenhaus, Carolina Chapel, Hill Carla, L. Hudson Kam & Psychology Department. 2007. If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension. 914–930. MIT Press.

Artstein, Ron & Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596.

Atkins, Sue, Jeremy Clear & Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7(1). 1–16. doi:10.1093/llc/7.1.1.

Atwell, Eric, Peter Howarth & Clive Souter. 2003. The ISLE Corpus: Italian and German Spoken Learner's English. *ICAME Journal* 27. 5–18.

Austin, John L. 1965. *How to do things with words*. New York: Oxford University Press.

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. 1st ed. Cambridge, UK ; New York: Cambridge University Press.

Bachman, Lyle F. 1990. *Fundamental Considerations in Language Testing*. Oxford: OUP.

Bachman, Lyle & Adrian Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

Baker, John Paul. 1997. Correcting automatically tagged data. In Roger Garside, Geoffrey Leech & Anthony McEnery (eds.), *Corpus annotation: Linguistic information from computer text corpora*, 243–250. Longman.

Baker-Smemoe, Wendy, Dan Dewey, Jennifer Bown & Rob Martinsen. 2014. Does Measuring L2 Utterance Fluency Equal Measuring Overall L2 Proficiency? Evidence From Five Languages. *Foreign Language Annals* 47(4). 707–728. doi:10.1111/flan.12110.

Ballier, Nicolas & Philippe Martin. 2015. Speech annotation of learner corpora. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, 107–134. (Cambridge Handbooks in Language and Linguistics). Cambridge: Cambridge University Press.

Barker, Fiona. 2010. How can corpora be used in language testing? In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*, 633–645. (Routledge Handbooks in Applied Linguistics). Abingdon: Routledge.

Barker, Fiona. 2013. Corpus-Based Testing. In Carol A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*, 1360–1366. Blackwell Publishing Ltd. doi:10.1002/9781405198431.wbeal0263. http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0263/abstract (16 November, 2017).

Barlow, Michael, Rod Ellis & Gary Barkhuizen. 2005. Computer-based analyses of learner language. *Analysing Learner Language*, 335–357. Oxford: Oxford University Press. http://michaelbarlow.com/chapter14.pdf.

Barr, Dale. 2001. Trouble in Mind: Paralinguistic Indices of Effort and Uncertainty in Communication. *Oralité et gestualité, communication multimodal, intéraction*. L'Harmattan. Paris. http://www.psy.gla.ac.uk/docs/download.php?type=PUBLS&id=1584 (29 October, 2013).

Bataineh, Ahmad Mousa. 2014. The Effect of Using Audiovisual Chat on Developing English as a Foreign Language Learners' Fluency and Productivity of Authentic Oral Texts. *International Journal of Linguistics* 6(3). 85–108. doi:10.5296/ijl.v6i3.5563.

Bavelas, Janet B. 2000. Nonverbal aspects of fluency. In Heidi Riggenbach (ed.), *Perspectives on fluency*, 91–127. Ann Arbor: University of Michigan Press.

Bear, John, John Dowding, Elizabeth Shriberg & Patti Price. 1993. *A system for labeling self-repairs in speech*.

Beattie, Geoffrey W. & Brian L. Butterworth. 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* 22(3). 201–211.

Beaufort, Richard & Alain Ruelle. 2006. eLite: système de synthèse de la parole à orientation linguistique. *Proceedings of JEP*. 509–512.

Beliao, Julie & Anne Lacheret. 2013. Disfluency and discursive markers: when prosody and syntax plan discourse. *DiSS 2013: The 6th Workshop on Disfluency in Spontaneous Speech*, vol. 54, 5–9. http://hal.archives-ouvertes.fr/halshs-00869849/ (8 May, 2014).

Bell, Célia D.S. 2003. L2 speech rate in monologic and dialogic activities. *Linguagem & Ensino* 6(2). 55–79.

Besser, Jana. 2006. A Corpus-Based Approach to the Classification and Correction of Disfluencies in Spontaneous Speech. Saarbrücken: Saarland University Bachelor Thesis. http://ami.dfki.de/pdf/besserBachelor06.pdf (15 December, 2014).

Besser, Jana & Jan Alexandersson. 2007. A comprehensive disfluency model for multi-party interaction. *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, 182–189. Antwerp: Association for Computational Linguistics. http://sigdial.org/workshops/workshop8/Proceedings/SIGdial31.pdf (15 December, 2014).

Bestgen, Yves. 1998. Segmentation markers as trace and signal of discourse structure. *Journal of Pragmatics* 29(6). 753–763.

Biber, Douglas. 1988. *Variation across speech and writing*. London: Cambridge university press.

Biber, Douglas, Stig Johansson, Geoffrey Neil Leech & Randolph Quirk (eds.). 1999. *Longman grammar of spoken and written English*. 4th impr. Harlow: Pearson education.

Bigi, Brigitte. 2015. SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician, International Society of Phonetic Sciences* 111–112. 54–69.

Bilá, Magdaléna & Anna Džambová. 2011. A Preliminary Study on Silent Pauses in L1 and L2 Speakers of English and German. *Brno Studies in English* 37(1). 109–118.

Blackmer, E R & J L Mitton. 1991. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition* 39(3). 173–194.

Blake, Christopher Grant. 2006. The potential of text-based Internet chats for improving ESL oral fluency. http://docs.lib.purdue.edu/dissertations/AAI3239774.

Blakemore, Diane. 1988. "So" as a constraint on relevance. In R.M. Kempson (ed.), *Mental representations: The interface between language and reality*, 183–195. New York: Cambridge University Press.

Blakemore, Diane. 2002. *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*. New York: Cambridge University Press. http://assets.cambridge.org/052164/0075/sample/0521640075WS.pdf (1 March, 2018).

Blanche-Benveniste, Claire. 1997. *Approches de la langue parlée en français*. Editions OPHRYS.

Blanche-Benveniste, Claire. 2000. Transcription de l'oral et morphologie. In M. Gille & R. Kiesler (eds.), *Romania Una et diversa, Phililigische Studien für Theodor Berchem*, 61–74. Tübingen: Gunter Narr Verlag.

Blanche-Benveniste, Claire, Colette Jeanjean & Jacques Monfrin. 1987. *Le français parlé: transcription et édition*. (CNRS. Institut national de la langue française. Publications du Trésor général des langues et parlers français). Paris: Didier.

Bley-Vroman, Robert. 1983. The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning* 33(1). 1–17. doi:10.1111/j.1467-1770.1983.tb00983.x.

Boersma, Paul & David Weenink. 2013. *Praat: Doing phonetics by computer [Computer program]*. WindowsEnglish.

Bolden, Galina B. 2006. Little Words That Matter: Discourse Markers "So" and "Oh" and the Doing of Other-Attentiveness in Social Interaction. *Journal of Communication* 56(4). 661–688. doi:10.1111/j.1460-2466.2006.00314.x.

Bolden, Galina B. 2009. Implementing incipient actions: The discourse marker 'so' in English conversation. *Journal of Pragmatics* 41(5). 974–998. doi:10.1016/j.pragma.2008.10.004.

Boomer, Donald S. & Allen T. Dittmann. 1962. Hesitation Pauses and Juncture Pauses in Speech. *Language and Speech* 5(4). 215–220.

Bortfeld, Heather, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober & Susan E. Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech* 44(2). 123–147.

Bosker, Hans R. & Eva Reinisch. 2015. Normalization for speech rate in native and nonnative speech. In M. Wolters, B. Livingstone, B. Beattie, R. Smith, M. MacMahon, J. Stuart-Smith & J. Scobbie (eds.), *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*, online proceedings. London: International Phonetic Association. http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:2193307 (26 January, 2017).

Bosker, Hans Rutger. 2014. The processing and evaluation of fluency in native and non-native speech. Utrecht: LOT, Netherlands Graduate School PhD Thesis.

Bosker, Hans Rutger, Anne-France Pinget, Hugo Quené, Ted Sanders & Nivja De Jong. 2013. What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing* 30(2). 159–175. doi:10.1177/0265532212455394.

Bosker, Hans Rutger, Hugo Quené, Ted Sanders & Nivja H. de Jong. 2014. The Perception of Fluency in Native and Non-native Speech. *Language Learning* 64(3). 579–614. doi:10.1111/lang.12067.

Bot, Kees de. 1992. A bilingual production model: Levelt's "speaking" model adapted. *Applied Linguistics* 13(1). 1–24. doi:10.1093/applin/13.1.1.

Brand, Christiane & Sandra Götz. 2011. Fluency versus accuracy in advanced spoken learner language: A multi-method approach. *International Journal of Corpus Linguistics* 16(2). 255–275. doi:10.1075/ijcl.16.2.05bra.

Brennan, Susan E. & Michael F. Schober. 2001. How Listeners Compensate for Disfluencies. *Journal of Memory and Language* 44. 274–296.

Broek, Simon & Inge van den Enden. 2013. The implemention of the Common European Framework for Languages in European education systems. http://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/495871/IPOL-CULT_ET(2013)495871_EN.pdf (3 November, 2016).

Brognaux, Sandrine, Sophie Roekhaut, Thomas Drugman & Richard Beaufort. 2012a. Automatic Phone Alignment. http://tcts.fpms.ac.be/~drugman/files/JAPTAL12.pdf (23 December, 2013).

Brognaux, Sandrine, Sophie Roekhaut, Thomas Drugman & Richard Beaufort. 2012b. Train&align: A new online tool for automatic phonetic alignment. *Spoken Language*

Technology Workshop (SLT), 2012 IEEE, 416–421. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6424260 (23 December, 2013).

Brumfit, Christopher. 1984. *Communicative methodology in language teaching: The roles of fluency and accuracy*. Cambridge: Cambridge University Press.

Butterworth, Brian L. 1980. Evidence from pauses in speech. *Language Production. Speech and talk*, 155–176. (1). London: London Academic Press. https://www.researchgate.net/publication/238229597_Evidence_from_pauses_in_speech (14 June, 2016).

Buysse, Lieven. 2007. Discourse marker so in the English of Flemish university students. *Belgian Journal of English Language and Literature (BELL)* 5(Thematic Issue). 79–95.

Buysse, Lieven. 2009. So as a marker of elaboration in native and non-native speech. http://www.academia.edu/3194496/So_as_a_marker_of_elaboration_in_native_and_non-native_speech (22 August, 2013).

Buysse, Lieven. 2010. Discourse Markers in the English of Flemish University Students. In Iwona Witzcak-Plisiecka (ed.), *Pragmatic perspectives on language and linguistics. Vol. 1: Speech actions in theory and applied studies*, 461–484. Newcastle upon Tyne: Cambridge Scholars Publishing. http://www.academia.edu/3194494/Discourse_Marker_So_in_Native_and_Non-Native_Spoken_English (22 August, 2013).

Buysse, Lieven. 2012. So as a multifunctional discourse marker in native and learner speech. *Journal of Pragmatics* 44(13). 1764–1782. doi:10.1016/j.pragma.2012.08.012.

Buysse, Lieven. 2014. "So what's a year in a lifetime so." Non-prefatory use of so in native and learner English. *Text & Talk - An Interdisciplinary Journal of Language Discourse Communication Studies* 34(1). 23–47. doi:10.1515/text-2013-0036.

Buysse, Lieven. 2015. 'Well it's not very ideal …' The pragmatic marker well in learner English. *Intercultural Pragmatics* 12(1). 59–89. doi:10.1515/ip-2015-0003.

Callies, Marcus, Maria Belen Díez-Bedmar & Ekaterina Zaytseva. 2014. Using learner corpora for testing and assessing L2 proficiency. In Pascale Leclercq, Amanda Edmonds & Heather Hilton (eds.), *Measuring L2 proficiency: Perspectives from SLA*, 71–90. (Second Language Acquisition Series). Clevedon: Multilingual Matters.

Callies, Marcus & Sandra Götz. 2015a. Learner corpora in language testing and assessment: Prospects and challenges. In Marcus Callies & Sandra Götz (eds.), *Learner corpora in language testing and assessment*, 1–9. Amsterdam: John Benjamins. http://www.academia.edu/10267452/Callies_Marcus_and_Sandra_G%C3%B6tz_2015_Learner_corpora_in_language_testing_and_assessment_Prospects_and_challenges (28 May, 2015).

Callies, Marcus & Sandra Götz. 2015b. *Learner Corpora in Language Testing and Assessment*. John Benjamins Publishing Company.

Cameron, Deborah. 2001. *Working with Spoken Discourse*. 1 edition. London ; Thousand Oaks: SAGE Publications Ltd.

Campillos Llanos, Leonardo & Paula González Gómez. 2014. Oral Production of Discourse Markers by Intermediate Learners of Spanish: A Corpus Perspective. In Jesús Romero-

Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2014: New Empirical and Theoretical Paradigms*, vol. 2, 239–259. Cham: Springer. doi:10.1007/978-3-319-06007-1_11. http://link.springer.com/10.1007/978-3-319-06007-1_11 (18 August, 2017).

Campione, Estelle & Jean Véronis. 2002. A large-scale multilingual study of silent pause duration. *Proceedings of the Speech Prosody 2002 Conference*, 199–202. Aix-en-Provence. http://sprosig.isle.illinois.edu/sp2002/pdf/campione-veronis.pdf (26 April, 2016).

Campione, Estelle & Jean Véronis. 2004. Pauses et hésitations en français spontané. *Actes des 25èmes Journées d'Études sur la Parole (JEP), Fès, Maroc*. 109–112.

Campione, Estelle & Jean Véronis. 2005. Pauses and hesitations in French spontaneous speech. *Disfluency in Spontaneous Speech*. http://www.isca-speech.org/archive_open/diss_05/dis5_043.html (9 September, 2015).

Campoy, Mari Carmen & María José Luzón (eds.). 2007. *Spoken Corpora in Applied Linguistics*. . Vol. 51. (Linguistic Insights). Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien: Peter Lang. http://www.peterlang.com/index.cfm?event=cmp.ccc.seitenstruktur.detailseiten&seitentyp=produkt&pk=13820 (11 February, 2015).

Candea, Maria. 2000. Contribution à l'étude des pauses silencieuses et des phénomènes dits d'hésitation en français oral spontané. Etude sur un corpus de récits en classe de français. Université de la Sorbonne nouvelle-Paris III. http://halshs.archives-ouvertes.fr/tel-00290143/ (11 February, 2014).

Carlsen, Cecilie. 2009. Proficiency levels in learner corpora – a source of error or an asset in SLA-research. Paper presented at the GURT 2009, Georgetown University.

Carlsen, Cecilie. 2012. Proficiency Level - a Fuzzy Variable in Computer Learner Corpora. *Applied Linguistics* 33(2). 161–183. doi:10.1093/applin/amr047.

Carter, Ronald & Michael McCarthy. 2006. *Cambridge grammar of English: a comprehensive guide: spoken and written English, grammar and usage*. 5th print. Cambridge: Cambridge university press.

Chafe, Wallace. 1980. Some reasons for hesitating. *Temporal Variables in Speech*, 168–180. Den Haag: Mouton de Gruyter.

Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen (ed.), *Spoken and Written Language: Exploring Orality and Literacy*, 35–53. Norwood: Ablex.

Chalhoub-Deville, Micheline. 1995. A Contextualized Approach to Describing Oral Language Proficiency. *Language Learning* 45(2). 251–281. doi:10.1111/j.1467-1770.1995.tb00440.x.

Chambers, Francine. 1997. What do we mean by fluency? *System* 25(4). 535–544. doi:10.1016/S0346-251X(97)00046-8.

Chapelle, Carol A. & Lia Plakans. 2013. Assessment and testing: Overview. In Carol A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*, 241–244. New York: Wiley-Blackwell.

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau & Azam Niknafs. 2014. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software* 61(6). 1–36.

Chenoweth, N. Ann & John R. Hayes. 2001. Fluency in Writing: Generating Text in L1 and L2. *Written Communication* 18(1). 80–98. doi:10.1177/0741088301018001004.

Christenfeld, Nicholas & Beth Creager. 1996. Anxiety, alcohol, aphasia, and ums. *Journal of Personality and Social Psychology* 70(3). 451–460. doi:10.1037/0022-3514.70.3.451.

Clark, Herbert H. & Eve V. Clark. 1977. *Psychology and language: an introduction to psycholinguistics*. New York (N.Y.): Harcourt, Brace and Jovanovich.

Clark, Herbert H. & Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition* 84(1). 73–111. doi:10.1016/S0010-0277(02)00017-3.

Clark, Herbert H. & Thomas Wasow. 1998. Repeating Words in Spontaneous Speech. http://citeseer.uark.edu:8080/citeseerx/viewdoc/summary?doi=10.1.1.130.7663.

Cohen, Jacob. 1977. The t Test for Means. *Statistical Power Analysis for the Behavioral Sciences (Revised Edition)*, 19–74. Academic Press. http://www.sciencedirect.com/science/article/pii/B9780121790608500074 (25 October, 2016).

Corley, Martin. 2010. Making predictions from speech with repairs: Evidence from eye movements. *Language and Cognitive Processes* 25(5). 706–727. doi:10.1080/01690960903512489.

Corley, Martin & Robert J. Hartsuiker. 2003. Hesitation in speech can... um... help a listener understand. *Proceedings of the 25th Meeting of the Cognitive Science Society*, 276–281. http://csjarchive.cogsci.rpi.edu/Proceedings/2003/pdfs/70.pdf (25 September, 2013).

Corley, Martin, Lucy J. MacGregor & David I. Donaldson. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition* 105(3). 658–668. doi:10.1016/j.cognition.2006.10.010.

Corley, Martin & Oliver W. Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass* 2(4). 589–602.

Cosme, Christelle. 2007. Clause Linking across Languages. A corpus-based study of coordination and subordination in English, French and Dutch. Louvain-la-Neuve: Université catholique de Louvain PhD thesis.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. 3rd printing. Cambridge: Cambridge University Press.

Council of Europe (ed.). 2017. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion volume with new descriptors*. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/168074a4e2 (10 October, 2017).

Coupland, Douglas. 1995. *Microserfs*. London: Flamingo.

Cox, Troy & Wendy Baker-Smemoe. 2013. The Relationship between L1 Fluency and L2 Fluency across different Proficiency. Conference Presentation. Paper presented at the

AAAL2013, Dallas, Texas. http://troycox.byu.edu/Conference_Presentations_files/FluencyAAAL.pdf.

Craggs, Richard & Mary McGee Wood. 2005. Evaluating Discourse and Dialogue Coding Schemes. *Computational Linguistics* 31(3). 289–296. doi:10.1162/089120105774321109.

Crawley, Michael J. 2007. *The R book*. Chichester: John Wiley.

Crible, Ludivine. forthcoming. Emplois sous-spécifiés des marqueurs discursifs et / and : stratégie (inter)subjective et variation en genre. *Cahiers du FoReLL*. https://dial.uclouvain.be/pr/boreal/object/boreal:192384 (14 March, 2018).

Crible, Ludivine. 2017a. Discourse markers and (dis)fluency across registers: A contrastive usage-based study in English and French. Louvain-la-Neuve: Université catholique de Louvain Doctoral Dissertation.

Crible, Ludivine. 2017b. Discourse markers and (dis)fluency in English and French: Variation and combination in the DisFrEn corpus. *International Journal of Corpus Linguistics* 22(2). 242–269. doi:10.1075/ijcl.22.2.04cri.

Crible, Ludivine. 2018. *Discourse Markers and (Dis)fluency. Forms and Functions across Languages and Registers*. (Pragmatics and Beyond New Series). Amsterdam: John Benjamins. https://dial.uclouvain.be/pr/boreal/object/boreal:192383 (14 March, 2018).

Crible, Ludivine & Maria Josep Cuenca. 2018. Co-occurrence of discourse markers: from juxtaposition to composition. https://dial.uclouvain.be/pr/boreal/object/boreal:193573 (14 March, 2018).

Crible, Ludivine, Liesbeth Degand & Gaëtanelle Gilquin. 2017. The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency. *Languages in Contrast : international journal for contrastive linguistics* 17(1). 69–95. doi:10.1075/lic.17.1.04cri.

Crible, Ludivine, Amandine Dumont, Iulia Grosman & Ingrid Notarrigo. 2014. *Set d'étiquettes des données et métadonnées*. Internal report. Louvain-la-Neuve (Belgium): Université catholique de Louvain.

Crible, Ludivine, Amandine Dumont, Iulia Grosman & Ingrid Notarrigo. 2015a. *Annotation des marqueurs de fluence et disfluence dans des corpus multilingues et multimodaux, natifs et non natifs*. Louvain-la-Neuve: Université catholique de Louvain.

Crible, Ludivine, Amandine Dumont, Iulia Grosman & Ingrid Notarrigo. 2015b. One protocol to rule them all. Paper presented at the ARC workshop, Louvain-la-Neuve (Belgium).

Crible, Ludivine, Amandine Dumont, Iulia Grosman & Ingrid Notarrigo. 2017. (Dis)fluency across spoken and signed languages. Paper presented at the (Dis)fluency conference, Louvain-la-Neuve.

Crookes, Graham. 1989. Planning and interlanguage variation. *Studies in second language acquisition* 11(4). 367–383.

Crookes, Graham. 1990. The utterance, and other basic units for second language discourse analysis. *Applied Linguistics* 11(2). 183–199. doi:10.1093/applin/11.2.183.

Crystal, David. 1988. Another look at, well, you know … *English Today*(13). 47–49.

Cucchiarini, Catia, Joost van Doremalen & Helmer Strik. 2010. Fluency in non-native read and spontaneous speech. *DiSS-LPSS*, 15–18. https://www.researchgate.net/profile/Helmer_Strik/publication/228734985_Fluency_in_non-native_read_and_spontaneous_speech/links/00b7d51950900 6d72e000000.pdf (8 March, 2016).

Cucchiarini, Catia, Helmer Strik & Lou Boves. 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America* 107(2). 989–999. doi:10.1121/1.428279.

Cucchiarini, Catia, Helmer Strik & Lou Boves. 2002. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America* 111(6). 2862–2873. doi:10.1121/1.1471894.

Cuenca, Maria Josep. 2013. The fuzzy boundaries between discourse marking and modal marking. In Liesbeth Degand, Bert Cornillie & Paola Pietrandrea (eds.), *Discourse markers and modal particles. Categorization and description*, 287–294. Amsterdam: John Benjamins.

Cumming, Alister. 2007. Book reviews: Lumley, T. 2005: Assessing second language writing: the rater's perspective. Frankfurt: Peter Lang (Volume 3, Language Testing and Evaluation Series, edited by Rudiger Grotjahn and Gunther Sigott). 368 pp. ISBN 3-631-53327-6 US-ISBN 0-8204-7655-2 US$62.95. *Language Testing* 24(2). 287–291. doi:10.1177/0265532207076366.

Davidson, Fred & Glenn Fulcher. 2007. The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching* 40(03). 231. doi:10.1017/S0261444807004351.

Davies, Alan. 2003. *The Native Speaker: Myth and Reality*. Multilingual Matters.

Davies, Alan, Annie Brown, Cathie Elder, Kathryn Hill, Tom Lumley & Tim McNamara. 1999. *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

De Cock, Sylvie. 2003. Recurrent sequences of words in native speaker and advanced learner spoken and written English. A Corpus-driven approach. Louvain-la-Neuve, Belgium: Université catholique de Louvain.

De Cock, Sylvie. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literature (BELL)* 2. 225–246.

De Cock, Sylvie. 2015a. An exploration of the use of foreign words in interviews with EFL learners: a(n) (effective) communication strategy? Nijmegen. http://dial.uclouvain.be/pr/boreal/object/boreal:167288 (24 November, 2016).

De Cock, Sylvie. 2015b. Foreign words in interviews with EFL learners: bridging lexical gaps? Trier. https://dial.uclouvain.be/pr/boreal/fr/object/boreal%3A162359 (24 January, 2018).

De Cock, Sylvie. 2017a. Fluency and the use of foreign words in interviews with EFL learners. Louvain-la-Neuve. https://dial.uclouvain.be/pr/boreal/object/boreal:187470 (24 January, 2018).

De Cock, Sylvie. 2017b. "Speaking in tongues": EFL learners' use of "foreign words" in informal interviews. Bolzano. https://dial.uclouvain.be/pr/boreal/fr/object/boreal%3A188024 (24 January, 2018).

De Gaulmyn, M.-M. 1987. Actes de reformulation et processus de reformulation. In Pierre Bange (ed.), *L'Analyse des Interactions Verbales. La Dame de Caluire: Une Consultation*, 83–98. Bern: Peter Lang.

De Jong, Nivja. 2016. Fluency in second language assessment. In Dina Tsagari & Jayanti Banerjee (eds.), *Handbook of Second Language Assessment*, 203–218. Berlin, Boston: Mouton de Gruyter. https://www.researchgate.net/publication/301296350_Fluency_in_second_language_assessment (5 September, 2016).

De Jong, Nivja H. 2016. Predicting pauses in L1 and L2 speech: the effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching* 54(2). 113–132. doi:10.1515/iral-2016-9993.

De Jong, Nivja H. & Hans Rutger Bosker. 2013. Choosing a threshold for silent pauses to measure second language fluency. In Robert Eklund (ed.), *Proceedings of Disfluency in Spontaneous Speech, DiSS 2013*, vol. TMH-QPSR 54(1), 17–20. Stockholm. http://www.ida.liu.se/~g-robek/conferences/diss2013/pdf/DiSS2013_Proceedings.pdf#page=25 (29 November, 2013).

De Jong, Nivja H., Rachel Groenhout, Rob Schoonen & Jan H. Hulstijn. 2015. Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics* 36(2). 223–243. doi:10.1017/S0142716413000210.

De Jong, Nivja H., Rob Schoonen & Jan H. Hulstijn. 2009. Fluency in L2 is related to fluency in L1. Utrecht.

De Jong, Nivja H., Margarita P. Steinel, Arjen Florijn, Rob Schoonen & Jan H. Hulstijn. 2012a. Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics* 34(05). 893–916. doi:10.1017/S0142716412000069.

De Jong, Nivja H., Margarita P. Steinel, Arjen Florijn, Rob Schoonen & Jan H. Hulstijn. 2012b. The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In Alex Housen, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA*, 123–142. (Language Learning & Language Teaching 32). Amsterdam & Philadelphia: John Benjamins Publishing Company.

De Jong, Nivja H. & Ton Wempe. 2007. Automatic measurement of speech rate in spoken Dutch. *ACLC Working Papers* 2(2). 49–58.

De Jong, Nivja H. & Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41(2). 385–390. doi:10.3758/BRM.41.2.385.

Dechert, Hans-Wilhelm & Manfred Raupach. 1980. *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*. Mouton de Gruyter.

Degand, Liesbeth. 2000. Causal connectives or causal prepositions? Discursive constraints. *Journal of Pragmatics* 32(6). 687–707. doi:10.1016/S0378-2166(99)00066-1.

Degand, Liesbeth & Anne Catherine Simon. 2005. Minimal Discourse Units: Can we define them, and why should we. *Proceedings of SEM-05. Connectors, discourse framing and discourse structure: from corpus-based and experimental analyses to discourse theories, Biarritz*. 14–15.

Degand, Liesbeth & Anne-Catherine Simon. 2009. On identifying basic discourse units in speech: theoretical and empirical issues. *Discours*(4). doi:10.4000/discours.5852. http://discours.revues.org/5852 (10 November, 2016).

Delais-Roussarie, Elisabeth & Brechtje Post. 2014. Corpus annotation. In Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.), *The Oxford Handbook of Corpus Phonology*, 46–88. Oxford University Press.

Delais-Roussarie, Elisabeth & Hiyon Yoo. 2010a. COREIL, un corpus pour l'étude de l'acquisition de la prosodie en Français et Anglais Langue Etrangère. https://halshs.archives-ouvertes.fr/halshs-00751096 (17 March, 2017).

Delais-Roussarie, Elisabeth & Hiyon Yoo. 2010b. The COREIL corpus: a learner corpus designed for studying phrasal phonology and intonation. *Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech*. Poznań, Poland.

Denke, Annika. 2009. *Nativelike performance: a corpus study of pragmatic markers, repair and repetition in native and non-native English speech*. Saarbrücken: Dr. Müller.

Dér, Csilla Ilona & Alexandra Markó. 2010. A Pilot Study of Hungarian Discourse Markers. *Language and Speech* 53(2). 135–180. doi:10.1177/0023830909357162.

Derwing, Tracey M., Murray J. Munro, Ronald I. Thomson & Marian J. Rossiter. 2009. The Relationship between L1 Fluency and L2 Fluency Development. *Studies in Second Language Acquisition* 31(04). 533–557. doi:10.1017/S0272263109990015.

Derwing, Tracey M. & Marian J. Rossiter. 2003. The Effects of Pronunciation Instruction on the Accuracy, Fluency and Complexity of L2 Accented Speech. *Applied Language Learning*(13). 1–18.

Derwing, Tracey M., Marian J. Rossiter, Murray J. Munro & Ron I. Thomson. 2004. Second Language Fluency: Judgments on Different Tasks. *Language Learning* 54(4). 655–679. doi:10.1111/j.1467-9922.2004.00282.x.

Deschamps, A. 1980. The syntactical distribution of pauses in English spoken as a second language by French students. In Hans W. Dechert & Manfred Raupach (eds.), *Temporal Variables in Speech*. The Hague: Mouton de Gruyter.

Dewaele, Jean-Marc. 1996. Les phénomènes d'hésitation dans l'interlangue française: analyse de la variation interstylistique et interindividuelle. *Rassegna Italiana da Linguistica Applicata* 28(1). 87–103.

Dewaele, Jean-Marc & Adrian Furnham. 2000. Personality and speech production: a pilot study of second language learners. *Personality and Individual Differences* 28(2). 355–365. doi:10.1016/S0191-8869(99)00106-3.

DiStefano, Christine, Min Zhu & Diana Mindrila. 2009. Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation* 14(20). 1–11.

Dister, Anne. 2007. De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL. Louvain-la-Neuve, Belgium: Université catholique de Louvain.

Dister, Anne & Anne-Catherine Simon. 2008. La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé. *Arena Romanistica* 1(1). 54–79.

Divjak, Dagmar & Stefan Th. Gries. 2006. Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1). 23–60. doi:10.1515/CLLT.2006.002.

Dolnicar, Sara & Bettina Grün. 2008. Challenging "Factor Cluster Segmentation." *Journal of Travel Research* 47(1). 63–71.

Dore, Cecilia. 2016. Perceptions of fluency. Reading: University of Reading MA dissertation. https://englishagenda.britishcouncil.org/sites/default/files/attachments/dissertation_design_for_publication_2016_reading_university_cecilia_dore.pdf (16 March, 2018).

Du Bois, John. 1974. *Syntax in mid-sentence*. (Ed.) Charles J. Fillmore, George Lakoff & Robin Lakoff. . Vol. I. (Berkeley Studies in Syntax and Semantics). Berkeley: University of California, Institute of Human Learning and Department of Linguistics.

Du Bois, John W., Stephan Schuetze-Coburm, Susanna Cumming & Danae Paolino. 1993. Outline of discourse transcription. In Jane A. Edwards & Martin D. Lampert (eds.), *Talking Data: Transcription and Coding in Discourse Research*, 45–89. Hillsdale: Lawrence Erlbaum Associates. (3 March, 2014).

Duez, Danielle. 1985. Perception of Silent Pauses in Continuous Speech. *Language and Speech* 28(4). 377–389.

Duez, Danielle. 1998. The aims of SPoSS. Introductory remarks. *Proceedings of the ESCA Workshop on the Sound Patterns of Spontaneous Speech*. 7–9.

Duez, Danielle. 2001. Signification des hésitations dans la parole spontanée. *Parole* 17–19. 113–138.

Dumont, Amandine. 2015. Designing and implementing a multilayer annotation system for (dis)fluency features in learner and native corpora. In Federica Formato & Andrew Hardie (eds.), *Corpus Linguistics 2015: Abstract Book*, 96–98. Lancaster (UK).

Dumont, Amandine. 2016. A corpus-driven approach to native and learner spoken fluency. The contribution of pauses. Paper presented at the Nouvelles Approches de Corpus Linguistique Anglaise (NACLA1), Avignon, France.

Dumont, Amandine. 2017a. The contribution of learner corpora to the substantiation of fluency levels. In Pieter De Haan, Sanne Van Vuuren & Rina De Vries (eds.), *Language, learners and levels. Progression and variation*, vol. Proceedings 3, 281–308. Presses universitaires de Louvain. (Corpora and Language in Use). Louvain-la-Neuve.

Dumont, Amandine. 2017b. The effects of speaking task on L2 fluency. Bolzano. https://dial.uclouvain.be/pr/boreal/en/object/boreal%3A189830 (29 November, 2017).

Dybkjaer, Laila & Niels Ole Bernsen. 2000. The MATE markup framework. *Proceedings of the 1st SIGdial workshop on Discourse and dialogue-Volume 10*, 19–28. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1117739 (28 February, 2017).

Eckart, Kerstin. 2012. Resource annotations. *CLARIN-D User Guide*, 30–42. http://media.dwds.de/clarin/userguide/userguide-1.0.1.pdf (20 August, 2014).

Eco, Umberto. 1984. *The Name of the Rose*. London: Picador Edition.

Edwards, Jane A. 1992. Principles and contrasting systems of discourse transcription. In Jane A. Edwards & Martin D. Lampert (eds.), *Talking data. Transcription and coding in discourse research*, 3–32. Hillsdale: Lawrence Erlbaum Associates.

Edwards, Jane A. 2001. The transcription of discourse. In Deborah Schiffrin, Deborah Tannen & Heidi E. Hamilton (eds.), *The Handbook of Discourse Analysis*, 321–348. Oxford: Blackwell Publishing Ltd. doi:10.1111/b.9780631205968.2003.00018.x. http://doi.wiley.com/10.1111/b.9780631205968.2003.00018.x (30 January, 2017).

Ejzenberg, Roseli. 1996. *Understanding nonnative oral fluency: the role of task structure and discourse variability*. MI: University Microfilms International. Ann Arbor.

Ejzenberg, Roseli. 1997. The role of task structure in oral fluency assessment. Paper presented at the Presented at the 28th Annual Meeting of TESOL USA, Baltimore. (11 July, 2017).

Ejzenberg, Roseli. 2000. The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In Heidi Riggenbach (ed.), *Perspectives on fluency*, 287–313. Ann Arbor. University of Michigan Press.

Eklund, Robert. 2000. Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and Tok Pisin Human–Human ATIS Dialogues. *Proceedings of ICSLP 2000*, 991–994. Beijing, China: China Military Friendship Publish.

Eklund, Robert. 2001. Prolongations: A dark horse in the disfluency stable. *Proceedings of DISS'01*, 5–8. Edinburgh, UK. http://www.isca-speech.org/archive_open/diss_01/dis1_005.html (19 May, 2017).

Eklund, Robert. 2004. Disfluency in Swedish human–human and human–machine travel booking dialogues. http://www.diva-portal.org/smash/record.jsf?pid=diva2:20923 (3 June, 2014).

Eklund, Robert & Elizabeth Shriberg. 1998. Crosslinguistic Disfluency Modelling: A Comparative Analysis of Swedish and American English Human-Human and Human-Machine Dialogues. *Proceedings of ICSLP 98*, 2631–2634. Sydney. http://mirlab.org/conference_papers/International_Conference/ICSLP%201998/PDF/SCAN/SL980805.PDF (19 May, 2017).

Eldridge, John. 1996. Code-switching in a Turkish secondary school. *ELT journal* 50(4). 303–311.

Ellis, Rod. 1985. *Understanding Second Language Acquisition*. Oxford University Press.

Ellis, Rod. 2009. The Differential Effects of Three Types of Task Planning on the Fluency, Complexity, and Accuracy in L2 Oral Production. *Applied Linguistics*. 474–509. doi:10.1093/applin/amp042.

Ellis, Rod & Gary Patrick Barkhuizen. 2005. *Analysing Learner Language*. Oxford: Oxford University Press.

Ellis, Rod & Fangyuan Yuan. 2004. The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition* 26(01). 59–84. doi:10.1017/S0272263104026130.

Erard, Michael. 2004. THINK TANK; Just Like, Er, Words, Not, Um, Throwaways. *The New York Times*, sec. Arts. https://www.nytimes.com/2004/01/03/arts/think-tank-just-like-er-words-not-um-throwaways.html (1 February, 2018).

Erman, Britt. 1987. *Pragmatic Expressions in English: A Study of You Know, You See, and I Mean in Face-to-face Conversation*. Almqvist & Wiksell International.

Erman, Britt. 2001. Pramatic markers revisited with a focus on "you know" in adult and adolescent talk. *Journal of Pragmatics* 33. 1337–1359.

ESOL Examinations (ed.). 2011. *Using the CEFR: Principles of good practice*. Cambridge: University of Cambridge. http://www.cambridgeenglish.org/images/126011-using-cefr-principles-of-good-practice.pdf (3 November, 2016).

Fathman, Ann K. 1980. Repetition and correction as an indication of speech planning and execution processes among second language learners. In Hans-Wilhelm Dechert & Manfred Raupach (eds.), *Towards a Cross-Linguistic Assessment of Speech Production*, 77–86. (Kasseler Arbeiten Zur Sprache Und Literatur 7). Frankfurt a. M: Lang.

Fayer, Joan M. & Emily Krasinski. 1987. Native and nonnative judgments of intelligibility and irritation. *Language Learning* 37(3). 313–326. doi:10.1111/j.1467-1770.1987.tb00573.x.

Field, Andy. 2013. *Discovering Statistics Using IBM SPSS Statistics, 4th Edition*. 4th edition. Los Angeles: SAGE Publications Ltd.

Fielder, Grace E. 2008. Bulgarian adversative connectives: Conjunctions or discourse markers? In Ritva Laury (ed.), *Crosslinguistic Studies of Clause Combining*, 79–97. Amsterdam: John Benjamins Publishing Company. doi:10.1075/tsl.80.05fie. https://benjamins.com/catalog/tsl.80.05fie (18 August, 2017).

Fillmore, Charles J. 1979. On fluency. In Charles J. Fillmore, Daniel Kempler & William S. Wang (eds.), *Individual Differences in Language Ability and Language Behavior*, 85–101. New York: Academic Press.

Fillmore, Charles J. 2000. On fluency. In Heidi Riggenbach (ed.), *Perspectives on fluency*, 252–261. Ann Arbor. University of Michigan Press.

Fokes, Joann & Z. S. Bond. 1989. The Vowels of Stressed and Unstressed Syllables in Nonnative English*. *Language Learning* 39(3). 341–373. doi:10.1111/j.1467-1770.1989.tb00596.x.

Foster, Pauline & Peter Skehan. 1996. The Influence of Planning and Task Type on Second Language Performance. *Studies in Second Language Acquisition* 18(3). 299–323. doi:10.1017/S0272263100015047.

Foster, Pauline & Peter Skehan. 1999. The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research* 3(3). 215–247. doi:10.1177/136216889900300303.

Foster, Pauline & Peter Skehan. 2009. The influence of planning and task type on second language performance. In Kris Van den Branden, Martin Bygate & John Michael Norris (eds.), *Task-based language teaching: a reader*, 275–300. (Task-Based Language Teaching : Issues, Research and Practice 1). Amsterdam: Benjamins.

Foster, Pauline & Parvaneh Tavakoli. 2009. Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language learning* 59(4). 866–896.

Foster, Pauline, Alan Tonkyn & Gillian Wigglesworth. 2000. Measuring spoken language: a unit for all reasons. *Applied Linguistics* 21(3). 354–375. doi:10.1093/applin/21.3.354.

Fox, Barbara & Robert Jasperson. 1995. A syntactic exploration of repair in English conversation. In Philip Davis (ed.), *Alternative linguisitcs: Descriptive and theoretical modes*, 77–134. Amsterdam: John Benjamins.

Fox Tree, Jean E. 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language* 34. 709–738.

Fox Tree, Jean E. 2001. Listeners' uses of "um" and "uh" in speech comprehension. *Memory and Cognition* 29(2). 320–326.

Fox Tree, Jean E. 2006. Placing "like" in telling stories. *Discourse Studies* 8(6). 723–743. doi:10.1177/1461445606069287.

Fox Tree, Jean E. & Herbert H. Clark. 1997. Pronouncing "the" as "thee" to signal problems in speaking. *Cognition* 62(2). 151–167. doi:10.1016/S0010-0277(96)00781-0.

Fox Tree, Jean E. & Josef C. Schrock. 2002. Basic meanings of "you know" and "I mean." *Journal of Pragmatics* 34. 727–747.

Fraser, Bruce. 1988. Types of English discourse markers. *Acta Linguistica Hungarica* 38(1–4). 19–33.

Fraser, Bruce. 1990. An approach to discourse markers. *Journal of pragmatics* 14(3). 383–398.

Fraser, Bruce. 1999. What are discourse markers? *Journal of pragmatics* 31(7). 931–952.

Fraser, Bruce. 2005. Towards a theory of discourse markers. *Approaches to discourse particles* 1. 189–204.

Fraundorf, Scott H. & Duane G. Watson. 2011. The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language* 65(2). 161–175. doi:10.1016/j.jml.2011.03.004.

Freed, Barbara F. 1995. What makes us think that students who study abroad become fluent? In Barbara F. Freed (ed.), *Second Language Acquisition in a Study Abroad Context*, 123–148. (Studies in Bilingualism 9). Amsterdam & Philadelphia: John Benjamins Publishing Company. http://libra.msra.cn/Publication/2020650/what-makes-us-think-that-students-who-study-abroad-become-fluent (10 March, 2016).

Freed, Barbara F. 2000. Is fluency, like beauty, in the eyes (and ears) of the beholder? In Heidi Riggenbach (ed.), *Perspectives on fluency*, 243–265. Ann Arbor. University of Michigan Press.

Freed, Barbara F., Norman Segalowitz & Dan P. Dewey. 2004. Context of Learning and Second Language Fluency in French: Comparing Regular Classroom, Study Abroad, and Intensive Domestic Immersion Programs. *Studies in Second Language Acquisition* 26(02). 275–301. doi:10.1017/S0272263104262064.

Friedman, Ernest H. 1991a. Speech pauses and diagnosis. *Journal of Clinical Psychiatry* 52(4). 181–182.

Friedman, Ernest H. 1991b. Speech hesitation pauses as markers for mood disorder in stroke patients. *The Journal of Clinical Psychiatry* 52(3). 140.

Friginal, Eric, Man Li & Sara C. Weigle. 2014. Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing* 23. 1–16. doi:10.1016/j.jslw.2013.10.001.

Fulcher, Glenn. 1996. Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13(2). 208–238.

Fulcher, Glenn. 2003. *Testing second language speaking*. (Applied Linguistics and Language Study). Harlow: Pearson Education.

Fuller, Janet M. 2003. Discourse marker use across speech contexts: A comparison of native and non-native speaker performance. *Multilingua* 22(2). 185–208.

Fung, Loretta & Ronald Carter. 2007. Discourse Markers and Spoken English: Native and Learner Use in Pedagogic Settings. *Applied Linguistics* 28(3). 410–439. doi:10.1093/applin/amm030.

Garner, W. R. 1960. Rating scales, discriminability, and information transmission. *Psychological Review* 67(6). 343–352. doi:10.1037/h0043047.

Garside, Roger, Geoffrey Leech & Anthony McEnery. 1997. *Corpus annotation: Linguistic information from computer text corpora*. Longman.

Gass, Susan M. & Larry Selinker. 2008. *Second Language Acquisition: An Introductory Course*. 3rd edition. New York: Routledge.

Gelderen, Amos van & Ron Oostdam. 2002. Improving Linguistic Fluency for Writing: Effects of Explicitness and Focus of Instruction. *L1-Educational Studies in Language and Literature* 2(3). 239–270. doi:10.1023/A:1021304027877.

Gilquin, Gaëtanelle. 2008. Hesitation markers among EFL learners: Pragmatic deficiency or difference? *Pragmatics and corpus linguistics: A mutualistic entente* 2. 119–149.

Gilquin, Gaëtanelle. 2015. From design to collection of learner corpora. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, 9–34. Cambridge: Cambridge University Press.

Gilquin, Gaëtanelle. 2016. Discourse markers in L2 English: From classroom to naturalistic input. In Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja & Sarah Chevalier (eds.), *New Approaches to English Linguistics: Building Bridges*, vol. 177, 213–

249. Amsterdam: John Benjamins Publishing Company. doi:10.1075/slcs.177.09gil. https://benjamins.com/catalog/slcs.177.09gil (18 August, 2017).

Gilquin, Gaëtanelle, Yves Bestgen & Sylviane Granger. 2016. Assessing the CEFR assessment grid for spoken language use: A learner corpus-based approach. https://dial.uclouvain.be/pr/boreal/object/boreal:179650 (8 November, 2017).

Gilquin, Gaëtanelle, Sylvie De Cock & Sylviane Granger (eds.). 2010. *LINDSEI. Louvain International Database of Spoken English Interlanguage*. Presses Universitaires de Louvain. Louvain-la-Neuve.

Gilquin, Gaëtanelle & Sylviane Granger. 2015. Learner language. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of Corpus Linguistics*, 418–435. Cambridge: Cambridge University Press. http://dial.academielouvain.be/downloader/downloader.php?pid=boreal%3A145508&datastream=PDF_01&disclaimer=9c9c5684143ca560d241ccb2eb8120523c6f4586754d10d872a87d14984ff2fd (9 January, 2015).

Ginther, April, Slobodanka Dimova & Rui Yang. 2010. Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing* 27(3). 379–399. doi:10.1177/0265532210364407.

Goldman, Jean-Philippe, Antoine Auchlin & Anne-Catherine Simon. 2013. Les variables temporelles dans le dialogue. In P. Mertens & Anne-Catherine Simon (eds.), *Proceedings of the Prosody-Discourse Interface Conference (IDP 2013)*, 39–44. Leuven. http://dial.uclouvain.be/handle/boreal:133776 (10 November, 2016).

Goldman, Jean-Philippe, Thomas François, Sophie Roekhaut & Anne-Catherine Simon. 2010. Étude statistique de la durée pausale dans différents styles de parole. *Actes des 28e Journées d'Etude sur la Parole (JEP)*, 161–164. Mons (Belgium). http://www.researchgate.net/profile/Anne-Catherine_Simon/publication/228861968_tude_statistique_de_la_dure_pausale_dans_diffrents_styles_de_parole/links/00b49519cc222dba09000000.pdf (9 September, 2015).

Goldman, J.-Ph. 2011. EasyAlign: an automatic alignment tool under Praat. *Proceedings of InterSpeech*. Fire ze. http://latlcui.unige.ch/phonetique/easyalign.php (23 December, 2013).

Goldman-Eisler, Frieda. 1951. The Measurement of Time Sequences in Conversational Behaviour. *British Journal of Psychology. General Section* 42(4). 355–362. doi:10.1111/j.2044-8295.1951.tb00314.x.

Goldman-Eisler, Frieda. 1954a. A study of individual differences and of interaction in the behaviour of some aspects of language in interviews. *Journal of Mental Science* 100. 177–197.

Goldman-Eisler, Frieda. 1954b. On the variability of the speech of talking and on its relation to the length of utterances in conversations. *British Journal of Psychology* 45. 94–107.

Goldman-Eisler, Frieda. 1956. The determinants of the rate of speech output and their mutual relations. *Journal of Psychosomatic Research* 1(2). 137–143. doi:10.1016/0022-3999(56)90015-0.

Goldman-Eisler, Frieda. 1958a. Speech analysis and mental processes. *Language and Speech* 1. 59–75.

Goldman-Eisler, Frieda. 1958b. The predictability of words in context and the length of pauses in speech. *Language and Speech* 1. 226–231.

Goldman-Eisler, Frieda. 1961a. The distribution of pause durations in speech. *Language and Speech* 4(4). 232–237.

Goldman-Eisler, Frieda. 1961b. The significance of changes in the rate of articulation. *Language and Speech* 4(4). 171–174.

Goldman-Eisler, Frieda. 1968. *Psycholinguistics: Experiments in spontaneous speech*. London & New York: Academic Press.

Good, David & Brian L. Butterworth. 1980. Hesitancy as a conversational resource: some methodological implications. In Hans-Wilhelm Dechert & Manfred Raupach (eds.), *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*, 146–152. The Hague: Mouton de Gruyter. https://www.researchgate.net/publication/230876073_Hesitancy_as_a_conversational_resource_some_methodological_implications (18 July, 2016).

Götz, Sandra. 2011. Fluency in Native and Nonnative English Speech: Theory, description, implications. Giessen, Germany Inaugural Dissertation.

Götz, Sandra. 2013a. *Fluency in Native and Nonnative English Speech*. (Studies in Corpus Linguistics (SCL) 53). Amsterdam & Philadelphia: John Benjamins Publishing Company.

Götz, Sandra. 2013b. How fluent are advanced German learners of English (perceived to be)? Corpus findings vs. native-speaker perception. In Magnus Huber & Joybrato Mukherjee (eds.), *Studies in Variation, Contacts and Change in English*, vol. 13. Giessen: University of Giessen. (9 September, 2013).

Götz, Sandra. 2017. Do (non-linguistic) variables affect learners' (dis)fluency? A learner corpus-based perspective. Louvain-la-Neuve.

Gráf, Tomáš. 2015. Accuracy and fluency in the speech of the advanced learner of English. Prague, Czech Republic: Univerzita Karlova v Praze Unpublished PhD thesis.

Gráf, Tomáš. 2017. Repeats in advanced spoken English of learners with Czech as L1. *Acta Universitatis Carolinae* 2017(3). 65–78. doi:10.14712/24646830.2017.34.

Gráf, Tomáš & Lan Fen Huang. 2017. Repeats in native and learner English. Louvain-la-Neuve.

Granger, Sylviane. 1998. The computerized learner corpus: a versatile new source of data for SLA research. In Sylviane Granger (ed.), *Learner English on computer*, 3–18. London & New York: Addison Wesley Longman.

Granger, Sylviane. 2008. Learner Corpora in Foreign Language Education. In Nelleke Van Deusen-Scholl & Hornberger (eds.), *Encyclopedia of Language and Education*, vol. 4, 1427–1441. Boston, MA: Springer US. doi:10.1007/978-0-387-30424-3_109 (17 March, 2017).

Granger, Sylviane. 2012. Learner Corpora. In Carol A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*. Oxford, UK: Blackwell Publishing Ltd.

doi:10.1002/9781405198431.wbeal0669.
http://doi.wiley.com/10.1002/9781405198431.wbeal0669 (24 April, 2017).

Granger, Sylviane. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1). 7–24.

Granger, Sylviane, Gaëtanelle Gilquin & Fanny Meunier (eds.). 2015a. *The Cambridge Handbook of Learner Corpus Research*. (Cambridge Handbooks in Language and Linguistics). Cambridge: Cambridge University Press.

Granger, Sylviane, Gaëtanelle Gilquin & Fanny Meunier. 2015b. Introduction: learner corpus research - past, present and future. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, 1–5. (Cambridge Handbooks in Language and Linguistics). Cambridge: Cambridge University Press.

Granger, Sylviane & Jennifer Thewissen. 2005. Towards a reconciliation of a Can Do and Can't Do approach to language assessment. Voss (Norway). https://dial.uclouvain.be/pr/boreal/object/boreal:75893 (25 April, 2017).

Greggio, Saionara & Gloria Gil. 2007. Teacher's and Learners' Use of Code Switching in the English as a Foreign Language Classroom: A Qualitative Study. *Linguagem & Ensino* 10(2). 371–393.

Gries, Stefan Th. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1(2). 109–151.

Gries, Stefan Th. 2010. Corpus linguistics and theoretical linguistics. A love-hate relationship? Not necessarily… *International Journal of Corpus Linguistics* 15(3). 327–343.

Gries, Stefan Th. 2013. *Statistics for Linguistics with R*. 2nd revised edition. De Gruyter Mouton.

Gries, Stefan Th & Andrea L. Berez. 2015. Linguistic annotation in/for corpus linguistics. In Nancy Ide & James Pustejovsky (eds.), *Handbook of Linguistic Annotation*, 379–409. Berlin: Springer. http://www.linguistics.ucsb.edu/faculty/stgries/research/2017_STG_ALB_LingAnnotCorpLing_HbOfLingAnnot.pdf (7 June, 2017).

Griffiths, Roger. 1991. Pausological Research in an L2 Context: A Rationale, and Review of Selected Studies. *Applied Linguistics* 12(4). 345–364.

Griffiths, Roger & Alan Beretta. 1991. A Controlled Study of Temporal Variables in NS-NNS Lectures. *RELC Journal* 22(1). 1–19. doi:10.1177/003368829102200101.

Grosjean, François. 1972. Analyse des variables temporelles du français spontané. *Phonetica* 26(3). 129–157.

Grosjean, François. 1980a. Comparative studies of temporal variables in spoken and sign languages: A short review. In Hans W. Dechert & Manfred Raupach (eds.), *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*, 307–312. The Hague: Mouton.

Grosjean, François. 1980b. Linguistic structures and performance structures: Studies in pause distribution. In Hans W. Dechert & Manfred Raupach (eds.), *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*, 91–111. The Hague: Mouton de Gruyter.

Grosjean, François. 1980c. Temporal variables within and between languages. In Hans-Wilhelm Dechert & Manfred Raupach (eds.), *Towards a Cross-Linguistic Assessment of Speech Production*, 39–54. (7). Frankfurt a. M: Lang.

Grosjean, François & Alain Deschamps. 1975. Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica* 31(3–4). 144–184. doi:10.1159/000259667.

Grosman, Iulia. forthcoming. Evaluation contextuelle de la (dis)fluence en production et perception. Pratiques communicatives et formes prosodico-syntaxiques en français. Louvain-la-Neuve: Université catholique de Louvain PhD Thesis.

Guadagnoli, Edward & Wayne Velicer. 1988. Relation of Sample Size to the Stability of Component Patterns. *Psychological Bulletin* 103(2). 265–275. doi:10.1037//0033-2909.103.2.265.

Gumperz, John Joseph. 1982. *Discourse strategies*. Repr. (Studies in Interactional Sociolinguistics 1). Cambridge: Cambridge university press.

Gut, Ulrike. 2004. The LeaP corpus. A phonetically annotated corpus of non-native speech. http://www.philhist.uni-augsburg.de/de/lehrstuehle/anglistik/applied-linguistics/workshop/pdfs/LeapCorpus_Manual.pdf (23 February, 2017).

Gut, Ulrike. 2009. *Non-native Speech: A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. (Ed.) Thomas Kohnen & Joybrato Mukherjee. (English Corpus Linguistics 9). Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.

Gut, Ulrike. 2012. The LeaP corpus: A multilingual corpus of spoken learner German and learner English. In Thomas Schmidt & Kai Wörner (eds.), *Hamburg Studies on Multilingualism*, vol. 14, 3–23. Amsterdam: John Benjamins Publishing Company. doi:10.1075/hsm.14.03gut. https://benjamins.com/catalog/hsm.14.03gut (23 February, 2017).

Gut, Ulrike. 2014. Corpus Phonology and Second Language Acquisition. In Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.), *The Oxford Handbook of Corpus Phonology*, 286–301. Oxford: Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199571932.001.0001/oxfordhb-9780199571932-e-027 (24 April, 2017).

Gut, Ulrike & Robert Fuchs. 2017. Exploring speaker fluency with phonologically annotated ICE corpora. *World Englishes*. http://www.academia.edu/download/46824793/Gut_Fuchs17_Exploring_speaker_fluency_with_phonologically_annotated_ICE_corpora.pdf (22 September, 2016).

Guz, Ewa. 2015. Establishing the Fluency Gap Between Native and Non-Native-Speech. *Research in Language* 13(3). 230–247. doi:10.1515/rela-2015-0021.

Hair, Joseph F., Rolph E. Anderson, Ronald L. Tatham & William C. Black (eds.). 1995. *Multivariate data analysis: with readings*. 4th ed. Upper Saddle River (N.J.): Prentice Hall.

Hallgren, Kevin A. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology* 8(1). 23–34.

Halliday, Michael Alexander Kirkwood & Ruqaiya Hasan. 1976. *Cohesion in English*. 1st, 2nd impr ed. (Longman Paperbacks 9). London: Longman.

Hansson, Petra. 1999. Discourse markers in dialogue. *Proceedings FONETIK*, vol. 99, 65–68. https://pdfs.semanticscholar.org/05ac/38704f8ad79e6768516551 9fe22eeca912e0.pdf (7 June, 2017).

Hasselgren, Angela. 1994. Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4(2). 237–258. doi:10.1111/j.1473-4192.1994.tb00065.x.

Hasselgren, Angela. 1998. Smallwords and valid testing. Bergen: University of Bergen.

Hasselgren, Angela. 2002. Learner corpora and language testing. Smallwords as markers of learner fluency. In Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 143–174. (Language Learning & Language Teaching). Amsterdam & Philadelphia: John Benjamins Publishing Company.

Hasselgren, Angela. 2005. *Testing the spoken English of young norwegians: A study of test validity and the role of "smallwords" in contributing to pupils' fluency*. Cambridge: Cambridge University Press.

Hawkins, P.R. 1971. The syntactic location of hesitation pauses. *Language and Speech* 14(3). 277–288.

Hedeland, Hanna & Thomas Schmidt. 2012. Technological and methodological challenges in creating, annotating and sharing a learner corpus of spoken German. In Thomas Schmidt & Kai Wörner (eds.), *Multilingual Corpora and Multilingual Corpus Analysis*, 25–46. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Hedström, Karl. 1984. A Study of Repairs in Speech. *Stockholm Papers in English Language and Literature* 4. 69–101.

Hee Jeon, Eun. 2015. Multiple regression. In Luke Plonsky (ed.), *Advancing Quantitative Methods in Second Language Research*, 131–158. (Second Language Acquisition Research Series). New York ; London: Routledge.

Heeman, Peter A., Andy McMillin & J. Scott Yaruss. 2006. An annotation scheme for complex disfluencies. *INTERSPEECH*. (30 May, 2014).

Henderson, Alan, Frieda Goldman-Eisler & Andrew Skarbek. 1966. Sequential Temporal Patterns in Spontaneous Speech. *Language and Speech* 9(4). 207–216. doi:10.1177/002383096600900402.

Henry, Sandrine. 2002. Quelles répétitions à l'oral? Esquisse d'une typologie. *Actes des 2èmes Journées de Linguistique de Corpus, Lorient, France*. http://sites.univ-provence.fr/delic/papiers/Henry-2002lorient.pdf (11 February, 2014).

Henry, Sandrine & Bertille Pallaud. 2004. Amorces de mots et répétitions: des hésitations plus que des erreurs en français parlé. *Amorces et répétitions de mots*, vol. 2, 848–858. Louvain-la-Neuve (Belgium): Presses Universitaires de Louvain. http://hal.archives-ouvertes.fr/docs/00/28/35/79/PDF/1494.pdf (18 February, 2014).

Hieke, Adolf E. 1984. Linking as a marker of fluent speech. *Language and Speech* 27(4). 343–354. doi:10.1177/002383098402700405.

Hilton, Heather E. 2008. The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal* 36(2). 153–166.

Hilton, Heather E., John Osborne, Marie-Jo Derive, Nejma Suco, Jean O'Donnell, Sandrine Rutigliano & Sandra Billard. 2008. *Corpus PAROLE. Architecture du corpus & conventions de transcriptions*. Chambéry: Université de Savoie. http://archive.sfl.cnrs.fr/sites/sfl/IMG/pdf/PAROLE_manual.pdf (23 February, 2017).

Hincks, Rebecca. 2010. Speaking rate and information content in English lingua franca oral presentations. *English for Specific Purposes* 29(1). 4–18. doi:10.1016/j.esp.2009.05.004.

Honal, Matthias & Tanja Schultz. 2003. Correction of disfluencies in spontaneous speech using a noisy-channel approach. *INTERSPEECH*. http://pdf.aminer.org/003/045/589/correction_of_disfluencies_in_spontaneous_speech_using_a_noisy_channel.pdf (30 May, 2014).

Horowitz, Rosalind & S. Jay Samuels (eds.). 2005. *Comprehending oral and written language*. San Diego: Academic press.

House, Juliane. 2013. Developing pragmatic competence in English as a lingua franca: Using discourse markers to express (inter)subjectivity and connectivity. *Journal of Pragmatics* 59. (Pragmatic Development in L1, L2, L3: Its Biological and Cultural Foundations). 57–67. doi:10.1016/j.pragma.2013.03.001.

Housen, Alex, Folkert Kuiken & Ineke Vedder (eds.). 2012. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. (Language Learning & Language Teaching 32). Amsterdam: Benjamins.

Howe, Mary. 1991. Topic change in conversation. University of Kansas Unpublished PhD dissertation.

Howell, David C. 2013. *Statistical methods for Psychology*. 8th edition (International edition). Cengage Learning.

Huddleston, Rodney D. & Geoffrey K. Pullum (eds.). 2002. *The Cambridge grammar of the English language*. Repr. Cambridge: Cambridge university press.

Huensch, Amanda & Nicole Tracy-Ventura. 2016. Understanding second language fluency behavior: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics*. 1–31. doi:10.1017/S0142716416000424.

Hulstijn, Jan H. 2007. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal* 91(4). 663–667.

Hunston, Susan. 2002. *Corpora in applied linguistics*. 3rd print. (Cambridge Applied Linguistics). Cambridge: Cambridge university press.

IBM Corp. 2013. *IBM SPSS Statistics for Windows*. Armonk, New York: IBM.

Ide, Nancy & Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. *Proceedings of the Linguistic Annotation Workshop*, 1–8. Association for Computational Linguistics. (5 February, 2014).

Isaacs, Talia. 2010. Issues and arguments in the measurement of second language pronunciation. Montreal: McGill University.

Isaacs, Talia & Ron I. Thomson. 2013. Rater Experience, Rating Scale Length, and Judgments of L2 Pronunciation: Revisiting Research Conventions. *Language Assessment Quarterly* 10(2). 135–159. doi:10.1080/15434303.2013.769545.

Isaacs, Talia & Pavel Trofimovich. 2011. Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics* 32(01). 113–140. doi:10.1017/S0142716410000317.

Iwashita, Noriko, Annie Brown, Tim McNamara & Sallie O'Hagan. 2008. Assessed Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics* 29(1). 24–49. doi:10.1093/applin/amm017.

Izumi, Emi, Kiyotaka Uchimoto & Hitoshi Isahara. 2004. The NICT JLE Corpus: Exploiting the language learner's speech database for research and education. *International Journal of the Computer, the Internet and Management* 12(2). 119–125.

Izumi, Emi, Kiyotaka Uchimoto & Hitoshi Isahara. 2012. The NICT JLE Corpus (version 4.1). (23 February, 2017).

Jacewicz, Ewa, Robert Allen Fox & Lai Wei. 2010. Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America* 128(2). 839–850. doi:10.1121/1.3459842.

Jarvis, Scott. 2002. Topic continuity in L2 English article use. *Studies in Second Language Acquisition* 24(03). 387–418. doi:10.1017/S0272263102003029.

Jarvis, Scott, Leslie Grant, Dawn Bikowski & Dana Ferris. 2003. Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing* 12(4). 377–403. doi:10.1016/j.jslw.2003.09.001.

Jefferson, Gail. 1983. Issues in the transcription of naturally-occurring talk: Caricature vs. capturing pronunciational particulars. *Tilburg Papers in Language and Literature* 34. 1–12.

Jenks, Christopher Joseph. 2011. *Transcribing talk and interaction: issues in the representation of communication data*. Amsterdam: Benjamins.

Jin, Tan, Barley Mak & Pei Zhou. 2012. Confidence scoring of speaking performance: How does fuzziness become exact? *Language Testing* 29(1). 43–65. doi:10.1177/0265532211404383.

Jisa, Harriet. 1984. French preschoolers' use of et pis ('and then'). *First Language* 5(15). 169–184. doi:10.1177/014272378400501501.

Jo, Jinyoung. 2015. Native and non-native speakers' perceptions of fluency in L2 speech. *SNU Working Papers in English Linguistics and Language*, 40–62. (13). Seoul. http://s-space.snu.ac.kr/bitstream/10371/96069/1/5_%EC%A1%B0%EC%A7%84%EC%98%81.pdf (10 November, 2017).

Johnson, Alison. 2002. "So"… ?: Pragmatic implications of 'so'-prefaced questions in formal police interviews. In Janet Cotterill (ed.), *Language in the legal process*, 91–110. New York: Palgrave Macmillan.

Johnson, Wendell. 1959. *The Onset of Stuttering: Research Findings and Implications*. Minneapolis: University of Minnesota Press.

Johnson, Wendell. 1961. Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers. *The Journal of Speech and Hearing Disorders* Monograph supplement 7. 1–20.

Johnson, Wendell, Spencer F. Brown, James F. Curtis, Clarence W. Edney & Jacqueline Keaster. 1948. Stuttering. In Johnson Wendell, Spencer F. Brown, James F. Curtis, Clarence W. Edney & Jacqueline Keaster (eds.), *Speech Handicapped School Children*, 179–257. New York: Harper & Brothers Publishers.

Jucker, Andreas H. & Yael Ziv. 1998. *Discourse Markers: Descriptions and Theory*. Amsterdam & Philadelphia: John Benjamins Publishing.

Kahneman, D. 1973. *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.

Kahng, Jimin. 2014. Exploring Utterance and Cognitive Fluency of L1 and L2 English Speakers: Temporal Measures and Stimulated Recall: Utterance and Cognitive Fluency in L2. *Language Learning* 64(4). 809–854. doi:10.1111/lang.12084.

Kapatsinski, Vsevolod. 2004. Measuring the relationship of structure to use: Determinants of the extent of recycle in repetition repair. *Annual Meeting of the Berkeley Linguistics Society*, vol. 30, 481–492. http://journals.linguisticsociety.org/proceedings/index.php/BLS/article/viewFile/949/730 (8 July, 2016).

Kapatsinski, Vsevolod. 2010. Frequency of Use Leads to Automaticity of Production: Evidence from Repair in Conversation. *Language and Speech* 53(1). 71–105. doi:10.1177/0023830909351220.

Kasper, Gabriele & Eric Kellerman. 1997. Introduction: Approaches to communication strategies. In Gabriele Kasper & Eric Kellerman (eds.), *Communication strategies: psycholinguistic and sociolinguistic perspectives*, 1–16. (Applied Linguistics and Language Study). London: Longman.

Kaufman, Leonard & Peter Rousseeuw. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. *Biometrics* 47(2). (Wiley Series in Probability and Statistics). 788. doi:10.2307/2532178.

Kendall, Tyler. 2008. On the History and Future of Sociolinguistic Data. *Language and Linguistics Compass* 2(2). 332–351. doi:10.1111/j.1749-818X.2008.00051.x.

Kennedy, Sara & Pavel Trofimovich. 2008. Intelligibility, Comprehensibility, and Accentedness of L2 Speech: The Role of Listener Experience and Semantic Context. *Canadian Modern Language Review*. doi:10.3138/cmlr.64.3.459. http://www.utpjournals.press/doi/abs/10.3138/cmlr.64.3.459 (30 January, 2018).

Kirsner, Kim, John Dunn & Kathryn Hird. 2003. Fluency: Time for a paradigm shift. *ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*. (19 July, 2016).

Kohler, Klaus J. 2006. What is Emphasis and how is it coded? *Procedings of Speech Prosody 2006*, 748–751. Dresden. http://www.ipds.uni-kiel.de/kjk/pub_exx/kk2006_2/sp2006.html (19 May, 2017).

Koponen, Matti & Heidi Riggenbach. 2000. Overview: Varying perspectives on fluency. *Perspectives on fluency*, 5–24. Ann Arbor. University of Michigan Press.

Kormos, Judit. 1999. Monitoring and self-repair in L2. *Language Learning* 49(2). 303–342.

Kormos, Judit. 2000. The timing of self-repairs in second language speech production. *Studies in Second Language Acquisition* 22(02). 145–167.

Kormos, Judit. 2006. *Speech Production and Second Language Acquisition*. Psychology Press.

Kormos, Judit & Mariann Dénes. 2004. Exploring measures and perceptions of fluency in the speech of second language learners. *System* 32(2). 145–164. doi:10.1016/j.system.2004.01.001.

Koster, Cor J. & Ton Koet. 1993. The Evaluation of Accent in the English of Dutchmen. *Language Learning* 43(1). 69–92. doi:10.1111/j.1467-1770.1993.tb00173.x.

Kowal, Sabine, Richard Wiese & Daniel C. O'Connell. 1983. The use of time in storytelling. *Language and Speech* 26(4). 377–392.

Krashen, Stephen. 1983. *Second language acquisition and second language learning*. (Pergamon Institute of English. Language Teaching Methodology Series). Oxford: Pergamon.

Krashen, Stephen D. 1982. *Principles and practice in second language acquisition*. (Pergamon Institute of English. Language Teaching Methodology Series). Oxford: Pergamon.

Kübler, Sandra & Heike Zinsmeister (eds.). 2014. *Corpus Linguistics and Linguistically Annotated Corpora*. London, New Delhi, New York, Sydney: Bloomsbury. http://www.bloomsbury.com/uk/corpus-linguistics-and-linguistically-annotated-corpora-9781441164476/ (11 February, 2015).

Lacheret, Anne & Bernard Victorri. 2002. La période intonative comme unité d'analyse pour l'étude du français parlé: modélisation prosodique et enjeux linguistiques. *Verbum* 1(24). 55–72.

Ladd, D. Robert. 1996. *Intonational phonology*. (Cambridge Studies in Linguistics 79). Cambridge: Cambridge university press.

Lahmann, Cornelia, Rasmus Steinkrauss & Monika S. Schmid. 2015. Speed, breakdown, and repair: An investigation of fluency in long-term second-language speakers of English. *International Journal of Bilingualism*. doi:10.1177/1367006915613162 (8 September, 2016).

Landis, J. Richard & Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1). 159–174. doi:10.2307/2529310.

Lane, Harlan, François Grosjean, J. Le Berre & Erik Lewin. 1973. Exploring some properties of foreign-language utterances that control their comprehension. *Linguistics* 11(112). doi:10.1515/ling.1973.11.112.15. https://www.degruyter.com/view/j/ling.1973.11.issue-112/ling.1973.11.112.15/ling.1973.11.112.15.xml (11 July, 2017).

Larsen-Freeman, Diane. 2006. The Emergence of Complexity, Fluency, and Accuracy in the Oral and Written Production of Five Chinese Learners of English. *Applied Linguistics* 27(4). 590–619. doi:10.1093/applin/aml029.

Larsen-Freeman, Diane. 2009. Adjusting Expectations: The Study of Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics* 30(4). 579–589. doi:10.1093/applin/amp043.

Larsen-Freeman, Diane. 2014. Another step to be taken – Rethinking the end point of the interlanguage continuum. In Z. Han & Elizabeth Tarone (eds.), *Interlanguage. Forty years later*, 203–220. Amsterdam & Philadelphia: John Benjamins. doi:10.1075/lllt.39.11ch9.

Larsson Aas, Hege & Sylvi Rørvik. 2017. Investigating individual pause profiles through the use of a comparable NL1/IL corpus. *Proceedings of LCR2015*.

Lee, Akinobu, Tatsuya Kawahara & Kiyohiro Shikano. 2001. Julius - an open source real-time large vocabulary recognition engine. *Proceedings of Eurospeech 2001*, 1691–1694. (2 March, 2017).

Leech, Geoffrey. 2000. Grammars of spoken English: New outcomes of corpus-oriented research. *Language learning* 50(4). 675–724.

Leech, Geoffrey. 2010. Grammar on the move : recent changes in English grammatical usage. *JACET summer seminar proceedings 9 : English grammar 1900-2000 and Politeness in English*, 1–11. Tokyo: The Japan Association of College English Teachers. http://eprints.lancs.ac.uk/35636/ (8 March, 2018).

Leech, Geoffrey N, Marianne Hundt, Christian Mair & Nicholas Smith. 2009. *Change in contemporary English: a grammatical study*. Leiden: Cambridge University Press. http://public.eblib.com/choice/publicfullrecord.aspx?p=464842 (8 March, 2018).

Leech, Geoffrey N., Greg Myers & Jenny Thomas. 1995. *Spoken English on Computer: Transcription, Mark-up, and Application*. Longman.

Leech, Geoffrey Neil. 1997. Introducing corpus annotation. In Roger Garside, Geoffrey Leech & Anthony McEnery (eds.), *Corpus annotation: Linguistic information from computer text corpora*, 1–18. Longman.

Leeuw, Esther de. 2007. Hesitation Markers in English, German, and Dutch. *Journal of Germanic Linguistics* 19(02). 85–114. doi:10.1017/S1470542707000049.

Leijten, Mariëlle & Luuk Van Waes. 2013. Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication* 30(3). 358–392.

Lennon, Paul. 1990. Investigating Fluency in EFL: A Quantitative Approach. *Language Learning* 40(3). 387–417. doi:10.1111/j.1467-1770.1990.tb00669.x.

Lennon, Paul. 1995. Assessing Short-term Change in Advanced Oral Proficiency. *ITL - International Journal of Applied Linguistics* 109(1). 75–109. doi:10.1075/itl.109-110.04len.

Levelt, Willem J.M. 1983. Monitoring and self-repair in speech. *Cognition* 14(1). 41–104.

Levelt, Willem J.M. 1989. *Speaking: From intention to articulation*. (ACL-MIT Press Series in Natural-Language Processing). Cambridge (Mass.): MIT Press.

Levinson, Stephen C. 1983. *Pragmatics*. (Cambridge Textbooks in Linguistics). Cambridge [Cambridgeshire] ; New York: Cambridge University Press.

Levkina, Mayya & Roger Gilabert. 2012. The effects of cognitive task complexity on L2 oral production. In Alex Housen, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, 171–197. (Language Learning & Language Teaching 32). Amsterdam: Benjamins.

Lewkowicz, Jo A. 1997. Authenticity for whom? Does authenticity really matter? In A. Huhta, V. Kohnen, L. Lurki-Suonio & Sari Luoma (eds.), *Current developments and alternatives in language assessment (Proceedings of LTRC96)*, 165–184. Jyväskylä: Jyväskylä University Press.

Lewkowicz, Jo A. 2000. Authenticity in language testing: some outstanding questions. *Language testing* 17(1). 43–64.

Liberman, Mark. 2015a. Trump's eloquence. *Language log*. http://languagelog.ldc.upenn.edu/nll/?p=20492 (15 February, 2018).

Liberman, Mark. 2015b. Trump's rhetorical style. http://languagelog.ldc.upenn.edu/nll/?p=23057 (18 February, 2018).

Liberman, Mark. 2017. Presidential fluency. *Language log*. http://languagelog.ldc.upenn.edu/nll/?p=35174 (15 February, 2018).

Lickley, Robin. 1994. Detecting disfluency in spontaneous speech. Edinburgh, UK: University of Edinburgh PhD Thesis. https://www.researchgate.net/profile/Robin_Lickley/publication/243643325_Detecting_disfluency_in_spontaneous_speech/links/547ce1d10cf2cfe203c1fe4b/Detecting-disfluency-in-spontaneous-speech.pdf?origin=publication_detail&ev=pub_int_prw_xdl&msrp=SizBdlQfyBt2tpB-_uMrpcvDOJiWdtDlSoU6Nn2CM3ycpW-pifKa5nB_YoGlc37m-9DXPYLE0PdeufDwtvE8HU46gnMO_s9WdcsFQrNeLis.qwlTBiRI_ddBghkpWlZdYuR29YrTiA9yPX8MDoHd_NpzamijBaTwaQQYFIZBr7dHM4mobmJyJOCzrA953vOAXA.eewUF6LnVCuSNGJ3varmP4EZIGSHFcgYYTEDr9Fik1rr6F3myoUVOaCoL59RMxnJFCL88akpo9Xwuai-62IV-A (1 March, 2017).

Liddicoat, Anthony J. 2007. *An Introduction to Conversation Analysis*. London: Continuum. http://uclouvain.summon.serialssolutions.com/?q=conversation+analysis#!/search/document?ho=t&fvf=Language,English,f%7CContentType,Newspaper%20Article,t%7CContentType,Book%20Review,t&l=fr-FR&q=(conversation%20analysis)%20AND%20(AuthorCombined:(Liddicoat))&id=FETCHMERGED-proquest_abstracts_7423348912 (22 February, 2017).

Liebscher, Grit & Jennifer Dailey–O'Cain. 2005. Learner code-switching in the content-based foreign language classroom. *The Modern Language Journal* 89(2). 234–247.

Little, David. 2007. The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal* 91(4). 645–655.

Liu, Yang, Elizabeth Shriberg & Andreas Stolcke. 2003. Automatic disfluency identification in conversational speech using multiple knowledge sources. *Proceedings of InterSpeech*. (15 December, 2014).

Loewen, Shawn & Talip Gonulal. 2015. Exploratory factor analysis and principal components analysis. In Luke Plonsky (ed.), *Advancing Quantitative Methods in Second Language*

Research, 182–212. (Second Language Acquisition Research Series). New York ; London: Routledge.

Lounsbury, Floyd G. 1969. Pausal, juncture and hesitation phenomena. In Charles Egerton Osgood & Thomas Albert Sebeok (eds.), *Psycholinguistics: a survey of theory and research problems*. 4th print. (Indiana University Studies in the History and Theory of Linguistics). Bloomington (Ind.): Indiana university press.

Lowder, Matthew W. & Fernanda Ferreira. 2016. Prediction in the Processing of Repair Disfluencies. *Language, cognition and neuroscience* 31(1). 73–79. doi:10.1080/23273798.2015.1036089.

Lumley, Tom. 2005. *Assessing second language writing: the rater's perspective*. Frankfurt am Main: Peter Lang. https://works.bepress.com/tom_lumley/5/ (4 May, 2017).

MacGregor, Lucy J., Martin Corley & David I. Donaldson. 2009. Not all disfluencies are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language* 111(1). 36–45. doi:10.1016/j.bandl.2009.07.003.

Maclay, Howard & Charles E. Osgood. 1959. Hesitation Phenomena in Spontaneous English Speech. *WORD* 15(1). 19–44. doi:10.1080/00437956.1959.11659682.

Manning, Christopher D. & Hinrich Schütze. 2000. *Foundations of statistical natural language processing*. Cambridge: The MIT Press. http://cdn.preterhuman.net/texts/science_and_technology/artificial_intelligence/Foundations%20of%20Statistical%20Natural%20Language%20Processing%20-%20Christopher%20D.%20Manning.pdf (19 September, 2016).

McCarthy, Michael. 1998. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press. http://www.cambridge.org/ms/cambridgeenglish/catalog/teacher-training-development-and-research/spoken-language-and-applied-linguistics (31 January, 2017).

McCarthy, Michael. 2010. Spoken fluency revisited. *English Profile Journal* 1(01). 1–15. doi:10.1017/S2041536210000012.

McEnery, Anthony. 2003. Corpus linguistics. In Ruslan Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, 448–463. Oxford: Oxford University Press.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies: an advanced resource book*. (Routledge Applied Linguistics). Abingdon: Routledge.

Megyesi, Beáta & Sofia Gustafson-Capkova. 2002. Production and perception of pauses and their linguistic context in read and spontaneous speech in Swedish. *Proceedings of InterSpeech*, 2153–2156. https://pdfs.semanticscholar.org/7d79/389b1a6882ee8624of5a6481461fa1d4e243.pdf (25 January, 2017).

Mehnert, Uta. 1998. The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition* 20(1). 83–108.

Mehrang, Faezeh & Massoud Rahimpour. 2010. The impact of task structure and planning conditions on oral performance of EFL learners. *Procedia - Social and Behavioral Sciences* 2(2). 3678–3686. doi:10.1016/j.sbspro.2010.03.572.

Meisel, J. 1987. A note on second language speech production. In Hans W. Dechert & Manfred Raupach (eds.), *Psycholinguistic models of production*, 83–90. Norwood (N.J.): Ablex Publishing Corporation.

Mello, Heliana. 2014. Methodological issues for spontaneous speech corpora compilation. The case of C-ORAL-BRASIL. In Tommaso Raso & Heliana Mello (eds.), *Spoken corpora and linguistic studies*, 27–68. (Studies in Corpus Linguistics 61). Amsterdam & Philadelphia: John Benjamins Publishing Company.

Mertens, Piet. 1997. De la chaîne linéaire à la séquence de tons. *Traitement automatique des langues* 38(1). 27–51.

Meteer, Marie. 1995. Dysfluency Annotation Stylebook for the Switchboard Corpus.

Meunier, Fanny, Sylviane Granger, Damien Littré & Magali Paquot. 2010. LONGDALE. http://www.uclouvain.be/en-cecl-longdale.html (23 February, 2017).

Michel, Marije. 2011. Effects of Task Complexity and Interaction on L2 Performance. In P. Robinson (ed.), *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*, 141–174. Amsterdam: John Benjamins Publishing Company. http://eprints.lancs.ac.uk/59267/1/Michel2011_Effects_of_Task_Complexity_and_Interaction_on_L2_Performance_preprint.pdf (18 January, 2017).

Miller, Joanne, François Grosjean & Concetta Lomanto. 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica* 41(4). 215–225.

Miller, Kristyan Spelman, Eva Lindgren & Kirk PH Sullivan. 2008. The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *TESOL Quarterly* 42(3). 433–454.

Mira, Ariel. 1998. Discourse Markers and Form-function Correlations. In Andreas H. Jucker & Yael Ziv (eds.), *Discourse Markers: Descriptions and Theory*, 223–259. Amsterdam & Philadelphia: John Benjamins Publishing.

Mitchell, Rosamond, Laura Domínguez, Maria Arche, Florence Myles & Emma Marsden. 2008. *SPLLOC I*. Final research report. http://www.splloc.soton.ac.uk/doc/SPLLOC_1_Report_Final.pdf (23 February, 2017).

Möhle, Dorothea. 1984. A comparison of the second language speech production of different native speakers. In Hans W. Dechert, Dorothea Möhle & Manfred Raupach (eds.), *Second Language Productions*, 26–49. Narr. (Tübinger Beiträge Zur Linguistik 240). Tübingen.

Molenda, Marek & Piotr Pęzik. 2014. Extending the definition of confluence. A corpus-based study of advanced learners' spoken language. In Anna Turula, Beata Mikołajewska & Danuta Stanulewicz (eds.), *Insights into Technology Enhanced Language Pedagogy*, vol. 4, 105–118. (Gdansk Studies in Language). Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang. https://www.researchgate.net/publication/274690958_Extending_the_definition_of_confluence_A_corpus-based_study_of_advanced_learners'_spoken_language (10 March, 2016).

Moniz, Helena Gorete Silva. 2013. Processing Disfluencies in European Portuguese. Lisbon (Portugal): Universidade Técnica de Lisboa Ph.D. Thesis. http://repositorio.ul.pt/handle/10451/9614 (1 March, 2017).

Mooi, Erik & Marko Sarstedt. 2010. Cluster Analysis. *A Concise Guide to Market Research*, 237–284. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-12541-6_9. http://link.springer.com/10.1007/978-3-642-12541-6_9 (29 May, 2017).

Mora, Joan C. 2006. Age effects on oral fluency development. In Carmen Muñoz (ed.), *Age and the rate of foreign language learning*, 65–88. Clevedon: Multilingual Matters.

Mullen, Karen. 1980. Rater reliability and oral proficiency examinations. In J. Oller & K. perkins (eds.), *Research in Language testing*, 91–101. Rowley: Newbury House Publishers.

Müller, Simone. 2005. *Discourse Markers in Native and Non-Native English Discourse*. John Benjamins Publishing.

Munro, Murray J. & Tracey M. Derwing. 2001. Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. *Studies in second language acquisition* 23(04). 451–468.

Nacey, Susan & Anne-Line Graedler. 2013. Communication strategies used by Norwegian students of English. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead. Proceedings of the First Learner Corpus Research Conference (LCR 2011)*, 345–356. (Corpora and Language in Use). Louvain-la-Neuve: Presses Universitaires de Louvain.

Nagy, Naomi & Devyani Sharma. 2013. Transcription. In Robert Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 235–256. Cambridge: Cambridge university press.

Nakakubo, Takako. 2011. The effects of planning on second language oral performance in Japanese: processes and production. Iowa: University of Iowa. (9 March, 2017).

Neary-Sundquist, Colleen. 2014. The use of pragmatic markers across proficiency levels in second language speech. *Studies in Second Language Learning & Teaching* 4(4). 637–663.

Neary-Sundquist, Colleen A. 2013. The development of cohesion in a learner corpus. *Studies in Second Language Learning and Teaching* 3(1). 109. doi:10.14746/ssllt.2013.3.1.6.

Nesselhauf, Nadja. 2004. Learner Corpora and their Potential for Language Teaching. In John Sinclair (ed.), *How to Use Corpora in Language Teaching*, 125–152. Amsterdam: Benjamins. https://www.researchgate.net/publication/246663764_Learner_Corpora_and_their_Potential_for_Language_Teaching (16 March, 2017).

Nooteboom, Sibout Govert. 1980. Speaking and Unspeaking: Detection and Correction of Phonological and Lexical Errors in Spontaneous Speech. In Victoria Fromkin (ed.), *Errors in linguistic performance : slips of the tongue, ear, pen and hand*, 87–95. London: Academic Press.

North, Brian. 2007. The CEFR Illustrative Descriptor Scales. *The Modern Language Journal* 91(4). 656–659. doi:10.1111/j.1540-4781.2007.00627_3.x.

North, Brian. 2014. Putting the Common European Framework of Reference to good use. *Language Teaching* 47(2). 1–22.

Notarrigo, Ingrid. 2017. Les marqueurs de (dis)fluence en langue des signes de Belgique francophone. Namur: Université de Namur Unpublished PhD thesis.

O'Brien, Irena, Norman Segalowitz, Barbara Freed & Joe Collentine. 2007. Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition* 29(04). 557–581.

Ochs, Elinor, Emanuel A. Schegloff & Sandra A. Thompson. 1996. *Interaction and Grammar*. Cambridge University Press.

O'Connell, Daniel C. & Sabine Kowal. 1995. Basic Principles of Transcription. *Rethinking Methods in Psychology*, 93–105. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd. doi:10.4135/9781446221792.n7. http://sk.sagepub.com/books/rethinking-methods-in-psychology/n7.xml (29 January, 2017).

O'Connell, Daniel & Sabine Kowal. 2008. *Communicating with One Another: Toward a Psychology of Spontaneous Spoken Discourse*. (Cognition and Language). Springer.

Oh, Saerhim. 2006. Investigating the relationship between fluency measures and second language writing placement test decisions. Hawaii: University of Hawauu at Mãnoa. http://scholarspace.manoa.hawaii.edu/handle/10125/20203 (8 November, 2016).

Oller, D. Kimbrough. 1973. The effect of position in utterance on speech segment duration in English. *The Journal of the Acoustical Society of America* 54(5). 1235–1247. doi:10.1121/1.1914393.

Ortega, Lourdes. 1999. Planning and focus on form in L2 oral performance. *Studies in second language acquisition* 21(01). 109–148.

Ortega, Lourdes & Heidi Byrnes. 2008. The longitudinal study of advanced L2 capacities: An introduction. In Lourdes Ortega & Heidi Byrnes (eds.), *The Longitudinal Study of Advanced L2 Capacities*, 3–20. New York: Routledge / Taylor & Francis. (16 March, 2017).

Osborne, John. 2010. Researching fluency and proficiency with multimodal corpora. Pdf.

Osborne, John. 2011a. Oral learner corpora and the assessment of fluency in the Common European Framework. In Ana Frankenberg-Garcia, Lynne Flowerdew & Guy Aston (eds.), *New Trends in Corpora and Language Learning*, 181–197. (Research in Corpus and Discourse 81). London & New-York: Continuum.

Osborne, John. 2011b. Fluency, complexity and informativeness in native and non-native speech. *International Journal of Corpus Linguistics* 16(2). 276–298. doi:10.1075/ijcl.16.2.060sb.

O'Shaughnessy, D. 1993. Analysis and automatic recognition of false starts in spontaneous speech. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93)*, vol. 2, 724–727. doi:10.1109/ICASSP.1993.319414.

O'Shaughnessy, Douglas. 1992. Analysis of False Starts in Spontaneous Speech. *International Conference on Spoken Language Processing (ICSLP'92)*. Banff, Alberta, Canada. http://eric.ed.gov/?id=ED356506 (4 March, 2014).

Ozono, Shuichi & Harumi Ito. 2003. Logical Connectives as Catalysts for Interactive L2 Reading. *System* 31(2). 283–297.

Pallaud, Bertille. 2002. Les amorces de mots comme faits autonymiques en langage oral. *Recherches sur le Français Parlé* 1(17). 79–102.

Pallaud, Bertille & Sandrine Henry. 1995. Amorces de mots et répétitions: des hésitations plus que des erreurs en français parlé. *Macromolecular Chemistry and Physics* 196(9). 2715–2735.

Pallaud, Bertille, Stéphane Rauzy & Philippe Blache. 2013. Auto-interruptions et disfluences en français parlé dans quatre corpus du CID. *TIPA. Travaux interdisciplinaires sur la parole et le langage*(29). [en ligne]. doi:10.4000/tipa.995.

Palmer, Martha & Nianwen Xue. 2010. Linguistic annotation. In Alexander Clark, Chris Fox & Shalom Lappin (eds.), *The Handbook of Computational Linguistics and Natural Language Processing*, 238–270. Willey-Blackwell. United Kingdom: Blackwell Publishing Ltd.

Park, Kwanghyun. 2014. Corpora and Language Assessment: The State of the Art. *Language Assessment Quarterly* 11(1). 27–44. doi:10.1080/15434303.2013.872647.

Pawley, Andrew & Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and communication* 191. 191–226.

Pawley, Andrew & Frances Hodgetts Syder. 2000. The one clause at a time hypothesis. In Heidi Riggenbach (ed.), *Perspectives on fluency*, 163–199. Ann Arbor: University of Michigan Press. https://www.researchgate.net/publication/247934030_The_one_clause_at_a_time_hypothesis (14 June, 2016).

Pedhazur, Elazar J. & Liora Pedhazur Schmelkin. 1991. *Measurement, design, and analysis: an integrated approach*. Hillsdale, N.J: Lawrence Erlbaum.

Pellegrino, François, Christophe Coupé & Egidio Marsico. 2011. Across-language perspective on speech information rate. *Language* 87(3). 539–558.

Peters, Ted & Barry Guitar. 1991. *Stuttering: An integrated approach to its nature and treatment*. Baltimore: William & Wilkins.

Peterson, Carole & Allyssa McCabe. 1987. The connective 'and': do older children use it less as they learn other connectives? *Journal of Child Language* 14(2). 375–381. doi:10.1017/S0305000900012988.

Pica, T., R. Kanagy & J. Falodun. 1993. Choosing and using communication tasks for second language instruction and research. In Graham Crookes & Susan M. Gass (eds.), *Tasks and Language Learning: Integrating Theory and Practice*, 9–34. Cleveland: Multilingual Matters.

Pillai, Stefanie. 2006. Self-monitoring and self-repair in spontaneous speech. *k@ta* 8(2). 114–126.

Pinget, Anne-France, Hans Rutger Bosker, Hugo Quené & Nivja H. de Jong. 2014. Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing* 31(3). 349–365. doi:10.1177/0265532214526177.

Podesva, Robert & Elizabeth Zsiga. 2013. Sound recordings: acoustic and articulatory data. In Robert Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 169–194. Cambridge: Cambridge university press.

Porte, Graeme Keith (ed.). 2012. *Replication research in applied linguistics*. (The Cambridge Applied Linguistics Series). New York: Cambridge university press.

Postma, Albert, Herman Kolk & Dirk-Jan Povel. 1990. On the relation among speech errors, disfluencies, and self-repairs. *Language and Speech* 33(1). 19–29.

Poulisse, Nanda. 1987. Problems and solutions in the classification of compensatory strategies. *Second Language Research* 3. 141–153.

Poulisse, Nanda, Theo Bongaerts & Eric Kellerman. 1984. On the use of compensatory strategies in second language performance. *Interlanguage Studies Bulletin*(8). 70–105.

Povolná, Renata. 2009. Exploring interactive discourse markers in academic spoken discourse. In Olga Dontcheva-Navratilova & Renata Povolná (eds.), *Coherence and Cohesion in Spoken and Written Discourse*, 60–80. Newcastle upon Tyne: Cambridge Scholars Publishing.

Préfontaine, Yvonne. 2013a. Perceptions of French Fluency in Second Language Speech Production. *Canadian Modern Language Review* 69(3). 324–348. doi:10.3138/cmlr.1748.

Préfontaine, Yvonne. 2013b. Fluency in French : A psycholinguistic study of second language speech production and perception. Lancaster: Lancaster University Ph.D. Thesis. (11 January, 2017).

Préfontaine, Yvonne & Judit Kormos. 2015. The Relationship Between Task Difficulty and Second Language Fluency in French: A Mixed Methods Approach. *The Modern Language Journal* 99(1). 96–112. doi:10.1111/modl.12186.

Préfontaine, Yvonne & Judit Kormos. 2016. A qualitative analysis of perceptions of fluency in second language French. *International Review of Applied Linguistics in Language Teaching* 54(2). 151–169. doi:10.1515/iral-2016-9995.

Préfontaine, Yvonne, Judit Kormos & Daniel Ezra Johnson. 2015. How do utterance measures predict raters perceptions of fluency in French as a second language? *Language Testing* 33(1). 53–73. doi:10.1177/0265532215579530.

Pye, Clifton, Kim A. Wilcox & Kathleen A. Siren. 1988. Refining transcriptions: the significance of transcriber 'errors.' *Journal of Child Language* 15(01). 17. doi:10.1017/S0305000900012034.

Quené, Hugo. 2008. Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America* 123(2). 1104–1113. doi:10.1121/1.2821762.

Quené, Hugo. 2013. Longitudinal trends in speech tempo: The case of Queen Beatrix. *The Journal of the Acoustical Society of America* 133(6). EL452–EL457. doi:10.1121/1.4802892.

Quirk, Randolph, Geoffrey Neil Leech, Sidney Greenbaum & David Crystal (eds.). 1995. *A comprehensive grammar of the English language*. London: Longman.

Raupach, Manfred. 1980. Temporal variables in first and second language speech production. In Manfred Raupach, Hans-Wilhelm Dechert & Frieda Goldman-Eisler (eds.), *Temporal variables in speech*, 271–285. (Janua Linguarum 86). The Hague: Mouton.

Raupach, Manfred. 1983. Analysis and evaluation of communication strategies. In C. Faerch & Gabriele Kasper (eds.), *Strategies in interlanguage communication*, 199–209. London: Longman.

Rehbein, Ines, Sören Schalowski & Heike Wiese. 2012. Annotating spoken language. *Best Practices for Speech Corpora in Linguistic Research*, 29. Istanbul, Turkey. http://www.corpora.uni-hamburg.de/lrec2012/Proceedings_Complete.pdf#page=35 (26 August, 2014).

Riazantseva, Anastasia. 2001. Second language proficiency and pausing: A Study of Russian Speakers of English. *Studies in Second Language Acquisition* 23(04). 497–526. doi:null.

Rietveld, Toni, Roeland Van Hout & Mirjam Ernestus. 2004. Pitfalls in Corpus Research. *Computers and the Humanities* 38(4). 343–362. doi:10.1007/s10579-004-1919-1.

Riggenbach, Heidi. 1991. Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes* 14(4). 423–441. doi:10.1080/01638539109544795.

Riggenbach, Heidi Ruth. 1989. *Nonnative Fluency in Dialogue Versus Monologue Speech: A Microanalytic Approach*. UMI.

Roberts, Benjamin & Kim Kirsner. 2000. Temporal cycles in speech production. *Language and Cognitive Processes* 15(2). 129–157. doi:10.1080/016909600386075.

Rodríguez, Luis Javier, Inés Torres & Amparo Varona. 2001. Annotation of disfluencies in Spanish dialogues. http://gtts.ehu.es/gtts/NT/fulltext/RodriguezEtalo1c.pdf (30 May, 2014).

Roekhaut, Sophie, Sandrine Brognaux, Richard Beaufort & Thierry Dutoit. 2014. eLite-HTS: a NLP tool for French HMM-based speech synthesis. Singapore.

Rohlfing, Katharina, Daniel Loehr, Susan Duncan, Amanda Brown, Amy Franklin, Irene Kimbara, Jan-Torsten Milde, et al. 2006. Comparison of multimodal annotation tools. *Gesprächsforschung* 7(100). http://www.gespraechsforschung-ozs.de/heft2006/tb-rohlfing.pdf (22 January, 2015).

Rohr, Jessica. 2017. A survey of prolongations in learner speech. *Proceedings of LCR2015*.

Romero-Trillo, Jesús. 2002. The pragmatic fossilization of discourse markers in non-native speakers of English. *Journal of Pragmatics* 34(6). 769–784.

Rose, Kenneth & Connie Ng. 2001. Inductive and deductive teaching of compliments and compliment responses. In Kenneth Rose & Gabriele Kasper (eds.), *Pragmatics in language teaching*, 145–170. Cambridge: Cambridge University Press.

Rose, Kenneth R. 2000. An exploratory cross-sectional study of interlanguage pragmatic development. *Studies in Second Language Acquisition* 22(01). 27–67. doi:10.1017/S0272263100001029.

Rose, R. 2015. Temporal Variables in First and Second Language Speech and Perception of Fluency. *Proceedings of the 18th International Congress of Phonetic Sciences*, 0405–1.

Glasgow, Scotland. http://www.roselab.sci.waseda.ac.jp/resources/file/2015_ichps_rose_slides.pdf (17 May, 2017).

Rose, Ralph. 2011. Investigating the Relationship between Hesitation Phenomena and L2 Accentedness. Lódz. http://www.roselab.sci.waseda.ac.jp/resources/file/accents_2011_rose_hp_and_l2_accentedness_slides.pdf (17 May, 2017).

Rose, Ralph. 2015. Fluidity: Real-time feedback for speaking fluency development. *The English Language Education Society of Japan (JELES) Annual Meeting*. Waseda University, Tokyo, Japan.

Rose, Ralph L. 1998. The communicative value of filled pauses in spontaneous speech. University of Birmingham Unpublished MA thesis. http://www.roselab.sci.waseda.ac.jp/resources/file/madissertation.pdf (17 May, 2017).

Rose, Ralph L. 2013. Crosslinguistic corpus of hesitation phenomena: a corpus for investigating first and second language speech performance. *INTERSPEECH*, 992–996. http://www.roselab.sci.waseda.ac.jp/resources/file/2013_interspeech_rose_cchp_final.pdf (17 May, 2017).

Rossiter, Marian J. 2009. Perceptions of L2 Fluency by Native and Non-native Speakers of English. *Canadian Modern Language Review* 65(3). 395–412. doi:10.3138/cmlr.65.3.395.

Rousseeuw, Peter. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20. 53–65.

Sajavaara, Kari. 1987. Factors affecting fluency. In Hans-Wilhelm Dechert & Manfred Raupach (eds.), *Psycholinguistic models of production*, 45–66. Norwood (N.J.): Ablex Publishing Corporation.

Santorini, Beatrice. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). (23 January, 2015).

Schachter, Stanley, Nicholas Christenfeld, Bernard Ravina & Frances Bilous. 1991. Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology* 60(3). 362.

Schegloff, Emanuel A., Gail Jefferson & Harvey Sacks. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53(2). 361. doi:10.2307/413107.

Schiffrin, Deborah. 1987. *Discourse Markers*. (Studies in Interactional Sociolinguistics 5). Cambridge: Cambridge University Press.

Schmidt, Thomas. 2001. The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse. *Proceedings of the IRCS Workshop On Linguistic Databases, Philadelphia*, 11–13. http://www1.uni-hamburg.de/exmaralda/Daten/4D-Literatur/Vortraege-Dokumente/IRCS_Paper.pdf (17 April, 2014).

Schmidt, Thomas. 2003. *Visualising linguistic annotation as interlinear text*. Sonderforschungsbereich 538. http://www1.uni-hamburg.de/exmaralda/files/Visualising-final.pdf (11 March, 2014).

Schmidt, Thomas & Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.), *The Oxford Handbook of Corpus Phonology*, 402–419. Oxford University Press.

Schmitt, Norbert. 2000. *Vocabulary in Language Teaching*. Cambridge University Press.

Schourup, Lawrence. 1985. *Common discourse particles in English conversation*. New York: Garland. http://catalog.hathitrust.org/api/volumes/oclc/12135530.html (26 August, 2013).

Schourup, Lawrence. 1999. Discourse markers. Tutorial overview. *Lingua* 107. 227–265.

Schourup, Lawrence. 2001. Rethinking "well." *Journal of Pragmatics* 33(7). 1025–1060.

Searle, John R. 1965. What is a speech act? In M. Black (ed.), *Philosophy in America*, 39–53. New York: Allen and Unwin.

Segalowitz, Norman. 2010. *Cognitive bases of second language fluency*. (Cognitive Science and Second Language Acquisition Series). New York ; London: Routledge.

Segalowitz, Norman & Barbara F. Freed. 2004. Context, Contact, and Cognition in Oral Fluency Acquisition: Learning Spanish in at Home and Study Abroad Contexts. *Studies in Second Language Acquisition* 26(2). 173–199.

Segalowitz, Norman & Jan H. Hulstijn. 2005. Automaticity in bilingualism and second language learning. In Judith F. Kroll & Annette M.B. de Groot (eds.), *Handbook of bilingualism. Psycholinguistic approaches*, 371–388. Oxford: Oxford University Press. https://www.researchgate.net/profile/Albert_Costa/publication/246682573_Lexical_access_in_bilingual_production/links/54ddcc1focf2814662eb6880.pdf#page=386.

Sert, Olcay. 2005. The Functions of Code-Switching in ELT Classrooms. *The Internet TESL Journal* 11(8).

Shaw, Stuart. 2004. IELTS writing: revising assessment criteria and scales (Phase 3). *Research Notes*(16). 3–7.

Shriberg, Elizabeth. 1994. Preliminaries to a Theory of Speech Disfluencies.

Shriberg, Elizabeth. 2001. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31(01). 153–169. doi:10.1017/S0025100301001128.

Shriberg, Elizabeth, Rebecca Bates & Andreas Stolcke. 1997. A prosody-only decision-tree model for disfluency detection. *Proc. EUROSPEECH*, 2383–2386.

Sinclair, John. 1992. The automatic analysis of corpora. In Jan Svartvik (ed.), *Directions in corpus linguistics: Proceedings of the Nobel Symposium 82*, 379–397. New York: Mouton de Gruyter.

Sinclair, John. 1996. Preliminary recommendations on corpus typology. http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html (31 January, 2017).

Skehan, Peter. 1997. Task characteristics, fluency, and oral performance testing. *Thames Valley Working Papers in Applied Linguistics*, vol. 5.

Skehan, Peter. 1998. Task-Based Instruction. *Annual Review of Applied Linguistics* 18. 268–286. doi:10.1017/S0267190500003585.

Skehan, Peter. 1999. *A cognitive approach to language learning*. (Oxford Applied Linguistics). Oxford: Oxford University Press. http://bib.uclouvain.be/opac/ucl/fr/chamo/chamo%3A796671?i=0 (13 October, 2016).

Skehan, Peter. 2003. Task-based instruction. *Language Teaching* 36(1). 1–14. doi:10.1017/S026144480200188X.

Skehan, Peter. 2009. Modelling Second Language Performance: Integrating Complexity, Accuracy, Fluency, and Lexis. *Applied Linguistics* 30(4). 510–532. doi:10.1093/applin/amp047.

Skehan, Peter (ed.). 2014. *Processing perspectives on task performance*. (Task-Based Language Teaching 5). Amsterdam: John Benjamins.

Skehan, Peter & Pauline Foster. 1997. Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research* 1(3). 185–211. doi:10.1177/136216889700100302.

Skehan, Peter & Pauline Foster. 1999. The Influence of Task Structure and Processing Conditions on Narrative Retellings. *Language Learning* 49(1). 93–120. doi:10.1111/1467-9922.00071.

Skehan, Peter & Pauline Foster. 2007. Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing research. (Ed.) S. Van Daele, Alex Housen, Folkert Kuiken, Michel Pierrard & Ineke Vedder. *Complexity, accuracy, and fluency in second language use, learning, and teaching*. 207–226.

Skehan, Peter & Pauline Foster. 2012. Complexity, accuracy, fluency and lexis in task-based performance. A synthesis of the Ealing project. In Alex Housen, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, 199–220. (Language Learning & Language Teaching 32). Amsterdam: Benjamins.

Skehan, Peter, Pauline Foster & Sabrina Shum. 2016. Ladders and Snakes in Second Language Fluency. *International Review of Applied Linguistics in Language Teaching* 54(2). 97–111. doi:10.1515/iral-2016-9992.

Sloetjes, Han & P. Wittenburg. 2008. Annotation by category - ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Spooren, Wilbert. 1997. The processing of underspecified coherence relations. *Discourse Processes* 24(1). 149–168. doi:10.1080/01638539709545010.

Spooren, Wilbert & Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory* 6(2). doi:10.1515/cllt.2010.009. http://www.degruyter.com/view/j/cllt.2010.6.issue-2/cllt.2010.009/cllt.2010.009.xml (10 November, 2016).

Staples, Shelley & Douglas Biber. 2015. Cluster Analysis. In Luke Plonsky (ed.), *Advancing Quantitative Methods in Second Language Research*, 243–274. (Second Language Acquisition Research Series). New York ; London: Routledge.

Stenström, Anna-Brita. 1994. *An introduction to spoken interaction*. (Learning about Language). London: Longman.

Susca, Michael & E.Charles Healey. 2002. Listener perceptions along a fluency–disfluency continuum: A phenomenological analysis. *Journal of Fluency Disorders* 27(2). 135–161. doi:10.1016/S0094-730X(02)00126-2.

Svartvik, Jan. 1990. *The London-Lund Corpus of Spoken English: Description and Researcg*. (Lund Studies in English 82). Lund: Lund University Press.

Swales, John. 2009. The concept of task. In Kris Van den Branden, Martin Bygate & John Michael Norris (eds.), *Task-based language teaching: a reader*, 41–55. (Task-Based Language Teaching : Issues, Research and Practice 1). Amsterdam: Benjamins.

Sweetlove, Douglas, Kevin Kato, Daniel Leigh Paller & Matthew Taylor. 2015. Fluency Practice on Mobile Phones: Implementing a Recording System and Following Students' Respons. 金城学院大学論集. 人文科学編 11(2). 69–76.

Swerts, Marc. 1998. Filled pauses as markers of discourse structure. *Journal of pragmatics* 30(4). 485–496.

Sykes, Julie M. 2013. Synchronous CMC and Pragmatic Development: Effects of Oral and Written Chat. *CALICO Journal* 22(3). 399–431.

Tabachnick, Barbara G. & Linda S. Fidell. 1989. *Using multivariate statistics*. 2nd ed. New York (N.Y.): Harper and Row.

Tanguy, Noalig, Thomas Van Damme, Liesbeth Degand & Anne-Catherine Simon. 2012. Projet FRFC "Périphérie gauche des unités de discours" - Protocole de codage syntaxique. https://hal.archives-ouvertes.fr/halshs-00762866/ (20 March, 2017).

Tarone, Elizabeth. 1983. Some thoughts on the notion of "communication strategy." In C. Faerch & Gabriele Kasper (eds.), *Strategies in interlanguage communication*. London: Longman.

Tarone, Elizabeth. 2005. Speaking in a second language. In Eli Hinkel (ed.), *Handbook of research in second language teaching and learning*, 485–502. Mahwah, N.J: Erlbaum.

Tavakol, Mohsen & Reg Dennick. 2011. Making sense of Cronbach's alpha. *International Journal of Medical Education* 2. 53–55. doi:10.5116/ijme.4dfb.8dfd.

Tavakoli, Parvaneh. 2009. Assessing L2 task performance: Understanding effects of task design. *System* 37(3). 482–495. doi:10.1016/j.system.2009.02.013.

Tavakoli, Parvaneh. 2011. Pausing patterns: differences between L2 learners and native speakers. *ELT Journal* 65(1). 71–79. doi:10.1093/elt/ccq020.

Tavakoli, Parvaneh. 2016. Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching* 54(2). 133–150. doi:10.1515/iral-2016-9994.

Tavakoli, Parvaneh & Peter Skehan. 2005a. Planning, task structure, and performance testing. In Rod Ellis (ed.), *Planning and task performance in a second language*, 239–273. Amsterdam: John Benjamins.

Tavakoli, Parvaneh & Peter Skehan. 2005b. Strategic planning, task structure, and performance testing. In Rod Ellis (ed.), *Language Learning & Language Teaching*, 239–273. (11). Philadelphia: John Benjamins Publishing Company.

Taylor, Ann, Mitchell Marcus & Beatrice Santorini. 2003. The Penn treebank: an overview. *Treebanks*, 5–22. Springer. http://link.springer.com/chapter/10.1007/978-94-010-0201-1_1 (30 May, 2014).

Taylor, Calvin W. 1947. A factorial study of fluency in writing. *Psychometrika* 12(4). 239–262.

Taylor, Lynda & Fiona Barker. 2008. Using Corpora for Language Assessment. In Nancy H. Hornberger (ed.), *Encyclopedia of Language and Education*, 2377–2390. Springer US. http://link.springer.com/referenceworkentry/10.1007/978-0-387-30424-3_179 (28 May, 2015).

Temple, Liz. 1992. Disfluencies in Learner Speech. *Australian Review of Applied Linguistics* 15(2). 29–44.

Temple, Liz. 2000. Second language learner speech production. *Studia Linguistica* 54(2). 288–297.

Thewissen, Jennifer. 2012. Accuracy across proficiency levels : Insights from an error-tagged EFL learner corpus. Louvain-la-Neuve (Belgium): Université catholique de Louvain.

Thompson, Irene. 1991. Foreign Accents Revisited: The English Pronunciation of Russian Immigrants. *Language Learning* 41(2). 177–204. doi:10.1111/j.1467-1770.1991.tb00683.x.

Tian, Ye. 2016. Trump Clinton First Debate – disfluency and smile. *Linguistics and more*. https://yetianlinguistics.blog/2016/09/28/trump-clinton-first-debate-disfluency-and-laughter/ (15 February, 2018).

Tibshirani, Robert, Guenther Walther & Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(2). 411–423.

Tonkyn, Alan. 2012. Measuring and perceiving changes in oral complexity, accuracy and fluency. In Alex Housen, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, 221–245. (Language Learning & Language Teaching 32). Amsterdam: Benjamins.

Tortel, Anne. 2008. ANGLISH. Une base de données comparatives de l'anglais lu, répété et parlé en L1 & L2. *TIPA. Travaux interdisciplinaires sur la parole et le langage*(27). 111–122. doi:10.4000/tipa.321.

Tottie, Gunnel. 2011. Uh and Um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics* 16(2). 173–197. doi:10.1075/ijcl.16.2.02tot.

Tottie, Gunnel. 2014. On the use of uh and um in American English. *Functions of Language* 21(1). 6–29. doi:10.1075/fol.21.1.02tot.

Towell, Richard. 1987. Approaches to the Analysis of the Oral Language Development of the Advanced Learner. In Richard Towell & James Francis Coleman (eds.), *The advanced language learner*, 157–181. London: AFLS/SULFRA.

Towell, Richard. 2002. Relative degrees of fluency: A comparative case study of advanced learners of French. *IRAL: International Review of Applied Linguistics in Language Teaching* 40(2). 117.

Towell, Richard. 2012. Complexity, accuracy and fluency from the perspective of psycholinguistic second language acquisition research. In Alex Housen, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 performance and Proficiency. Complexity, Accuracy and Fluency in SLA*, 47–70. (Language Learning & Language Teaching 32). Amsterdam & Philadelphia: John Benjamins Publishing Company.

Towell, Richard, R. Hawkins & N. Bazergui. 1996. The Development of Fluency in Advanced Learners of French. *Applied Linguistics* 17(1). 84–119. doi:10.1093/applin/17.1.84.

Trofimovich, Pavel & Wendy Baker. 2006. Learning Second Language Suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech. *Studies in Second Language Acquisition* 28(01). 1–30. doi:10.1017/S0272263106060013.

Trofimovich, Pavel & Wendy Baker. 2007. Learning prosody and fluency characteristics of second language speech: The effect of experience on child learners' acquisition of five suprasegmentals. *Applied Psycholinguistics* 28(2). 251–276. doi:10.1017/S0142716407070130.

Trosborg, Anna. 1995. *Interlanguage Pragmatics: Requests, Complaints, and Apologies*. Walter de Gruyter.

Tsao, Ying-Chiao, Gary Weismer & Kamran Iqbal. 2006. Interspeaker variation in habitual speaking rate: additional evidence. *Journal of speech, language, and hearing research: JSLHR* 49(5). 1156–1164. doi:10.1044/1092-4388(2006/083).

Uchihara, Takumi & Kazuya Saito. 2016. Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal*. 1–12. doi:10.1080/09571736.2016.1191527.

Uluman, Müge & Deha C. Doğan. 2016. Comparison of Factor Score Computation Methods In Factor Analysis. *Australian Journal of Basic and Applied Sciences* 10(18). 143–151.

Upshur, John A. & Carolyn E. Turner. 1999. Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing* 16(1). 82–111.

Valdmets, Annika. 2013. Modal particles, discourse markers, and adverbs with lt-suffix in Estonian. In Liesbeth Degand, Bert Cornillie & Paola Pietrandrea (eds.), *Discourse markers and modal particles. Categorization and description*, 107–132. Amsterdam: John Benjamins.

Vinay, Jean-Paul & Jean Darbelnet. 1995. *Comparative stylistics of French and English: a methodology for translation*. (Trans.) Juan C. Sager & M.-J. Hamel. (Benjamins Translation Library 11). Amsterdam: Benjamins.

Walker, James. 2013. Variation analysis. In Robert Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 440–459. Cambridge: Cambridge university press.

Watanabe, Michiko, Keikichi Hirose, Yasuharu Den & Nobuaki Minematsu. 2008. Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication* 50(2). 81–94. doi:10.1016/j.specom.2007.06.002.

Watts, Richard J. 1988. A relevance-theoretic approach to commentary pragmatic markers: The case of "actually", "really" and "basically". *Acta Linguistica Hungarica* 19. 37–60.

Webber, Bonnie. 2001. Computational perspectives on discourse and dialogue. In Deborah Schiffrin, Deborah Tannen & Heidi E. Hamilton (eds.), *The Handbook of Discourse Analysis*, 798–817. Oxford: Blackwell Publishing Ltd. doi:10.1111/b.9780631205968.2003.00018.x. http://doi.wiley.com/10.1111/b.9780631205968.2003.00018.x (30 January, 2017).

Weir, C. 1993. *Understanding and developing language tests*. New York ; London: Prentice Hall International.

Wennerstrom, Ann. 2000. The role of intonation in second language fluency. In Heidi Riggenbach (ed.), *Perspectrives on Fluency*, 102–127. Ann Arbor: University of Michigan Press.

Whishaw, Ian Q., Lori-Ann R. Sacrey, Scott G. Travis, Gita Gholamrezaei & Jenni M. Karl. 2010. The functional origins of speech-related hand gestures. *Behavioural Brain Research* 214(2). 206–215. doi:10.1016/j.bbr.2010.05.026.

White, Ron. 1997. Back channelling, repair, pausing, and private speech. *Applied Linguistics* 18(3). 314–344. doi:10.1093/applin/18.3.314.

Wigglesworth, Gillian & Cathie Elder. 2010. An Investigation of the Effectiveness and Validity of Planning Time in Speaking Test Tasks. *Language Assessment Quarterly* 7(1). 1–24. doi:10.1080/15434300903031779.

Wilson, Thomas & Don Zimmerman. 1986. The structure of silence between rums in two-party conversation. *Discourse Processes* 9(4). 375–390.

Winke, Paula, Susan Gass & Carol Myford. 2013. Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30(2). 231–252. doi:10.1177/0265532212456968.

Wisniewski, Katrin. 2015. Empirical validity evidence for "Common European Framework of Reference" scales. Conference Presentation. Paper presented at the EALTA Conference. http://www.ealta.eu.org/conference/2015/presentations/Saturday/22.0.11/Wisniewski.pdf (6 July, 2016).

Wisniewski, Katrin. 2017. Disentangling CEFR scale validity: Level descriptions, learner language, and human ratings: An analysis of the empirical validity of the B2 vocabulary control level description. In Pieter De Haan, Rina De Vries & Sanne Van Vuuren (eds.), *Language, Learners and Levels: progression and Variation*. (Corpora and Language in Use).

Witton-Davies, Giles. 2010. The Role of Repair in Oral Fluency and Disfluency. *Selected papers from the nineteenth international symposium on English teaching*, 119–129. Taipei, Taiwan: Crane.

Witton-Davies, Giles. 2014. The study of fluency and its development in monologue and dialogue. Lancaster: Lancaster University Unpublished PhD thesis. http://www.forex.ntu.edu.tw/en/files/writing/4092_dc0088cd.pdf (18 January, 2017).

Wood, David Claude. 2010. *Formulaic language and second language speech fluency: background, evidence and classroom applications*. London: Continuum.

Wright, T. 1987. Instructional task and discoursal outcome in the L2 classroom. In C.N. Candlin & D.F. Murphy (eds.), *Language learning tasks*, vol. 7, 47–68. (Lancaster Practical Papers in English Language Education). Englewood Cliffs, NJ: Prentice Hall International.

Xiao, Richard. 2009. Theory-driven corpus research: Using corpora to inform aspect theory. In Anke Lüdeling & M. Kytö (eds.), *Corpus-linguistics: An international handbook*, 987–1008. Berlin/New York: Mouton de Gruyter.

Young, Steve J., Gunnar Evermann, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Daniel Povey, Valtcho Valtchev & Philip C. Woodland. 2013. *The HTK book (for HTK Version 3.2.1)*.

Yuan, Fangyuan & Rod Ellis. 2003. The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics* 24(1). 1–27. doi:10.1093/applin/24.1.1.

Chapter 9 APPENDICES

9.1 LINDSEI AND LOCNEC TRANSCRIPTION CONVENTIONS

Section delimitation	
Interview identification and delimitation	• Each interview is preceded by a code of the type: <h nt="FR" nr="FR+three-digit number"> *e.g. <h nt="FR" nr="FR004"> (4th interview with French mother tongue student)* • All interviews ends with the following tag (on a separate line): </h>
Task delimitation	• The three tasks making up the interview are marked by: o <S> (before the set topic), </S> (after the set topic), o <F> (before the free discussion), </F> (after the free discussion), o <P> (before the picture description), </P> (after the picture description). • These tags occupy a separate line and do not interrupt a turn. *e.g. <S>* *<A> did you . manage to choose a topic </A>*
Speaker turns	• Speaker turns are displayed in vertical format, i.e. one below the other. o <A> and </A> signify the beginning and end of the interviewer's turn; o <B> and </B> indicate the interviewee's turn. *e.g. <A> okay so which topic have you chosen </A>* *<B> the film or play that I thought was particularly good or bad really </B>*
General transcription conventions	
Punctuation	No punctuation marks are used to indicate sentence or clause boundaries.
Spelling and capitalization	British spelling conventions are followed.

	Capital letters are only kept when required by spelling conventions (proper names, I, Mrs etc.) – not at the beginning of turns.
Acronyms	• Pronounced as sequences of letters: transcribed as a series of upper-case letters separated by spaces. *e.g. <B> yes not really I did sort of basic G C S E French and German </B>* • Pronounced as words: transcribed as a series of upper-case letters not separated by spaces. *e.g. <A> (mhm) (er) you're doing a MAELT </A>*
Dates and numbers	Figures are written out in words. *e.g. <B> an awful lot of people complain and say well the grants were two thousand two hundred </B>*
Unclear passages	• <X> represents an unclear syllable or sound up to one word; • <XX> represents two unclear words; • <XXX> represents more than two words. *e.g. <B> <X> they're just begging <XX> there's there's honestly he did a course .. for a few weeks </B>* • If transcribers are not entirely sure of a word or word ending, they indicate this by having the word directly followed by the symbol <?>. *e.g. <B> I went to see a<?> friend at university there and stayed </B>* • Unclear names of towns or titles of films for example are indicated as <name of city> or <title of film>. *e.g. <B> where else did we go (er) <name of city> it's in Bolivia </B>*
Anonymisation	Data are anonymised: transcribers can use tags like <first name of interviewee>, <first name and full name of interviewer> or <name of professor> to replace names. *e.g. <A> I'm <first name of interviewer> . what's your name </A>*
Transcription of the features of spoken language	
Empty pauses	• One dot for a "short" pause (< 1 second); • Two dots for a "medium" pause (1-3 seconds);

	• Three dots for "long" pauses (> 3 seconds[178]). *e.g. <B> (erm) .. it's a British film there aren't many of those these days </B>*
Filled pauses and backchannelling	• Between brackets; • Marked as (eh) [brief], (er), (em), (erm), (mm), (uhu) and (mhm). *e.g. <B> yeah . well Namur was warmer (er) it was (eh) a really little town </B>*
Truncated word	Truncated words are immediately followed by an equals sign. *e.g. <B> it still resem= resembled the theatre </B>*
Contracted forms	All standard contracted forms are retained.
Non-standard forms	Non-standard forms that appear in the dictionary are transcribed orthographically in their dictionary accepted way: *cos*, *dunno*, *gonna*, *gotta*, *kinda*, *wanna* and *yeah*.
Foreign words and pronunciation	Foreign words are indicated by <foreign> (before the word) and </foreign> (after the word). *e.g. <B> we couldn't go with (er) knives and so on <foreign> enfin </foreign> we were (er) </B>* As a rule, foreign pronunciation is not noted, except in the case where the foreign word and the English word are identical. If in this case the word is pronounced as a foreign word, this is also marked using the <foreign> tag. *e.g. <B> I didn't have the (erm) . <foreign> distinction </foreign> </B>*
Phonetic features	(a) Syllable lengthening A colon is added at the end of a word to indicate that the last syllable is lengthened. Colons are not be inserted within words. *e.g. <B> that's something I'll I'll plan to: to learn </B>* (b) Articles

[178] Note, however, that the length of unfilled pauses is generally subjectively appreciated.

	• when pronounced as [ei], the article a is transcribed as a[ei]; *e.g. <B> and it's about (erm) . life in a[ei] (eh) public school in America I think </B>* • when pronounced as [i:], the article the is transcribed as the[i:]. *e.g. <B> and the[i:] villa we were staying in was in one of the valleys </B>*
Overlapping speech	The tag <overlap /> indicates the beginning of overlapping speech. It is indicated in both turns. The end of overlapping speech is not indicated. *e.g. <B> yeah I went on a bus to London once and I'll never <overlap /> do it again </B>* *<A> <overlap /> that's even worse </A>*
Contextual information	
Voice quality	If a particular stretch of text is said laughing or whispering for instance, this is marked by inserting <starts laughing> or <starts whispering> immediately before the specific stretch of speech and <stops laughing> or <stops whispering> at the end of it. *e.g. <B> <starts laughing> I don't have to assess it I only have to write it <stops laughing> </B>*
Non-verbal vocal sounds	Nonverbal vocal sounds are enclosed between angle brackets. *e.g. <B> I hope so I've I've got some <coughs> friends out there </B>* *e.g. <B> so I went back into Breda .. and sat down again <imitates the sound of a guitar> </B>*
Contextual comments	Non-linguistic events are indicated between angle brackets only if they are deemed relevant to the interaction (if one of the participants reacts to it, for example). *e.g. <A> no it's true it's nice to have your own bathroom </A>* *<somebody enters the room>* *<B> hi </B>*

Table 9-1: LINDSEI and LOCNEC transcription conventions (from Gilquin et al. 2010)

9.2 ARC SITUATIONAL VARIABLES

Table 9-2 shows the list of "situational variables" created by the four PhD students of the ARC Fluency project, namely (in alphabetical order) Ludivine Crible, Amandine Dumont, Iulia Grosman, and Ingrid Notarrigo. This list aims to provide a characterisation of speaking tasks to allow comparisons between (seemingly different) genres.

Situational variable	Description
D.ELICIT	**Degree of elicitation**
	The presence and weight of the experimental protocol as a constraint on the interaction.
SUPERVISED	Artificial production due to a rigid experimental protocol as part of scientific research which heavily constrains both the speaker's production and linguistic variables.
SEMI-SUPERVISED	Natural production in the framework of a flexible experimental protocol as part of scientific research which monitors the choice of the topic but allows the interlocutor to choose his/her own wording.
NATURAL	Authentic production out of experimental protocol, i.e. not generated by scientific research.
NB.SPK	**Number of speakers**
	The number of speakers actively taking part in the interaction
MONOLOGUE	There is one main speaker.
DIALOGUE	There are two main speakers.
MULTILOGUE	More than two speakers interact.
D.PREP	**Degree of preparedness**
	The extent of (spoken and/or written) preparation of the main speaker's discourse.
PREPARED	The speaker has entirely prepared both content and form of their speech. The speech was scripted and may be produced with textual support.
SEMI.PREPARED	The speaker has prepared the general frame of their speech, but the speech has not been fully scripted. The interaction may, however, include a visual support.
SPONTANEOUS	The speaker has not prepared their speech and improvises spontaneously as they speak.
D.INTERACT	**Degree of interactivity**
	The speakers' ability to adapt their speaking behavior to the other interlocutor(s) with respect to what is expected from their status in the interaction.
INTERACTIVE	Symmetrical relationship between the speakers. The situation allows all speakers to speak and hold the floor.
SEMI.INTERACTIVE	Asymmetrical relationship where one speaker holds the floor more than the others. The situation does not exclude punctual interventions from secondary speakers.
NON.INTERACTIVE	Asymmetrical relationship where one speaker keeps the floor nearly continuously, without leaving turn-taking opportunities for the other speakers.
D.MEDIA	**Degree of media coverage**
	The extent of broadcasting as the main aim of the interaction.

MEDIA	Broadcasting is the main aim of the interaction.
SEMI.MEDIA	The interaction would take place even without broadcasting.
NOT.MEDIA	The interaction is not broadcasted.
C.PRO	**Category of work-relatedness**
	Whether the situation is due to one speaker's professional activity or not
PRO	The situation is caused by one speaker's professional activity.
NON.PRO	The situation is not caused by one speaker's professional activity.

Table 9-2: ARC situational variables

9.3 CROSS-THESIS METADATA HOMOGENISATION

Table 9-3 to 9-5 provide generic categories of metadata for the corpora, texts, and speakers.

COR.MTD	**Corpus metadata**
COR.ID	Name of the corpus
SUBCOR.ID	Name of the source corpus
COR.AUTHOR	Name of the author of the corpus
COR.CONTACT.AUTHOR	Contact information of the author (email)
COR.DATE	Date of the creation of the corpus
COR.RIGHTS	Rights of use and/or license of the corpus
COR.INSTIT	Institution where the corpus was created
COR.DURATION	Total duration of the corpus (hh:mm:ss)
COR.NB.WORDS	Total number of words (tokens) in the corpus

Table 9-3: Corpus metadata homogenization

TXT.MTD	**Metadata of the transcription**
TXT.ID	ID of the text
TXT.COR.ID	Corpus name
SUB.CORPUS	Subcorpus the transcription comes from
DATE.RECORDING	Recording date (YYYY-MM-DD)
URL	Direct url to the sound or video if applicable
RIGHTS	Copyright/license
TXT.DURATION	Duration of the recording file (mm,ss)
TXT.NB.WORDS	Number of words (tokens) in the transcription
TXT.FILES	List of files associated to the text (.wav, .TextGrid,.ebt,.elan etc.)

Table 9-4: Metadata of the transcription files

SPK.MTD	**Speaker metadata**
SPK.CODE	Speaker ID in the corpus
SPK.TXT.ID	Text ID(s) in which the speaker participates
AGE.N	Speaker's age at the time of the recording
SPK.AGE	Age groups to which belongs the speaker at the time of the recording
18-25	The speaker belongs to the 18 to 25 year-old group.
26-45	The speaker belongs to the 26 to 45 year-old group.
46-65	The speaker belongs to the 46 to 65 year-old group.
66+	The speaker belongs to the 66+ year-old group.
?	The speaker's age is unknown.
SPK.GENDER	Speaker's gender
F	Female
M	Male
SPK.NATIONALITY	Home country of the speaker
SPK.1LG	Speaker's mother tongue

SPK.OTHERLG	Other languages spoken by the speaker
SPK.JOB	Whether the speaker works in the field of communication
COM	The speaker works in information and/or communication
NOT.COM	The speaker doesn't work in information and/or communication
?	The speaker's occupation is unknown.

Table 9-5: Speaker metadata

9.4 EDITING LINDSEI-FR+ AND LOCNEC+

Table 9-6 offers a full account of the checks and modifications LINDSEI-FR+ and LOCNEC+ underwent before and after their time alignment.

<table>
<tr><th></th><th>Description of the modification</th><th>Corpus</th></tr>
<tr><td colspan="3">BEFORE ALIGNMENT</td></tr>
<tr><td>1.</td><td>Save all the transcription files in .txt format</td><td>LOCNEC</td></tr>
<tr><td>2.</td><td>Save all the transcription files in UTF-8</td><td>LINDSEI-FR
LOCNEC</td></tr>
<tr><td>3.</td><td>Check and/or end transcriptions (esp. unclear passages)</td><td>LOCNEC</td></tr>
<tr><td>4.</td><td>Check the mark-up (cf. LINDSEI transcription guidelines)</td><td>LOCNEC</td></tr>
<tr><td>5.</td><td>Check spelling in the transcriptions:
1. missing or inverted letters;
2. missing spaces;
3. curled apostrophes instead of straight apostrophes
4. "…" [one symbol with three dots] instead of "…" [three dot symbols];
5. British vs. American spelling;
6. etc.</td><td>LINDSEI-FR
LOCNEC</td></tr>
<tr><td>6.</td><td>Create an excel document including the precise timings for the beginning and end of each speaking task.
• If the speaking task lasts less than c. 5 minutes: no change in the tags in the transcription, simply write up the precise timings of the beginning and end of the task in the excel document.
e.g.
<table><tr><td>Interview</td><td><S></td><td></S></td></tr><tr><td>FR001</td><td>00:00:00:000</td><td>00:02:51:905</td></tr></table>• If the duration of a given speaking tasks exceeds c. 5 minutes, split the task into sub-parts of less than 5 minutes by inserting tags in the transcription (<F1> </F1>; <F2> </F2>; <F3> </F3>). Insert the tags preferably in the middle of a silent pause in order not to distort the speech proper, or between speaker turns. Write up the precise timings of the beginning and end of each subtask in the excel document.
e.g.
<table><tr><td>Interview</td><td><F1></td><td></F1></td><td><F2></td><td></F2></td></tr><tr><td>FR001</td><td>00:02:51:900</td><td>00:07:02:911</td><td>00:07:02:911</td><td>00:13:40 :100</td></tr></table></td><td>LINDSEI-FR
LOCNEC</td></tr>
</table>

	Note that the time of the end of a (sub)task must always correspond to the beginning of next (sub)task.	
7.	Final check of the mark-up in transcriptions: 1. missing opening or closing brackets 2. missing or extra blank spaces 3. \ instead of / 4. missing closing turn tag (</A> or </B>), or task tag (</S[*n*]>; </F[*n*]>) 5. unclear passages 6. inconsistent tags, esp. paraverbal information (ex: <laughs>, <laughing>, <starts laughing>, <giggle>, <giggles>, <giggling>) 7. missing or extra overlapping tags 8. erroneous transcription of silent pauses as markers of the end of overlapping speech 9. …	LINDSEI-FR LOCNEC
POST-ALIGNMENT CORRECTIONS (i.e. during the annotation process)		
8.	Check/correct segmentation errors e.g. don't => don|'t	LINDSEI-FR LOCNEC
9.	Correct alignment errors, with specific focus on the boundaries and measurement of silent pauses	LINDSEI-FR LOCNEC
10.	Correct alignment bugs	LINDSEI-FR LOCNEC
11.	Correct overlapping speech (esp. end of overlapping speech)	LINDSEI-FR LOCNEC
12.	Check unclear passages	LINDSEI-FR LOCNEC
13.	Annotation of (dis)fluency features	LINDSEI-FR LOCNEC

Table 9-6: Editing LINDSEI-FR+ and LOCNEC+ for the time-alignement

9.5 EXMARALDA ANNOTATION SPECIFICATION

```
<annotation-specification>
<annotation-set exmaralda-tier-category="LEVEL 1">
<category name="DM">
<category name="DM monogram">
<tag name="&#60;DM&gt;"/>
</category>
<category name="DM beginning">
<tag name="&#60;DM"/>
</category>
<category name="DM middle">
<tag name="DM"/>
</category>
<category name="DM end">
<tag name="DM&gt;"/>
</category>
</category>
<category name="C">
<tag name="&#60;C&gt;"/>
</category>
<category name="Foreign Word">
<tag name="&#60;W&gt;"/>
</category>
<category name="Truncated word">
<category name="abandonned">
<tag name="&#60;T&gt;"/>
</category>
<category name="T completed">
<tag name="&#60;T"/>
</category>
<category name="T completion">
<tag name="T&gt;"/>
</category>
</category>
<category name="Repetition">
<category name="Repeatable 1">
<tag name="&#60;R0"/>
</category>
<category name="Repeatable 2">
<tag name="R0"/>
</category>
<category name="Repeated">
<category name="Repeated 1">
<tag name="R1"/>
</category>
<category name="Repeated 1 end">
<tag name="R1&gt;"/>
</category>
<category name="Repeated 2">
<tag name="R2"/>
</category>
<category name="Repeated 2 end">
<tag name="R2&gt;"/>
</category>
<category name="Repeated 3">
<tag name="R3"/>
</category>
<category name="Repeated 3 end">
<tag name="R3&gt;"/>
</category>
</category>
<category name="Embracing repetition">
<category name="beginning">
<tag name="&#60;RE0"/>
</category>
```

```
<category name="end">
<tag name="RE1&gt;"/>
</category>
</category>
</category>
<category name="Restart">
<category name="RS 1 word">
<tag name="&#60;RS&gt;"/>
</category>
<category name="RS beginning">
<tag name="&#60;RS"/>
</category>
<category name="RS middle">
<tag name="RS"/>
</category>
<category name="RS end">
<tag name="RS&gt;"/>
</category>
</category>
<category name="False start">
<tag name="&#60;FS&gt;"/>
</category>
</annotation-set>
<annotation-set exmaralda-tier-category="LEVEL 2">
<category name="Substitution morphosyntaxique">
<tag name="&#60;SM&gt;"/>
</category>
<category name="Substitution propositionnelle">
<tag name="&#60;SP&gt;"/>
</category>
<category name="Insertion">
<category name="Insertion 1 word">
<tag name="&#60;I&gt;"/>
</category>
<category name="Insertion beginning">
<tag name="&#60;I"/>
</category>
<category name="Insertion middle">
<tag name="I"/>
</category>
<category name="Insertion end">
<tag name="I&gt;"/>
</category>
</category>
<category name="Deletion">
<tag name="&#60;Del&gt;"/>
</category>
<category name="Nesting">
<tag name="&#60;N&gt;"/>
</category>
<category name="Order">
<category name="Order beginning">
<tag name="&#60;Or"/>
</category>
<category name="Order middle">
<tag name="Or"/>
</category>
<category name="Order end">
<tag name="Or&gt;"/>
</category>
</category>
</annotation-set>
<annotation-set exmaralda-tier-category="LEVEL 3 DESCRIPTION">
<category name="Editing term beginning">
<tag name="&#60;ET"/>
</category>
<category name="Editing term middle">
<tag name="ET"/>
```

```
</category>
<category name="Editing term end">
<tag name="ET&gt;"/>
</category>
<category name= "Recycling">
<tag name="G"/>
</category>
</annotation-set>
<annotation-set exmaralda-tier-category="PROSODY">
<category name ="UP">
<tag name="&#60;UP&gt;"/>
</category>
<category name="FP">
<tag name="&#60;FP&gt;"/>
</category>
<category name="Lengthening">
<tag name="&lt;L&gt;"/>
</category>
</annotation-set>
<annotation-set exmaralda-tier-category="SYMBOLS">
<category name="multi-tag">
<tag name="+"/>
</category>
<category name="opening bracket">
<tag name="&#60;"/>
</category>
<category name="closing bracket">
<tag name="&gt;"/>
</category>
</annotation-set>
</annotation-specification>
```

9.6 TESTING THE NORMALITY OF THE DATA

Table 9-7 displays the results of the Kolmogorov-Smirnov and the Shapiro-Wilk tests for the 14 (dis)fluency variables in LINDSEI-FR+ and LOCNEC+. Significant results are shown in bold font.

(Dis)fluency variable	Corpus	Kolmogorov-Smirnov	Shapiro-Wilk
C	LINDSEI-FR+	.200	**.009**
	LOCNEC+	**.015**	**.000**
DM	LINDSEI-FR+	**.008**	**.000**
	LOCNEC+	.200	.073
FP	LINDSEI-FR+	.200	.533
	LOCNEC+	**.001**	**.000**
FS	LINDSEI-FR+	.200	.135
	LOCNEC+	.200	.442
L	LINDSEI-FR+	.200	.219
	LOCNEC+	.145	.132
MLR	LINDSEI-FR+	**.029**	**.026**
	LOCNEC+	**.005**	**.000**
MLUP	LINDSEI-FR+	**.028**	.072
	LOCNEC+	.200	.061
PTR	LINDSEI-FR+	.200	.112
	LOCNEC+	.200	.078
Rep	LINDSEI-FR+	.200	.296
	LOCNEC+	**.007**	**.000**
RS	LINDSEI-FR+	.200	.434
	LOCNEC+	**.021**	**.046**
SR	LINDSEI-FR+	.200	.313
	LOCNEC+	.200	.577
T	LINDSEI-FR+	.082	.098
	LOCNEC+	**.024**	**.000**
UP	LINDSEI-FR+	.200	.676
	LOCNEC+	.148	.052
W	LINDSEI-FR+	**.000**	**.000**
	LOCNEC+	**.000**	**.000**

Table 9-7: Kolmogorov-Smirnov and Shapiro-Wilk test results in LINDSEI-FR+ and in LOCNEC+

The Shapiro-Wilk and the Kolmogorov-Smirnov tests are used to test the assumption that the sample data are drawn from a normally-distributed population. If the results are significant (i.e. $p < 0.05$), the assumption of normality of the distribution should be rejected. Both tests are, however, sensitive to sample size and, as strongly advised by Field (2013:185), should always be used in conjunction with a visual inspection of histograms and skewness and kurtosis measures to make an informed decision about the extent of non-normality based on converging evidence.

In this case, normality of distribution should not be assumed for some (dis)fluency variables (*cf.* bold figures in Table 9-7). However, after a visual inspection of the data, it appeared that departures from normality were not substantive. Moreover, often, only one "version" of the

(dis)fluency variable (that is, the LINDSEI-FR+ or the LOCNEC+ version of the variable) does not meet the assumption of normality of distribution.

9.7 PRINCIPAL COMPONENTS ANALYSIS

The scree plots of the eigenvalues of the learner and native speaker components are displayed in Figure 9-1 and Figure 9-2, respectively. Table 9-8 and Table 9-9 present the factor loadings of the (dis)fluency variables on each component before orthogonal rotation in LINDSEI-FR+ and in LOCNEC+, respectively.

9.7.1 LINDSEI-FR+

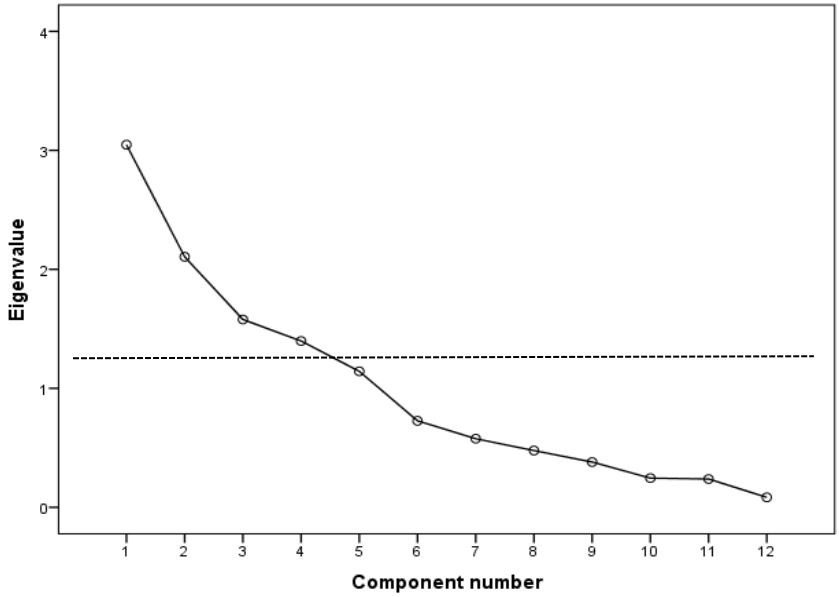


Figure 9-1: Scree plot for the final Principal Components Analysis in LINDSEI-FR+

(Dis)fluency variables	Factor loadings				
	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Unfilled pauses	**-.939**	.019	-.003	-.122	.108
Mean length of runs	**.885**	.213	-.080	-.004	.110
Phonation-time ratio	**.839**	-.053	-.173	.286	.099
Speech rate	**.784**	-.061	.301	-.175	-.096
Restarts	.040	**.871**	.074	.074	.053
Truncations	-.040	**.757**	-.051	-.218	-.345
Repetitions	-.004	**.572**	.101	**-.464**	-.011
Filled pauses	.040	.386	**-.769**	-.046	.081
Discourse markers	.163	-.093	**.634**	**-.530**	.078
False starts	-.115	.376	**.627**	**.456**	.003
Conjunctions	-.106	.296	.170	.391	**.763**

Foreign words	-.074	.091	.135	**.598**	**-.619**

Table 9-8: Factor loadings before orthogonal rotation in LINDSEI-FR+
Note: Loadings over .40 appear in bold; the variables are ranked in decreasing order of loading

9.7.2 LOCNEC+

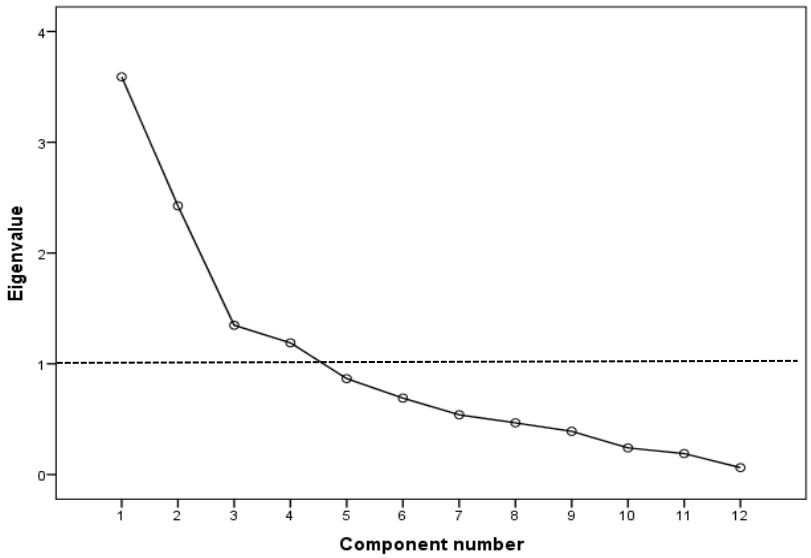


Figure 9-2: Scree plot for the final Principal Components Analysis in LOCNEC+

(Dis)fluency variables	Factor loadings			
	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Unfilled pauses	**.862**	-.154	-.118	.041
Phonation time ratio	**-.833**	.137	.295	.074
Mean length of runs	**-.767**	.350	.254	-.089
Speech rate	**-.743**	.307	.032	-.193
Filled pauses	**.686**	.129	.314	.215
Repetitions	**.489**	**.412**	.305	-.152
Restarts	.223	**.824**	-.178	-.225
Truncations	.303	**.753**	-.175	**-.413**
Lengthenings	.250	**.570**	**.442**	.130
Discourse markers	-.294	.241	**-.733**	.152
Connectors	-.114	.378	.172	**.755**
False starts	.030	**.455**	**-.406**	**.486**

Table 9-9: Factor loadings before orthogonal rotation in LOCNEC+
Note: Loadings > .40 appear in bold; the variables are ranked in decreasing order of loading

Figure 9-3 below is a scatterplot of the component scores in LINDSEI-FR+. Figure 9-4 is a scatterplot of the component scores in LOCNEC+. They show that there is no linear relationship between the components.

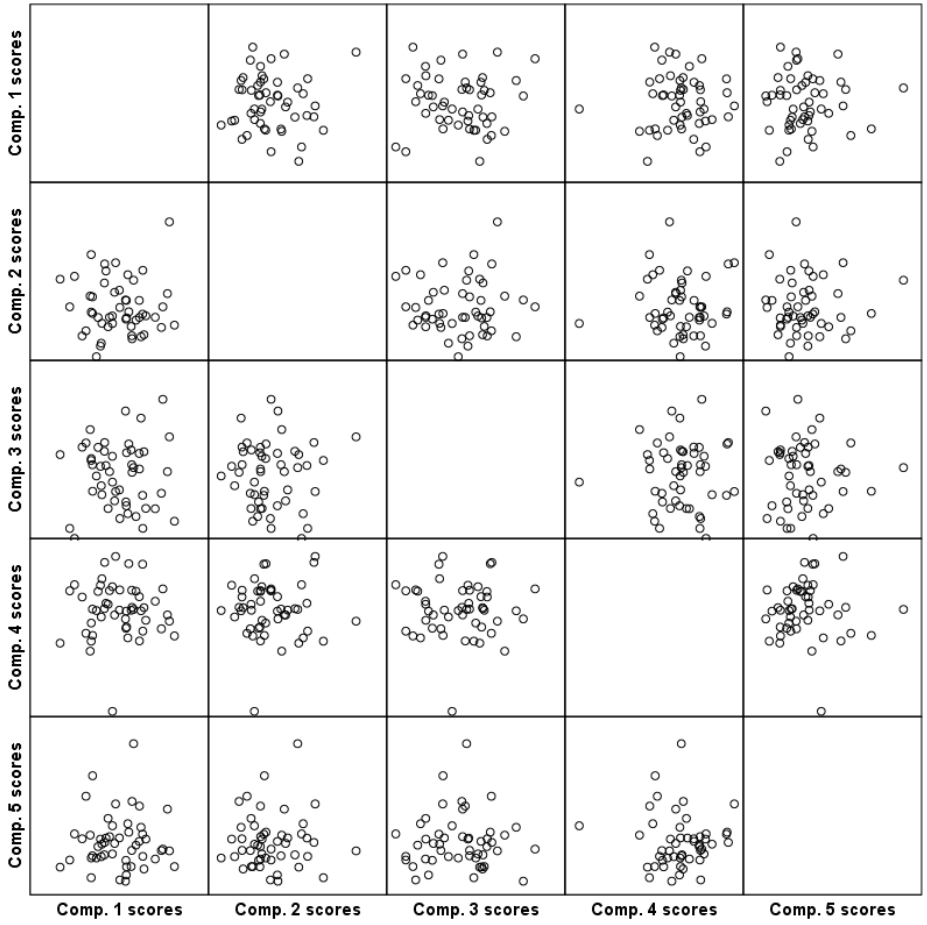


Figure 9-3: Scatterplots of component scores in LINDSEI-FR+

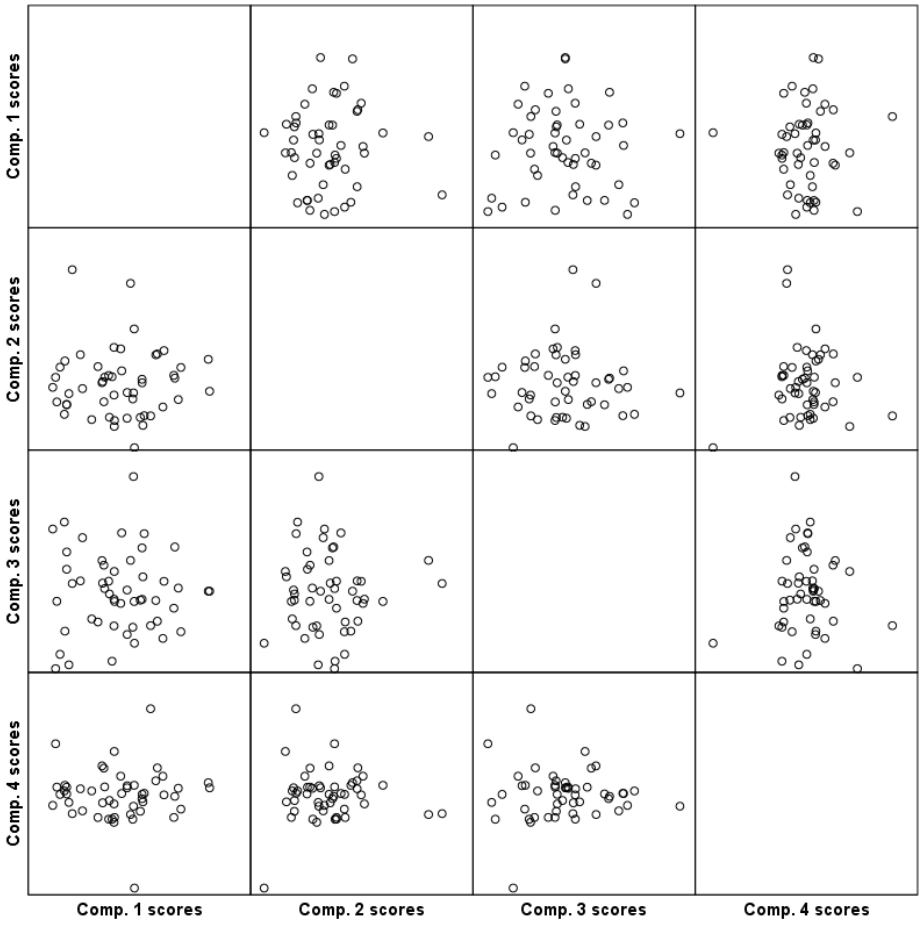


Figure 9-4: Scatterplots of components scores in LOCNEC+

9.8 CLUSTER ANALYSIS

9.8.1 LINDSEI-FR+

9.8.1.1 The make-up of the clusters

Table 9-10 shows the ID of the LINDSEI-FR+ learners included in the two clusters. Table 9-11 shows the ID of the learners in the 6 clusters.

Cluster 1			Cluster 2		
FR001	FR019	FR039	FR002	FR020	FR037
FR003	FR027	FR040	FR004	FR021	FR038
FR005	FR028	FR042	FR006	FR022	FR041
FR008	FR029	FR043	FR007	FR023	FR044
FR010	FR031	FR046	FR009	FR024	FR045
FR013	FR032		FR011	FR025	FR047
FR014	FR034		FR012	FR026	FR048
FR015	FR035		FR016	FR030	FR049
FR018	FR036		FR017	FR033	FR050
n = 23			*n* = 27		

Table 9-10: The make-up of the 2 main clusters in LINDSEI-FR+

Cluster A	Cluster B	Cluster C	Cluster D		Cluster E	Cluster F
FR001	FR005	FR010	FR002	FR033	FR006	FR009
FR003	FR008	FR013	FR004	FR037	FR022	FR012
FR019	FR018	FR014	FR007	FR044	FR025	FR016
FR027	FR028	FR015	FR011	FR047	FR030	FR020
FR029		FR035	FR017	FR048	FR038	FR026
FR031		FR039	FR021	FR049	FR041	FR045
FR032		FR042	FR023	FR050		
FR034		FR043	FR024			
FR036		FR046				
FR040						
n = 10	*n* = 4	*n* = 9	*n* = 15		*n* = 6	*n* = 6

Table 9-11: The make-up of the 6 clusters in LINDSEI-FR+

9.8.1.2 Cluster profiles per (dis)fluency component (6-cluster solution)

Figure 9-5 to 9-10 show the cluster profiles per (dis)fluency component for the 6-cluster solution in LINDSEI-FR+.

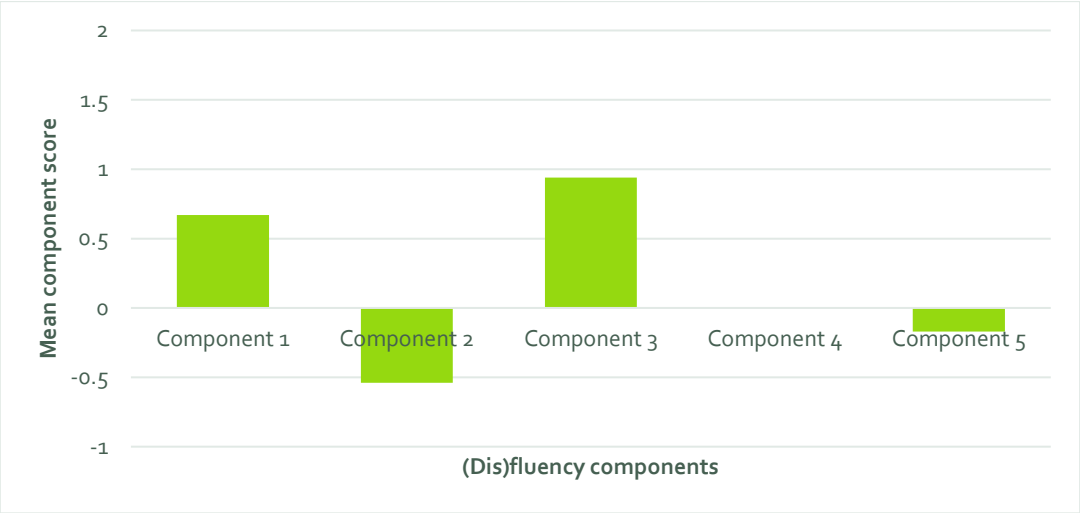


Figure 9-5: Cluster A profile per (dis)fluency components in LINDSEI-FR+

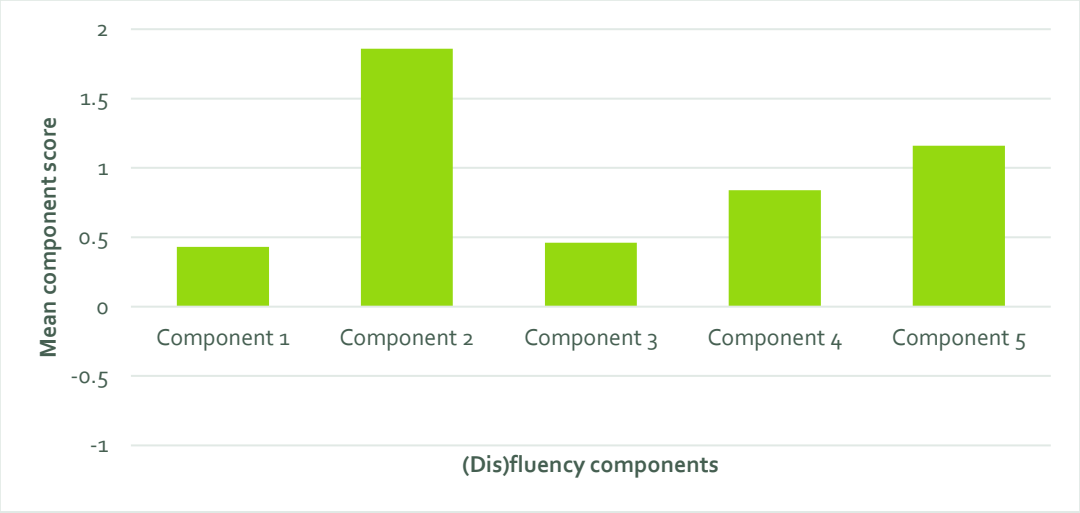


Figure 9-6: Cluster B profile per (dis)fluency components in LINDSEI-FR+

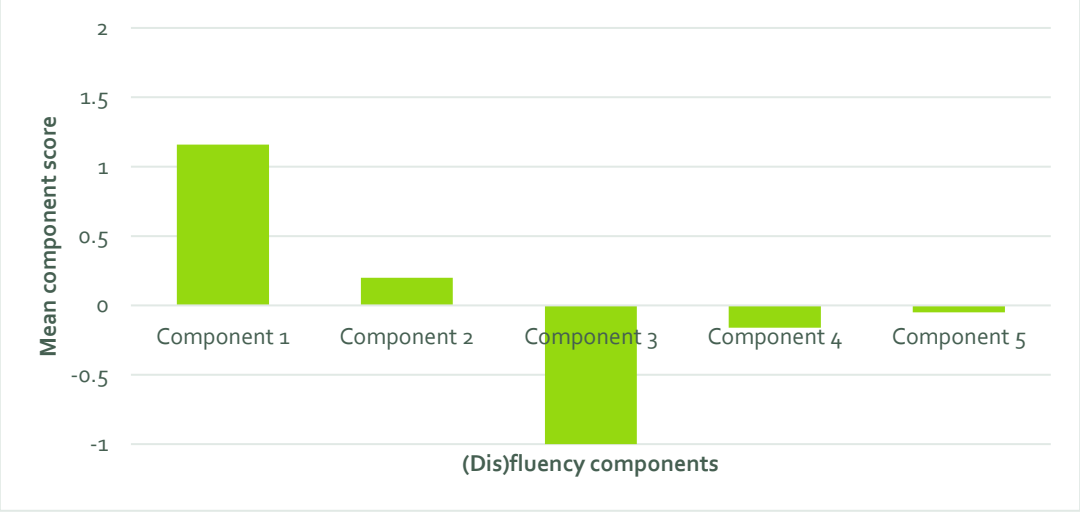


Figure 9-7: Cluster C profile per (dis)fluency components in LINDSEI-FR+

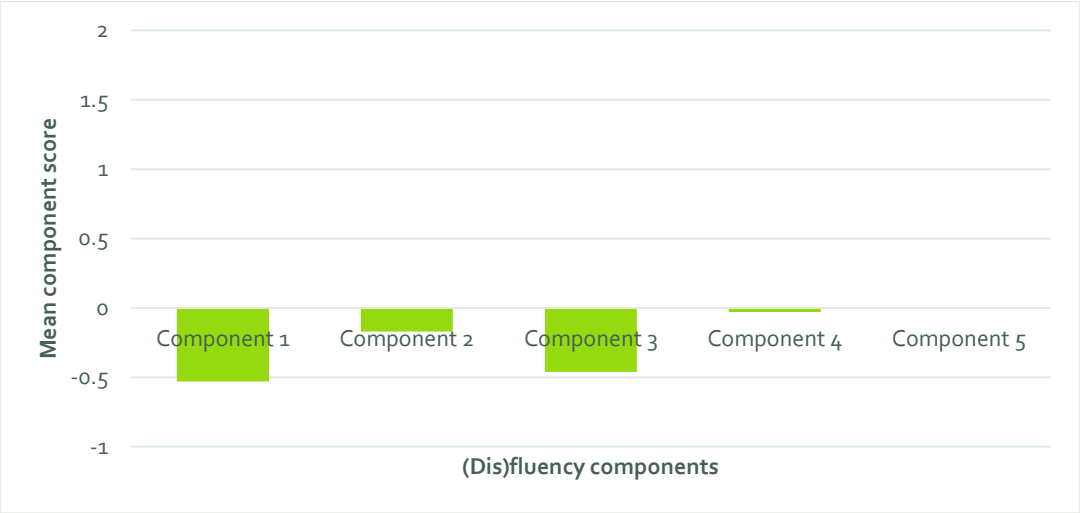


Figure 9-8: Cluster D profile per (dis)fluency components in LINDSEI-FR+

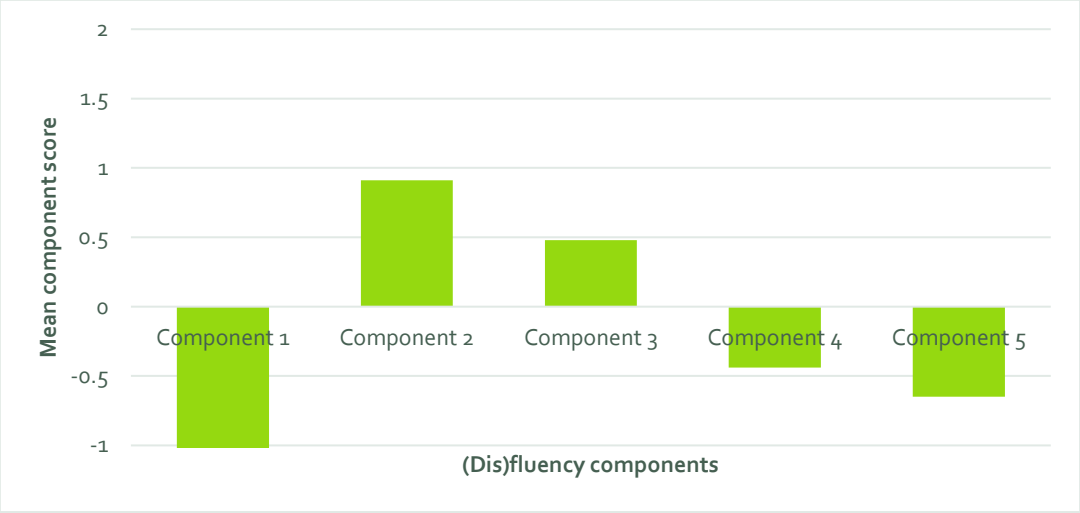


Figure 9-9: Cluster E profile per (dis)fluency components in LINDSEI-FR+

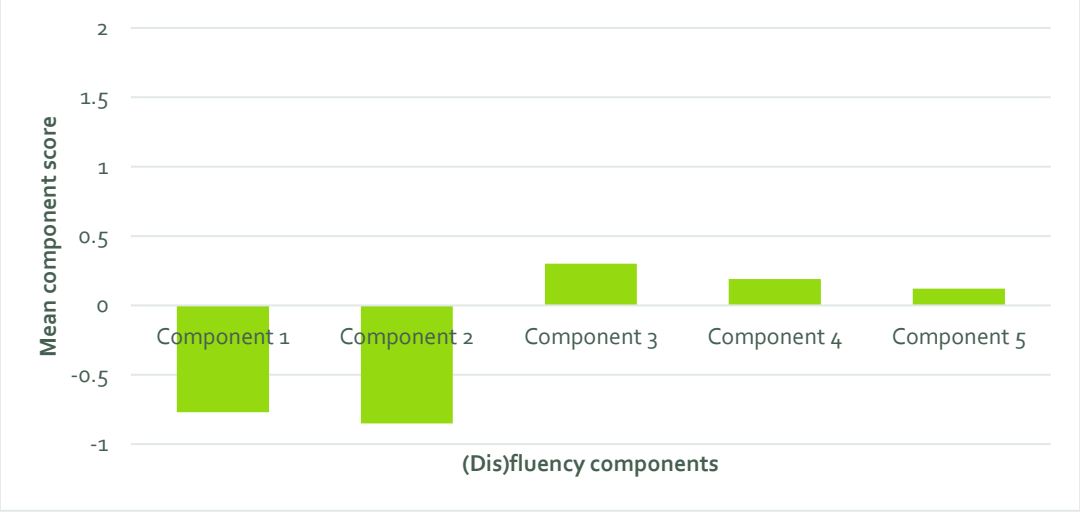


Figure 9-10: Cluster F profile per (dis)fluency components in LINDSEI-FR+

9.8.1.3 ANOVA results (14 (dis)fluency variables)

Table 9-12 shows the results for Levene's test of homogeneity of variances. Table 9-13 shows the results of the ANOVA test and Table 9-14 the results of Welch's test (for truncations and foreign words).

Note: significant results are in bold font.

Test d'homogénéité des variances				
	Statistique de Levene	ddl1	ddl2	Sig.
Score Z(C_phw)	2.116	5	44	.081
Score Z(DM_phw)	1.421	5	44	.235
Score Z(FP_phw)	.351	5	44	.879
Score Z(FS_phw)	.279	5	44	.922
Score Z(L_phw)	1.570	5	44	.188
Score Z(Mean_length_of_runs)	1.956	5	44	.104
Score Z(Mean_UP_length_sec)	1.651	5	44	.167
Score Z(Phonation_time_ratio)	1.458	5	44	.223
Score Z(Rep_phw)	1.427	5	44	.234
Score Z(RS_phw)	1.737	5	44	.146
Score Z(Speech_rate_wpm)	2.356	5	44	.056
Score Z(T_phw)	3.203	5	44	**.015**
Score Z(UP_phw)	1.236	5	44	.309
Score Z(W_phw)	4.816	5	44	**.001**

Table 9-12: Levene's test of homogeneity of variances
Note: significant results indicate that the variances are significantly different between the groups

ANOVA						
		Somme des carrés	ddl	Carré moyen	F	Sig.
Score Z(UP_phw)	Inter-groupes	26.231	5	5.246	10.138	**.000**
	Intragroupes	22.769	44	.517		
	Total	49.000	49			
Score Z(Mean_UP_length_sec)	Inter-groupes	25.564	5	5.113	9.599	**.000**
	Intragroupes	23.436	44	.533		
	Total	49.000	49			
Score Z(Mean_length_of_runs)	Inter-groupes	26.805	5	5.361	10.628	**.000**
	Intragroupes	22.195	44	.504		
	Total	49.000	49			
Score Z(Phonation_time_ratio)	Inter-groupes	29.721	5	5.944	13.566	**.000**
	Intragroupes	19.279	44	.438		
	Total	49.000	49			
Score Z(Speech_rate_wpm)	Inter-groupes	18.012	5	3.602	5.115	**.001**
	Intragroupes	30.988	44	.704		
	Total	49.000	49			
Score Z(C_phw)	Inter-groupes	3.406	5	.681	.657	.658

	Intragroupes	45.594	44	1.036		
	Total	49.000	49			
Score Z(DM_phw)	Inter-groupes	17.591	5	3.518	4.928	**.001**
	Intragroupes	31.409	44	.714		
	Total	49.000	49			
Score Z(FP_phw)	Inter-groupes	16.964	5	3.393	4.660	**.002**
	Intragroupes	32.036	44	.728		
	Total	49.000	49			
Score Z(FS_phw)	Inter-groupes	12.649	5	2.530	3.062	**.019**
	Intragroupes	36.351	44	.826		
	Total	49.000	49			
Score Z(L_phw)	Inter-groupes	29.136	5	5.827	12.908	**.000**
	Intragroupes	19.864	44	.451		
	Total	49.000	49			
Score Z(Rep_phw)	Inter-groupes	17.929	5	3.586	5.078	**.001**
	Intragroupes	31.071	44	.706		
	Total	49.000	49			
Score Z(RS_phw)	Inter-groupes	20.862	5	4.172	6.525	**.000**
	Intragroupes	28.138	44	.639		
	Total	49.000	49			
Score Z(T_phw)	*Inter-groupes*	*22.562*	*5*	*4.512*	*7.510*	*.000*
	Intragroupes	*26.438*	*44*	*.601*		
	Total	*49.000*	*49*			
Score Z(W_phw)	*Inter-groupes*	*7.185*	*5*	*1.437*	*1.512*	*.206*
	Intragroupes	*41.815*	*44*	*.950*		
	Total	*49.000*	*49*			

Table 9-13: ANOVA results

Tests robustes d'égalité des moyennes					
		Statistiques[a]	ddl1	ddl2	Sig.
Score Z(UP_phw)	*Welch*	*13.902*	*5*	*14.852*	*.000*
Score Z(Mean_UP_length_sec)	*Welch*	*17.480*	*5*	*15.739*	*.000*
Score Z(Mean_length_of_runs)	*Welch*	*9.143*	*5*	*14.637*	*.000*
Score Z(Phonation_time_ratio)	*Welch*	*12.816*	*5*	*15.403*	*.000*
Score Z(Speech_rate_wpm)	*Welch*	*7.165*	*5*	*15.480*	*.001*
Score Z(C_phw)	*Welch*	*.979*	*5*	*15.839*	*.461*
Score Z(DM_phw)	*Welch*	*4.637*	*5*	*14.655*	*.010*
Score Z(FP_phw)	*Welch*	*4.455*	*5*	*15.606*	*.010*
Score Z(FS_phw)	*Welch*	*1.707*	*5*	*14.670*	*.195*
Score Z(L_phw)	*Welch*	*13.148*	*5*	*15.361*	*.000*
Score Z(Rep_phw)	*Welch*	*4.713*	*5*	*14.414*	*.009*
Score Z(RS_phw)	*Welch*	*6.073*	*5*	*15.482*	*.003*
Score Z(T_phw)	Welch	21.170	5	15.272	**.000**
Score Z(W_phw)	Welch	3.654	5	14.803	**.024**
a. F distribué asymptotiquement					

Table 9-14: Welch's F (for T and W phw)

Table 9-15 shows the results of pariwise comparisons using Gabriel and Hochberg's procedure.

Note: 1 = cluster A; 2 = cluster D; 3 = cluster B; 4 = cluster E; 5 = cluster F; 6 = cluster C.

Comparaisons multiples :

Variable dépendante				Différence moyenne (I-J)	Erreur standard	Sig.	Intervalle de confiance à 95 %	
							Borne inférieure	Borne supérieure
Score Z(UP_phw)	Gabriel	1	2	-1.263*	0.294	**0.001**	-2.164	-0.362
			3	-0.638	0.426	0.852	-1.918	0.643
			4	-1.293*	0.371	**0.015**	-2.430	-0.157
			5	-1.380*	0.371	**0.008**	-2.517	-0.244
			6	0.441	0.331	0.941	-0.578	1.460
		2	1	1.263*	0.294	**0.001**	0.362	2.164
			3	0.626	0.405	0.800	-0.564	1.815
			4	-0.030	0.347	1.000	-1.076	1.015
			5	-0.117	0.347	1.000	-1.163	0.928
			6	1.703*	0.303	**0.000**	0.776	2.632
		3	1	0.638	0.426	0.852	-0.643	1.918
			2	-0.626	0.405	0.800	-1.815	0.564
			4	-0.656	0.464	0.910	-2.080	0.769
			5	-0.743	0.464	0.809	-2.167	0.682
			6	1.078	0.432	0.187	-0.229	2.385
		4	1	1.293*	0.371	**0.015**	0.157	2.430
			2	0.030	0.347	1.000	-1.015	1.076
			3	0.656	0.464	0.910	-0.769	2.080
			5	-0.087	0.415	1.000	-1.368	1.194
			6	1.734*	0.379	**0.001**	0.571	2.897
		5	1	1.380*	0.371	**0.008**	0.244	2.517
			2	0.117	0.347	1.000	-0.928	1.163
			3	0.743	0.464	0.809	-0.682	2.167
			4	0.087	0.415	1.000	-1.194	1.368
			6	1.820*	0.379	**0.000**	0.658	2.984
		6	1	-0.441	0.331	0.941	-1.460	0.578
			2	-1.703*	0.303	**0.000**	-2.632	-0.776
			3	-1.078	0.432	0.187	-2.385	0.229
			4	-1.734*	0.379	**0.001**	-2.897	-0.571
			5	-1.820*	0.379	**0.000**	-2.984	-0.658
	Hochberg	1	2	-1.263*	0.294	**0.001**	-2.169	-0.358
			3	-0.638	0.426	0.873	-1.950	0.675
			4	-1.293*	0.371	**0.017**	-2.439	-0.148
			5	-1.380*	0.371	**0.008**	-2.526	-0.235

			6	0.441	0.331	0.942	-0.579	1.460
		2	1	1.263*	0.294	**0.001**	0.358	2.169
			3	0.626	0.405	0.847	-0.623	1.874
			4	-0.030	0.347	1.000	-1.102	1.041
			5	-0.117	0.347	1.000	-1.189	0.954
			6	1.703*	0.303	**0.000**	0.768	2.639
		3	1	0.638	0.426	0.873	-0.675	1.950
			2	-0.626	0.405	0.847	-1.874	0.623
			4	-0.656	0.464	0.913	-2.088	0.776
			5	-0.743	0.464	0.815	-2.175	0.689
			6	1.078	0.432	0.209	-0.255	2.411
		4	1	1.293*	0.371	**0.017**	0.148	2.439
			2	0.030	0.347	1.000	-1.041	1.102
			3	0.656	0.464	0.913	-0.776	2.088
			5	-0.087	0.415	1.000	-1.368	1.194
			6	1.734*	0.379	**0.001**	0.565	2.903
		5	1	1.380*	0.371	**0.008**	0.235	2.526
			2	0.117	0.347	1.000	-0.954	1.189
			3	0.743	0.464	0.815	-0.689	2.175
			4	0.087	0.415	1.000	-1.194	1.368
			6	1.820*	0.379	**0.000**	0.652	2.990
		6	1	-0.441	0.331	0.942	-1.460	0.579
			2	-1.703*	0.303	**0.000**	-2.639	-0.768
			3	-1.078	0.432	0.209	-2.411	0.255
			4	-1.734*	0.379	**0.001**	-2.903	-0.565
			5	-1.820*	0.379	**0.000**	-2.990	-0.652
Score Z(Mean_UP_length_sec)	Gabriel	1	2	-0.549	0.298	0.632	-1.464	0.365
			3	0.261	0.432	1.000	-1.038	1.560
			4	-2.298*	0.377	**0.000**	-3.452	-1.146
			5	-0.437	0.377	0.979	-1.590	0.716
			6	-0.124	0.335	1.000	-1.158	0.909
		2	1	0.549	0.298	0.632	-0.365	1.464
			3	0.811	0.411	0.463	-0.396	2.017
			4	-1.749*	0.353	**0.000**	-2.810	-0.689
			5	0.113	0.353	1.000	-0.948	1.173
			6	0.425	0.308	0.920	-0.516	1.367
		3	1	-0.261	0.432	1.000	-1.560	1.038
			2	-0.811	0.411	0.463	-2.017	0.396
			4	-2.560*	0.471	**0.000**	-4.006	-1.115
			5	-0.698	0.471	0.877	-2.144	0.747
			6	-0.386	0.439	0.998	-1.712	0.941
		4	1	2.298*	0.377	**0.000**	1.146	3.452
			2	1.749*	0.353	**0.000**	0.689	2.810
			3	2.560*	0.471	**0.000**	1.115	4.006
			5	1.861*	0.421	**0.001**	0.563	3.161

			6	2.174*	0.385	**0.000**	0.994	3.355
		5	1	0.437	0.377	0.979	-0.716	1.590
			2	-0.113	0.353	1.000	-1.173	0.948
			3	0.698	0.471	0.877	-0.747	2.144
			4	-1.861*	0.421	**0.001**	-3.161	-0.563
			6	0.313	0.385	0.999	-0.868	1.493
		6	1	0.124	0.335	1.000	-0.909	1.158
			2	-0.425	0.308	0.920	-1.367	0.516
			3	0.386	0.439	0.998	-0.941	1.712
			4	-2.174*	0.385	**0.000**	-3.355	-0.994
			5	-0.313	0.385	0.999	-1.493	0.868
	Hochberg	1	2	-0.549	0.298	0.639	-1.468	0.369
			3	0.261	0.432	1.000	-1.070	1.593
			4	-2.298*	0.377	**0.000**	-3.461	-1.137
			5	-0.437	0.377	0.981	-1.599	0.725
			6	-0.124	0.335	1.000	-1.158	0.910
		2	1	0.549	0.298	0.639	-0.369	1.468
			3	0.811	0.411	0.538	-0.456	2.077
			4	-1.749*	0.353	**0.000**	-2.837	-0.662
			5	0.113	0.353	1.000	-0.975	1.200
			6	0.425	0.308	0.925	-0.524	1.374
		3	1	-0.261	0.432	1.000	-1.593	1.070
			2	-0.811	0.411	0.538	-2.077	0.456
			4	-2.560*	0.471	**0.000**	-4.013	-1.108
			5	-0.698	0.471	0.881	-2.151	0.754
			6	-0.386	0.439	0.999	-1.738	0.967
		4	1	2.298*	0.377	**0.000**	1.137	3.461
			2	1.749*	0.353	**0.000**	0.662	2.837
			3	2.560*	0.471	**0.000**	1.108	4.013
			5	1.861*	0.421	**0.001**	0.563	3.161
			6	2.174*	0.385	**0.000**	0.988	3.361
		5	1	0.437	0.377	0.981	-0.725	1.599
			2	-0.113	0.353	1.000	-1.200	0.975
			3	0.698	0.471	0.881	-0.754	2.151
			4	-1.861*	0.421	**0.001**	-3.161	-0.563
			6	0.313	0.385	0.999	-0.874	1.499
		6	1	0.124	0.335	1.000	-0.910	1.158
			2	-0.425	0.308	0.925	-1.374	0.524
			3	0.386	0.439	0.999	-0.967	1.738
			4	-2.174*	0.385	**0.000**	-3.361	-0.988
			5	-0.313	0.385	0.999	-1.499	0.874
Score Z(Mean_length_ of_runs)	Gabriel	1	2	.951*	0.290	**0.028**	0.062	1.842
			3	-0.021	0.420	1.000	-1.285	1.243
			4	1.072	0.367	0.071	-0.050	2.194
			5	1.311*	0.367	**0.012**	0.190	2.434

		6	-0.768	0.326	0.279	-1.774	0.238
	2	1	-.951*	0.290	**0.028**	-1.842	-0.062
		3	-0.973	0.400	0.183	-2.147	0.201
		4	0.120	0.343	1.000	-0.912	1.152
		5	0.360	0.343	0.990	-0.672	1.392
		6	-1.719*	0.299	**0.000**	-2.636	-0.804
	3	1	0.021	0.420	1.000	-1.243	1.285
		2	0.973	0.400	0.183	-0.201	2.147
		4	1.093	0.458	0.256	-0.313	2.500
		5	1.333	0.458	0.076	-0.074	2.740
		6	-0.747	0.427	0.685	-2.037	0.544
	4	1	-1.072	0.367	0.071	-2.194	0.050
		2	-0.120	0.343	1.000	-1.152	0.912
		3	-1.093	0.458	0.256	-2.500	0.313
		5	0.240	0.410	1.000	-1.025	1.504
		6	-1.84*	0.374	**0.000**	-2.989	-0.692
	5	1	-1.311*	0.367	**0.012**	-2.434	-0.190
		2	-0.360	0.343	0.990	-1.392	0.672
		3	-1.333	0.458	0.076	-2.740	0.074
		4	-0.240	0.410	1.000	-1.504	1.025
		6	-2.079*	0.374	**0.000**	-3.228	-0.931
	6	1	0.768	0.326	0.279	-0.238	1.774
		2	1.719*	0.299	**0.000**	0.804	2.636
		3	0.747	0.427	0.685	-0.544	2.037
		4	1.840*	0.374	**0.000**	0.692	2.989
		5	2.079*	0.374	**0.000**	0.931	3.228
Hochberg	1	2	.951*	0.290	**0.029**	0.058	1.846
		3	-0.021	0.420	1.000	-1.317	1.275
		4	1.072	0.367	0.076	-0.059	2.203
		5	1.311*	0.367	**0.013**	0.181	2.443
		6	-0.768	0.326	0.280	-1.774	0.238
	2	1	-.951*	0.290	**0.029**	-1.846	-0.058
		3	-0.973	0.400	0.237	-2.206	0.259
		4	0.120	0.343	1.000	-0.938	1.178
		5	0.360	0.343	0.992	-0.698	1.418
		6	-1.719*	0.299	**0.000**	-2.643	-0.796
	3	1	0.021	0.420	1.000	-1.275	1.317
		2	0.973	0.400	0.237	-0.259	2.206
		4	1.093	0.458	0.262	-0.320	2.507
		5	1.333	0.458	0.079	-0.081	2.747
		6	-0.747	0.427	0.711	-2.063	0.569
	4	1	-1.072	0.367	0.076	-2.203	0.059
		2	-0.120	0.343	1.000	-1.178	0.938
		3	-1.093	0.458	0.262	-2.507	0.320
		5	0.240	0.410	1.000	-1.025	1.504

			6	-1.840*	0.374	**0.000**	-2.994	-0.686
		5	1	-1.311*	0.367	**0.013**	-2.443	-0.181
			2	-0.360	0.343	0.992	-1.418	0.698
			3	-1.333	0.458	0.079	-2.747	0.081
			4	-0.240	0.410	1.000	-1.504	1.025
			6	-2.079*	0.374	**0.000**	-3.234	-0.925
		6	1	0.768	0.326	0.280	-0.238	1.774
			2	1.719*	0.299	**0.000**	0.796	2.643
			3	0.747	0.427	0.711	-0.569	2.063
			4	1.840*	0.374	**0.000**	0.686	2.994
			5	2.079*	0.374	**0.000**	0.925	3.234
Score Z(Phonation_time_ratio)	Gabriel	1	2	.955*	0.270	**0.013**	0.126	1.785
			3	0.218	0.392	1.000	-0.960	1.397
			4	2.148*	0.342	**0.000**	1.103	3.194
			5	0.895	0.342	0.151	-0.151	1.941
			6	-0.385	0.304	0.960	-1.323	0.553
		2	1	-.955*	0.270	**0.013**	-1.785	-0.126
			3	-0.737	0.372	0.460	-1.832	0.357
			4	1.192*	0.320	**0.006**	0.231	2.155
			5	-0.061	0.320	1.000	-1.023	0.901
			6	-1.340*	0.279	**0.000**	-2.194	-0.487
		3	1	-0.218	0.392	1.000	-1.397	0.960
			2	0.737	0.372	0.460	-0.357	1.832
			4	1.930*	0.427	**0.001**	0.619	3.241
			5	0.677	0.427	0.819	-0.634	1.988
			6	-0.603	0.398	0.846	-1.806	0.600
		4	1	-2.148*	0.342	**0.000**	-3.194	-1.103
			2	-1.192*	0.320	**0.006**	-2.155	-0.231
			3	-1.930*	0.427	**0.001**	-3.241	-0.619
			5	-1.253*	0.382	**0.029**	-2.432	-0.075
			6	-2.533*	0.349	**0.000**	-3.604	-1.463
		5	1	-0.895	0.342	0.151	-1.941	0.151
			2	0.061	0.320	1.000	-0.901	1.023
			3	-0.677	0.427	0.819	-1.988	0.634
			4	1.253*	0.382	**0.029**	0.075	2.432
			6	-1.279*	0.349	**0.009**	-2.350	-0.210
		6	1	0.385	0.304	0.960	-0.553	1.323
			2	1.340*	0.279	**0.000**	0.487	2.194
			3	0.603	0.398	0.846	-0.600	1.806
			4	2.533*	0.349	**0.000**	1.463	3.604
			5	1.279*	0.349	**0.009**	0.210	2.350
	Hochberg	1	2	.955*	0.270	**0.014**	0.122	1.789
			3	0.218	0.392	1.000	-0.989	1.426
			4	2.148*	0.342	**0.000**	1.094	3.203
			5	0.895	0.342	0.159	-0.159	1.949

			6	-0.385	0.304	0.960	-1.323	0.553
		2	1	-.955*	0.270	**0.014**	-1.789	-0.122
			3	-0.737	0.372	0.534	-1.886	0.411
			4	1.192*	0.320	**0.008**	0.207	2.179
			5	-0.061	0.320	1.000	-1.047	0.925
			6	-1.340*	0.279	**0.000**	-2.201	-0.480
		3	1	-0.218	0.392	1.000	-1.426	0.989
			2	0.737	0.372	0.534	-0.411	1.886
			4	1.930*	0.427	**0.001**	0.613	3.248
			5	0.677	0.427	0.824	-0.641	1.994
			6	-0.603	0.398	0.863	-1.830	0.623
		4	1	-2.148*	0.342	**0.000**	-3.203	-1.094
			2	-1.192*	0.320	**0.008**	-2.179	-0.207
			3	-1.930*	0.427	**0.001**	-3.248	-0.613
			5	-1.253*	0.382	**0.029**	-2.432	-0.075
			6	-2.533*	0.349	**0.000**	-3.609	-1.458
		5	1	-0.895	0.342	0.159	-1.949	0.159
			2	0.061	0.320	1.000	-0.925	1.047
			3	-0.677	0.427	0.824	-1.994	0.641
			4	1.253*	0.382	**0.029**	0.075	2.432
			6	-1.279*	0.349	**0.010**	-2.356	-0.204
		6	1	0.385	0.304	0.960	-0.553	1.323
			2	1.340*	0.279	**0.000**	0.480	2.201
			3	0.603	0.398	0.863	-0.623	1.830
			4	2.533*	0.349	**0.000**	1.458	3.609
			5	1.279*	0.349	**0.010**	0.204	2.356
Score Z(Speech_rate_wpm)	Gabriel	1	2	1.226*	0.343	**0.012**	0.175	2.278
			3	0.196	0.496	1.000	-1.297	1.690
			4	1.440*	0.433	**0.024**	0.115	2.766
			5	1.519*	0.433	**0.014**	0.194	2.845
			6	0.345	0.386	0.998	-0.844	1.534
		2	1	-1.226*	0.343	**0.012**	-2.278	-0.175
			3	-1.030	0.472	0.317	-2.418	0.357
			4	0.214	0.405	1.000	-1.006	1.434
			5	0.293	0.405	1.000	-0.927	1.512
			6	-0.881	0.354	0.202	-1.964	0.201
		3	1	-0.196	0.496	1.000	-1.690	1.297
			2	1.030	0.472	0.317	-0.357	2.418
			4	1.244	0.542	0.305	-0.418	2.906
			5	1.323	0.542	0.227	-0.339	2.985
			6	0.149	0.504	1.000	-1.376	1.674
		4	1	-1.440*	0.433	**0.024**	-2.766	-0.115
			2	-0.214	0.405	1.000	-1.434	1.006
			3	-1.244	0.542	0.305	-2.906	0.418
			5	0.079	0.485	1.000	-1.415	1.573

			6	-1.095	0.442	0.211	-2.452	0.262
		5	1	-1.519*	0.433	**0.014**	-2.845	-0.194
			2	-0.293	0.405	1.000	-1.512	0.927
			3	-1.323	0.542	0.227	-2.985	0.339
			4	-0.079	0.485	1.000	-1.573	1.415
			6	-1.174	0.442	0.141	-2.531	0.183
		6	1	-0.345	0.386	0.998	-1.534	0.844
			2	0.881	0.354	0.202	-0.201	1.964
			3	-0.149	0.504	1.000	-1.674	1.376
			4	1.095	0.442	0.211	-0.262	2.452
			5	1.174	0.442	0.141	-0.183	2.531
	Hochberg	1	2	1.226*	0.343	**0.013**	0.170	2.283
			3	0.196	0.496	1.000	-1.335	1.727
			4	1.440*	0.433	**0.026**	0.104	2.777
			5	1.519*	0.433	**0.016**	0.183	2.856
			6	0.345	0.386	0.998	-0.844	1.534
		2	1	-1.226*	0.343	**0.013**	-2.283	-0.170
			3	-1.030	0.472	0.386	-2.486	0.426
			4	0.214	0.405	1.000	-1.036	1.464
			5	0.293	0.405	1.000	-0.957	1.543
			6	-0.881	0.354	0.210	-1.972	0.210
		3	1	-0.196	0.496	1.000	-1.727	1.335
			2	1.030	0.472	0.386	-0.426	2.486
			4	1.244	0.542	0.312	-0.427	2.915
			5	1.323	0.542	0.233	-0.348	2.993
			6	0.149	0.504	1.000	-1.406	1.704
		4	1	-1.440*	0.433	**0.026**	-2.777	-0.104
			2	-0.214	0.405	1.000	-1.464	1.036
			3	-1.244	0.542	0.312	-2.915	0.427
			5	0.079	0.485	1.000	-1.415	1.573
			6	-1.095	0.442	0.217	-2.459	0.269
		5	1	-1.519*	0.433	**0.016**	-2.856	-0.183
			2	-0.293	0.405	1.000	-1.543	0.957
			3	-1.323	0.542	0.233	-2.993	0.348
			4	-0.079	0.485	1.000	-1.573	1.415
			6	-1.174	0.442	0.146	-2.538	0.190
		6	1	-0.345	0.386	0.998	-1.534	0.844
			2	0.881	0.354	0.210	-0.210	1.972
			3	-0.149	0.504	1.000	-1.704	1.406
			4	1.095	0.442	0.217	-0.269	2.459
			5	1.174	0.442	0.146	-0.190	2.538
Score Z(C_phw)	Gabriel	1	2	-0.150	0.416	1.000	-1.425	1.125
			3	-0.781	0.602	0.943	-2.593	1.031
			4	0.202	0.526	1.000	-1.406	1.810
			5	-0.094	0.526	1.000	-1.702	1.514

		6	0.229	0.468	1.000	-1.212	1.671
	2	1	0.150	0.416	1.000	-1.125	1.425
		3	-0.631	0.573	0.981	-2.314	1.052
		4	0.352	0.492	1.000	-1.127	1.831
		5	0.056	0.492	1.000	-1.423	1.535
		6	0.379	0.429	0.998	-0.934	1.692
	3	1	0.781	0.602	0.943	-1.031	2.593
		2	0.631	0.573	0.981	-1.052	2.314
		4	0.983	0.657	0.870	-1.033	2.999
		5	0.687	0.657	0.992	-1.329	2.703
		6	1.010	0.612	0.758	-0.840	2.860
	4	1	-0.202	0.526	1.000	-1.810	1.406
		2	-0.352	0.492	1.000	-1.831	1.127
		3	-0.983	0.657	0.870	-2.999	1.033
		5	-0.296	0.588	1.000	-2.108	1.516
		6	0.027	0.537	1.000	-1.619	1.673
	5	1	0.094	0.526	1.000	-1.514	1.702
		2	-0.056	0.492	1.000	-1.535	1.423
		3	-0.687	0.657	0.992	-2.703	1.329
		4	0.296	0.588	1.000	-1.516	2.108
		6	0.323	0.537	1.000	-1.323	1.969
	6	1	-0.229	0.468	1.000	-1.671	1.212
		2	-0.379	0.429	0.998	-1.692	0.934
		3	-1.010	0.612	0.758	-2.860	0.840
		4	-0.027	0.537	1.000	-1.673	1.619
		5	-0.323	0.537	1.000	-1.969	1.323
Hochberg	1	2	-0.150	0.416	1.000	-1.432	1.132
		3	-0.781	0.602	0.952	-2.638	1.076
		4	0.202	0.526	1.000	-1.419	1.823
		5	-0.094	0.526	1.000	-1.715	1.527
		6	0.229	0.468	1.000	-1.213	1.672
	2	1	0.150	0.416	1.000	-1.132	1.432
		3	-0.631	0.573	0.988	-2.397	1.136
		4	0.352	0.492	1.000	-1.164	1.868
		5	0.056	0.492	1.000	-1.460	1.572
		6	0.379	0.429	0.999	-0.944	1.703
	3	1	0.781	0.602	0.952	-1.076	2.638
		2	0.631	0.573	0.988	-1.136	2.397
		4	0.983	0.657	0.874	-1.044	3.009
		5	0.687	0.657	0.992	-1.339	2.713
		6	1.010	0.612	0.781	-0.876	2.897
	4	1	-0.202	0.526	1.000	-1.823	1.419
		2	-0.352	0.492	1.000	-1.868	1.164
		3	-0.983	0.657	0.874	-3.009	1.044
		5	-0.296	0.588	1.000	-2.108	1.516

			6	0.027	0.537	1.000	-1.627	1.682
		5	1	0.094	0.526	1.000	-1.527	1.715
			2	-0.056	0.492	1.000	-1.572	1.460
			3	-0.687	0.657	0.992	-2.713	1.339
			4	0.296	0.588	1.000	-1.516	2.108
			6	0.323	0.537	1.000	-1.331	1.978
		6	1	-0.229	0.468	1.000	-1.672	1.213
			2	-0.379	0.429	0.999	-1.703	0.944
			3	-1.010	0.612	0.781	-2.897	0.876
			4	-0.027	0.537	1.000	-1.682	1.627
			5	-0.323	0.537	1.000	-1.978	1.331
Score Z(DM_phw)	Gabriel	1	2	1.376*	0.345	**0.003**	0.319	2.435
			3	0.503	0.500	0.993	-1.000	2.007
			4	0.559	0.436	0.954	-0.776	1.894
			5	1.120	0.436	0.170	-0.214	2.455
			6	1.611*	0.388	**0.002**	0.415	2.809
		2	1	-1.376*	0.345	**0.003**	-2.435	-0.319
			3	-0.874	0.475	0.573	-2.271	0.523
			4	-0.818	0.408	0.477	-2.046	0.410
			5	-0.256	0.408	1.000	-1.484	0.971
			6	0.235	0.356	1.000	-0.855	1.325
		3	1	-0.503	0.500	0.993	-2.007	1.000
			2	0.874	0.475	0.573	-0.523	2.271
			4	0.056	0.545	1.000	-1.617	1.729
			5	0.617	0.545	0.984	-1.056	2.290
			6	1.109	0.508	0.356	-0.427	2.644
		4	1	-0.559	0.436	0.954	-1.894	0.776
			2	0.818	0.408	0.477	-0.410	2.046
			3	-0.056	0.545	1.000	-1.729	1.617
			5	0.561	0.488	0.982	-0.943	2.066
			6	1.053	0.445	0.267	-0.314	2.419
		5	1	-1.120	0.436	0.170	-2.455	0.214
			2	0.256	0.408	1.000	-0.971	1.484
			3	-0.617	0.545	0.984	-2.290	1.056
			4	-0.561	0.488	0.982	-2.066	0.943
			6	0.491	0.445	0.987	-0.875	1.858
		6	1	-1.611*	0.388	**0.002**	-2.809	-0.415
			2	-0.235	0.356	1.000	-1.325	0.855
			3	-1.109	0.508	0.356	-2.644	0.427
			4	-1.053	0.445	0.267	-2.419	0.314
			5	-0.491	0.445	0.987	-1.858	0.875
	Hochberg	1	2	1.376*	0.345	**0.004**	0.313	2.441
			3	0.503	0.500	0.995	-1.038	2.045
			4	0.559	0.436	0.956	-0.786	1.905
			5	1.120	0.436	0.178	-0.225	2.466

			6	1.611*	0.388	**0.002**	0.415	2.809
		2	1	-1.376*	0.345	**0.004**	-2.441	-0.313
			3	-0.874	0.475	0.644	-2.340	0.593
			4	-0.818	0.408	0.515	-2.076	0.441
			5	-0.256	0.408	1.000	-1.515	1.002
			6	0.235	0.356	1.000	-0.864	1.334
		3	1	-0.503	0.500	0.995	-2.045	1.038
			2	0.874	0.475	0.644	-0.593	2.340
			4	0.056	0.545	1.000	-1.626	1.738
			5	0.617	0.545	0.984	-1.065	2.299
			6	1.109	0.508	0.385	-0.457	2.674
		4	1	-0.559	0.436	0.956	-1.905	0.786
			2	0.818	0.408	0.515	-0.441	2.076
			3	-0.056	0.545	1.000	-1.738	1.626
			5	0.561	0.488	0.982	-0.943	2.066
			6	1.053	0.445	0.274	-0.320	2.426
		5	1	-1.120	0.436	0.178	-2.466	0.225
			2	0.256	0.408	1.000	-1.002	1.515
			3	-0.617	0.545	0.984	-2.299	1.065
			4	-0.561	0.488	0.982	-2.066	0.943
			6	0.491	0.445	0.987	-0.882	1.865
		6	1	-1.611*	0.388	**0.002**	-2.809	-0.415
			2	-0.235	0.356	1.000	-1.334	0.864
			3	-1.109	0.508	0.385	-2.674	0.457
			4	-1.053	0.445	0.274	-2.426	0.320
			5	-0.491	0.445	0.987	-1.865	0.882
Score Z(FP_phw)	Gabriel	1	2	-0.961	0.348	0.111	-2.029	0.108
			3	-0.718	0.505	0.892	-2.237	0.801
			4	-0.670	0.441	0.854	-2.018	0.678
			5	-0.086	0.441	1.000	-1.434	1.262
			6	-1.692*	0.392	**0.001**	-2.901	-0.484
		2	1	0.961	0.348	0.111	-0.108	2.029
			3	0.242	0.480	1.000	-1.168	1.653
			4	0.290	0.412	1.000	-0.950	1.531
			5	0.875	0.412	0.390	-0.365	2.115
			6	-0.732	0.360	0.480	-1.832	0.369
		3	1	0.718	0.505	0.892	-0.801	2.237
			2	-0.242	0.480	1.000	-1.653	1.168
			4	0.048	0.551	1.000	-1.642	1.738
			5	0.632	0.551	0.981	-1.057	2.322
			6	-0.974	0.513	0.567	-2.525	0.577
		4	1	0.670	0.441	0.854	-0.678	2.018
			2	-0.290	0.412	1.000	-1.531	0.950
			3	-0.048	0.551	1.000	-1.738	1.642
			5	0.584	0.493	0.977	-0.935	2.104

			6	-1.022	0.450	0.320	-2.402	0.358
		5	1	0.086	0.441	1.000	-1.262	1.434
			2	-0.875	0.412	0.390	-2.115	0.365
			3	-0.632	0.551	0.981	-2.322	1.057
			4	-0.584	0.493	0.977	-2.104	0.935
			6	-1.606*	0.450	**0.012**	-2.986	-0.227
		6	1	1.692*	0.392	**0.001**	0.484	2.901
			2	0.732	0.360	0.480	-0.369	1.832
			3	0.974	0.513	0.567	-0.577	2.525
			4	1.022	0.450	0.320	-0.358	2.402
			5	1.606*	0.450	**0.012**	0.227	2.986
	Hochberg	1	2	-0.961	0.348	0.114	-2.035	0.114
			3	-0.718	0.505	0.908	-2.275	0.839
			4	-0.670	0.441	0.861	-2.029	0.689
			5	-0.086	0.441	1.000	-1.445	1.273
			6	-1.692*	0.392	**0.001**	-2.901	-0.483
		2	1	0.961	0.348	0.114	-0.114	2.035
			3	0.242	0.480	1.000	-1.238	1.723
			4	0.290	0.412	1.000	-0.981	1.562
			5	0.875	0.412	0.427	-0.396	2.146
			6	-0.732	0.360	0.492	-1.841	0.378
		3	1	0.718	0.505	0.908	-0.839	2.275
			2	-0.242	0.480	1.000	-1.723	1.238
			4	0.048	0.551	1.000	-1.650	1.747
			5	0.632	0.551	0.982	-1.066	2.331
			6	-0.974	0.513	0.596	-2.555	0.607
		4	1	0.670	0.441	0.861	-0.689	2.029
			2	-0.290	0.412	1.000	-1.562	0.981
			3	-0.048	0.551	1.000	-1.747	1.650
			5	0.584	0.493	0.977	-0.935	2.104
			6	-1.022	0.450	0.327	-2.409	0.365
		5	1	0.086	0.441	1.000	-1.273	1.445
			2	-0.875	0.412	0.427	-2.146	0.396
			3	-0.632	0.551	0.982	-2.331	1.066
			4	-0.584	0.493	0.977	-2.104	0.935
			6	-1.606*	0.450	0.013	-2.993	-0.220
		6	1	1.692*	0.392	**0.001**	0.483	2.901
			2	0.732	0.360	0.492	-0.378	1.841
			3	0.974	0.513	0.596	-0.607	2.555
			4	1.022	0.450	0.327	-0.365	2.409
			5	1.606*	0.450	**0.013**	0.220	2.993
Score Z(FS_phw)	Gabriel	1	2	0.274	0.371	1.000	-0.865	1.412
			3	-1.491	0.538	0.093	-3.109	0.127
			4	0.127	0.469	1.000	-1.309	1.563
			5	-0.403	0.469	0.999	-1.839	1.033

		6	0.417	0.418	0.995	-0.871	1.704
	2	1	-0.274	0.371	1.000	-1.412	0.865
		3	-1.764*	0.511	**0.011**	-3.267	-0.262
		4	-0.147	0.439	1.000	-1.468	1.174
		5	-0.677	0.439	0.826	-1.998	0.644
		6	0.143	0.383	1.000	-1.029	1.316
	3	1	1.491	0.538	0.093	-0.127	3.109
		2	1.764*	0.511	**0.011**	0.262	3.267
		4	1.618	0.587	0.111	-0.183	3.418
		5	1.088	0.587	0.624	-0.713	2.888
		6	1.907*	0.546	**0.013**	0.256	3.559
	4	1	-0.127	0.469	1.000	-1.563	1.309
		2	0.147	0.439	1.000	-1.174	1.468
		3	-1.618	0.587	0.111	-3.418	0.183
		5	-0.530	0.525	0.995	-2.148	1.088
		6	0.290	0.479	1.000	-1.180	1.760
	5	1	0.403	0.469	0.999	-1.033	1.839
		2	0.677	0.439	0.826	-0.644	1.998
		3	-1.088	0.587	0.624	-2.888	0.713
		4	0.530	0.525	0.995	-1.088	2.148
		6	0.820	0.479	0.732	-0.650	2.290
	6	1	-0.417	0.418	0.995	-1.704	0.871
		2	-0.143	0.383	1.000	-1.316	1.029
		3	-1.907*	0.546	**0.013**	-3.559	-0.256
		4	-0.290	0.479	1.000	-1.760	1.180
		5	-0.820	0.479	0.732	-2.290	0.650
Hochberg	1	2	0.274	0.371	1.000	-0.871	1.418
		3	-1.491	0.538	0.110	-3.149	0.167
		4	0.127	0.469	1.000	-1.321	1.574
		5	-0.403	0.469	0.999	-1.851	1.044
		6	0.417	0.418	0.995	-0.871	1.705
	2	1	-0.274	0.371	1.000	-1.418	0.871
		3	-1.764*	0.511	**0.018**	-3.342	-0.187
		4	-0.147	0.439	1.000	-1.501	1.207
		5	-0.677	0.439	0.849	-2.031	0.677
		6	0.143	0.383	1.000	-1.039	1.325
	3	1	1.491	0.538	0.110	-0.167	3.149
		2	1.764*	0.511	**0.018**	0.187	3.342
		4	1.618	0.587	0.115	-0.192	3.427
		5	1.088	0.587	0.632	-0.722	2.897
		6	1.907*	0.546	**0.016**	0.223	3.592
	4	1	-0.127	0.469	1.000	-1.574	1.321
		2	0.147	0.439	1.000	-1.207	1.501
		3	-1.618	0.587	0.115	-3.427	0.192
		5	-0.530	0.525	0.995	-2.148	1.088

			6	0.290	0.479	1.000	-1.187	1.767
		5	1	0.403	0.469	0.999	-1.044	1.851
			2	0.677	0.439	0.849	-0.677	2.031
			3	-1.088	0.587	0.632	-2.897	0.722
			4	0.530	0.525	0.995	-1.088	2.148
			6	0.820	0.479	0.738	-0.657	2.297
		6	1	-0.417	0.418	0.995	-1.705	0.871
			2	-0.143	0.383	1.000	-1.325	1.039
			3	-1.907*	0.546	**0.016**	-3.592	-0.223
			4	-0.290	0.479	1.000	-1.767	1.187
			5	-0.820	0.479	0.738	-2.297	0.657
Score Z(L_phw)	Gabriel	1	2	-0.068	0.274	1.000	-0.909	0.774
			3	-0.255	0.398	1.000	-1.450	0.941
			4	-1.030	0.347	0.064	-2.091	0.032
			5	-2.358*	0.347	**0.000**	-3.420	-1.297
			6	-.959*	0.309	**0.047**	-1.912	-0.008
		2	1	0.068	0.274	1.000	-0.774	0.909
			3	-0.187	0.378	1.000	-1.298	0.924
			4	-0.962	0.325	0.056	-1.939	0.014
			5	-2.290*	0.325	**0.000**	-3.267	-1.314
			6	-.892*	0.283	**0.039**	-1.759	-0.025
		3	1	0.255	0.398	1.000	-0.941	1.450
			2	0.187	0.378	1.000	-0.924	1.298
			4	-0.775	0.434	0.675	-2.106	0.555
			5	-2.103*	0.434	**0.000**	-3.435	-0.773
			6	-0.705	0.404	0.687	-1.926	0.516
		4	1	1.030	0.347	0.064	-0.032	2.091
			2	0.962	0.325	0.056	-0.014	1.939
			3	0.775	0.434	0.675	-0.555	2.106
			5	-1.328*	0.388	**0.020**	-2.525	-0.132
			6	0.070	0.354	1.000	-1.017	1.157
		5	1	2.358*	0.347	**0.000**	1.297	3.420
			2	2.290*	0.325	**0.000**	1.314	3.267
			3	2.103*	0.434	**0.000**	0.773	3.435
			4	1.328*	0.388	**0.020**	0.132	2.525
			6	1.398*	0.354	**0.004**	0.312	2.485
		6	1	.959*	0.309	**0.047**	0.008	1.912
			2	.892*	0.283	**0.039**	0.025	1.759
			3	0.705	0.404	0.687	-0.516	1.926
			4	-0.070	0.354	1.000	-1.157	1.017
			5	-1.398*	0.354	**0.004**	-2.485	-0.312
	Hochberg	1	2	-0.068	0.274	1.000	-0.914	0.778
			3	-0.255	0.398	1.000	-1.480	0.971
			4	-1.030	0.347	0.068	-2.100	0.040
			5	-2.358*	0.347	**0.000**	-3.429	-1.289

			6	-.959*	0.309	**0.047**	-1.912	-0.008
		2	1	0.068	0.274	1.000	-0.778	0.914
			3	-0.187	0.378	1.000	-1.353	0.979
			4	-0.962	0.325	0.068	-1.963	0.039
			5	-2.290*	0.325	**0.000**	-3.292	-1.290
			6	-.892*	0.283	**0.042**	-1.766	-0.019
		3	1	0.255	0.398	1.000	-0.971	1.480
			2	0.187	0.378	1.000	-0.979	1.353
			4	-0.775	0.434	0.682	-2.113	0.562
			5	-2.103*	0.434	**0.000**	-3.441	-0.766
			6	-0.705	0.404	0.713	-1.951	0.540
		4	1	1.030	0.347	0.068	-0.040	2.100
			2	0.962	0.325	0.068	-0.039	1.963
			3	0.775	0.434	0.682	-0.562	2.113
			5	-1.328*	0.388	**0.020**	-2.525	-0.132
			6	0.070	0.354	1.000	-1.022	1.162
		5	1	2.358*	0.347	**0.000**	1.289	3.429
			2	2.290*	0.325	**0.000**	1.290	3.292
			3	2.103*	0.434	**0.000**	0.766	3.441
			4	1.328*	0.388	**0.020**	0.132	2.525
			6	1.398*	0.354	**0.004**	0.307	2.491
		6	1	.959*	0.309	**0.047**	0.008	1.912
			2	.892*	0.283	**0.042**	0.019	1.766
			3	0.705	0.404	0.713	-0.540	1.951
			4	-0.070	0.354	1.000	-1.162	1.022
			5	-1.398*	0.354	**0.004**	-2.491	-0.307
Score Z(Rep_phw)	Gabriel	1	2	-0.014	0.343	1.000	-1.066	1.039
			3	-1.553*	0.497	**0.036**	-3.049	-0.058
			4	-1.570*	0.434	**0.010**	-2.898	-0.243
			5	-0.467	0.434	0.989	-1.795	0.860
			6	-0.159	0.386	1.000	-1.349	1.032
		2	1	0.014	0.343	1.000	-1.039	1.066
			3	-1.539*	0.473	**0.020**	-2.929	-0.150
			4	-1.556*	0.406	**0.004**	-2.778	-0.335
			5	-0.453	0.406	0.983	-1.675	0.768
			6	-0.145	0.354	1.000	-1.229	0.939
		3	1	1.553*	0.497	**0.036**	0.058	3.049
			2	1.539*	0.473	**0.020**	0.150	2.929
			4	-0.017	0.542	1.000	-1.682	1.647
			5	1.086	0.542	0.509	-0.578	2.750
			6	1.395	0.505	0.099	-0.132	2.922
		4	1	1.570*	0.434	**0.010**	0.243	2.898
			2	1.556*	0.406	**0.004**	0.335	2.778
			3	0.017	0.542	1.000	-1.647	1.682
			5	1.103	0.485	0.326	-0.393	2.600

			6	1.411*	0.443	**0.036**	0.053	2.771
		5	1	0.467	0.434	0.989	-0.860	1.795
			2	0.453	0.406	0.983	-0.768	1.675
			3	-1.086	0.542	0.509	-2.750	0.578
			4	-1.103	0.485	0.326	-2.600	0.393
			6	0.309	0.443	1.000	-1.050	1.668
		6	1	0.159	0.386	1.000	-1.032	1.349
			2	0.145	0.354	1.000	-0.939	1.229
			3	-1.395	0.505	0.099	-2.922	0.132
			4	-1.411*	0.443	**0.036**	-2.771	-0.053
			5	-0.309	0.443	1.000	-1.668	1.050
	Hochberg	1	2	-0.014	0.343	1.000	-1.072	1.044
			3	-1.553*	0.497	**0.045**	-3.086	-0.020
			4	-1.570*	0.434	**0.011**	-2.909	-0.232
			5	-0.467	0.434	0.990	-1.805	0.871
			6	-0.159	0.386	1.000	-1.349	1.032
		2	1	0.014	0.343	1.000	-1.044	1.072
			3	-1.539*	0.473	**0.031**	-2.998	-0.081
			4	-1.556*	0.406	**0.006**	-2.808	-0.305
			5	-0.453	0.406	0.986	-1.705	0.798
			6	-0.145	0.354	1.000	-1.237	0.948
		3	1	1.553*	0.497	**0.045**	0.020	3.086
			2	1.539*	0.473	**0.031**	0.081	2.998
			4	-0.017	0.542	1.000	-1.690	1.655
			5	1.086	0.542	0.516	-0.587	2.759
			6	1.395	0.505	0.113	-0.163	2.952
		4	1	1.570*	0.434	**0.011**	0.232	2.909
			2	1.556*	0.406	**0.006**	0.305	2.808
			3	0.017	0.542	1.000	-1.655	1.690
			5	1.103	0.485	0.326	-0.393	2.600
			6	1.411*	0.443	**0.038**	0.046	2.778
		5	1	0.467	0.434	0.990	-0.871	1.805
			2	0.453	0.406	0.986	-0.798	1.705
			3	-1.086	0.542	0.516	-2.759	0.587
			4	-1.103	0.485	0.326	-2.600	0.393
			6	0.309	0.443	1.000	-1.057	1.674
		6	1	0.159	0.386	1.000	-1.032	1.349
			2	0.145	0.354	1.000	-0.948	1.237
			3	-1.395	0.505	0.113	-2.952	0.163
			4	-1.411*	0.443	**0.038**	-2.778	-0.046
			5	-0.309	0.443	1.000	-1.674	1.057
Score Z(RS_phw)	Gabriel	1	2	-0.398	0.326	0.969	-1.400	0.604
			3	-2.243*	0.473	**0.000**	-3.667	-0.820
			4	-0.465	0.413	0.984	-1.729	0.798

		5	0.395	0.413	0.997	-0.868	1.659
		6	-0.879	0.367	0.258	-2.012	0.254
	2	1	0.398	0.326	0.969	-0.604	1.400
		3	-1.845*	0.450	**0.001**	-3.168	-0.523
		4	-0.067	0.386	1.000	-1.230	1.095
		5	0.793	0.386	0.440	-0.369	1.955
		6	-0.481	0.337	0.902	-1.513	0.551
	3	1	2.243*	0.473	**0.000**	0.820	3.667
		2	1.845*	0.450	**0.001**	0.523	3.168
		4	1.777*	0.516	**0.018**	0.194	3.362
		5	2.638*	0.516	**0.000**	1.055	4.222
		6	1.364	0.481	0.081	-0.089	2.818
	4	1	0.465	0.413	0.984	-0.798	1.729
		2	0.067	0.386	1.000	-1.095	1.230
		3	-1.777*	0.516	**0.018**	-3.362	-0.194
		5	0.861	0.462	0.623	-0.563	2.285
		6	-0.414	0.421	0.996	-1.707	0.880
	5	1	-0.395	0.413	0.997	-1.659	0.868
		2	-0.793	0.386	0.440	-1.955	0.369
		3	-2.638*	0.516	**0.000**	-4.222	-1.055
		4	-0.861	0.462	0.623	-2.285	0.563
		6	-1.274	0.421	0.056	-2.567	0.019
	6	1	0.879	0.367	0.258	-0.254	2.012
		2	0.481	0.337	0.902	-0.551	1.513
		3	-1.364	0.481	0.081	-2.818	0.089
		4	0.414	0.421	0.996	-0.880	1.707
		5	1.274	0.421	0.056	-0.019	2.567
Hochberg	1	2	-.398	.326	.971	-1.405	.609
		3	-2.243*	.473	**.000**	-3.702	-.784
		4	-.465	.413	.985	-1.739	.808
		5	.395	.413	.997	-.878	1.669
		6	-.879	.367	.259	-2.012	.254
	2	1	.398	.326	.971	-.609	1.405
		3	-1.845*	.450	**.003**	-3.233	-.458
		4	-.067	.386	1.000	-1.259	1.124
		5	.793	.386	.477	-.398	1.985
		6	-.481	.337	.907	-1.521	.559
	3	1	2.243*	.473	**.000**	.784	3.702
		2	1.845*	.450	**.003**	.458	3.233
		4	1.777*	.516	**.019**	.186	3.370
		5	2.638*	.516	**.000**	1.047	4.231
		6	1.364	.481	.094	-.118	2.846

4	1	.465	.413	.985	-.808	1.739
	2	.067	.386	1.000	-1.124	1.259
	3	-1.777*	.516	**.019**	-3.370	-.186
	5	.861	.462	.623	-.563	2.285
	6	-.414	.421	.996	-1.713	.886
5	1	-.395	.413	.997	-1.669	.878
	2	-.793	.386	.477	-1.985	.398
	3	-2.638*	.516	**.000**	-4.231	-1.047
	4	-.861	.462	.623	-2.285	.563
	6	-1.274	.421	.059	-2.574	.025
6	1	0.879	0.367	0.259	-0.254	2.012
	2	0.481	0.337	0.907	-0.559	1.521
	3	-1.364	0.481	0.094	-2.846	0.118
	4	0.414	0.421	0.996	-0.886	1.713
	5	1.274	0.421	0.059	-0.025	2.574

*. La différence moyenne est significative au niveau 0.05.

Table 9-15: Pairwise comparisons (Gabriel's procedure and Hochberg GT2)

Table 9-16 shows the pairwise comparisons for truncations and foreign words using the Games-Howell procedure.

Comparaisons multiples :							
Games-Howell							
Variable dépendante	(I) 6-cluster solution with z-scores	(J) 6-cluster solution with z-scores	Différence moyenne (I-J)	Erreur standard	Sig.	Intervalle de confiance à 95 %	
						Borne inférieure	Borne supérieure
Score Z(T_phw)	1	2	-.611	.196	**.048**	-1.220	-.003
		3	-2.174*	.208	**.000**	-2.973	-1.375
		4	-1.175	.447	.224	-2.968	.616
		5	.480	.360	.761	-.920	1.880
		6	-.367	.411	.939	-1.805	1.070
	2	1	.611*	.196	**.048**	.003	1.220
		3	-1.562*	.223	**.001**	-2.363	-.762
		4	-.564	.454	.805	-2.346	1.218
		5	1.091	.369	.137	-.302	2.486
		6	.244	.418	.990	-1.199	1.688
	3	1	2.174*	.208	**.000**	1.375	2.973
		2	1.562*	.223	**.001**	.762	2.363
		4	.998	.459	.360	-.795	2.792

		5	2.654*	.375	**.002**	1.231	4.077
		6	1.807*	.424	**.015**	.338	3.275
	4	1	1.175	.447	.224	-.616	2.968
		2	.564	.454	.805	-1.218	2.346
		3	-.998	.459	.360	-2.792	.795
		5	1.655	.545	.101	-.260	3.572
		6	.808	.580	.731	-1.149	2.766
	5	1	-.480	.360	.761	-1.880	.920
		2	-1.091	.369	.137	-2.486	.302
		3	-2.654*	.375	**.002**	-4.077	-1.231
		4	-1.655	.545	.101	-3.572	.260
		6	-.847	.516	.589	-2.562	.867
	6	1	.367	.411	.939	-1.070	1.805
		2	-.244	.418	.990	-1.688	1.199
		3	-1.807*	.424	**.015**	-3.275	-.338
		4	-.808	.580	.731	-2.766	1.149
		5	.847	.516	.589	-.867	2.5622
Score Z(W_phw)	1	2	-.269	.321	.957	-1.273	.734
		3	-1.265	1.210	.878	-7.794	5.263
		4	.368	.245	.670	-.466	1.202
		5	-.410	.563	.972	-2.539	1.719
		6	-.246	.270	.938	-1.127	.634
	2	1	.269	.321	.957	-.734	1.273
		3	-.996	1.208	.945	-7.536	5.543
		4	.637	.238	.132	-.124	1.399
		5	-.140	.560	1.000	-2.266	1.984
		6	.023	.264	1.000	-.802	.848
	3	1	1.265	1.210	.878	-5.263	7.794
		2	.996	1.208	.945	-5.543	7.536
		4	1.633	1.190	.745	-5.086	8.353
		5	.855	1.294	.978	-5.172	6.882
		6	1.019	1.196	.937	-5.643	7.682
	4	1	-.368	.245	.670	-1.202	.466
		2	-.637	.238	.132	-1.399	.124
		3	-1.633	1.190	.745	-8.353	5.086
		5	-.778	.520	.681	-2.954	1.397
		6	-.614*	.162	**.024**	-1.159	-.068
	5	1	.410	.563	.972	-1.719	2.539
		2	.140	.560	1.000	-1.984	2.266
		3	-.855	1.294	.978	-6.882	5.172
		4	.778	.520	.681	-1.397	2.954
		6	.164	.532	.999	-1.987	2.316

	6	1	.246	.270	.938	-.634	1.127
		2	-.023	.264	1.000	-.848	.802
		3	-1.019	1.196	.937	-7.682	5.643
		4	.614*	.162	**.024**	.068	1.159
		5	-.164	.532	.999	-2.316	1.987
*. La différence moyenne est significative au niveau 0.05.							

Table 9-16: Games-Howell results (for T and W phw)

9.8.1.4 ANOVA results (5 (dis)fluency components)

Table 9-17 shows the results of Levene's test of homogeneity of variances in LINDSEI-FR+ with the 5 (dis)fluency components. Table 9-18 displays the results of the ANOVA using the 5 (dis)fluency components.

Note: significant results are in bold font.

Test d'homogénéité des variances				
	Statistique de Levene	ddl1	ddl2	Sig.
factor score component 1 (Anderson Rubin)	1.297	5	44	**.282**
factor score component 2 (Anderson Rubin)	.504	5	44	**.772**
factor score component 3 (Anderson Rubin)	1.884	5	44	**.117**
factor score component 4 (Anderson Rubin)	1.261	5	44	**.298**
factor score component 5 (Anderson Rubin)	1.836	5	44	**.126**

Table 9-17: Levene's test of homogeneity of variances

ANOVA						
		Somme des carrés	ddl	Carré moyen	F	Sig.
factor score component 1 (Anderson Rubin)	Inter-groupes	31.970	5	6.394	16.520	**.000**
	Intragroupes	17.030	44	.387		
	Total	49.000	49			
factor score component 2 (Anderson Rubin)	Inter-groupes	26.532	5	5.306	10.391	**.000**
	Intragroupes	22.468	44	.511		
	Total	49.000	49			
factor score component 3 (Anderson Rubin)	Inter-groupes	23.728	5	4.746	8.262	**.000**
	Intragroupes	25.272	44	.574		
	Total	49.000	49			
factor score component 4 (Anderson Rubin)	Inter-groupes	4.414	5	.883	.871	.508
	Intragroupes	44.586	44	1.013		
	Total	49.000	49			
	Inter-groupes	8.314	5	1.663	1.824	.128
	Intragroupes	40.103	44	.911		

factor score component 5 (Anderson Rubin)	Total	48.417	49				

Table 9-18: ANOVA results

Table 9-19 displays the results of pairwise comparisons using the Gabriel and Hochberg's procedures. Note: 1 = cluster A; 2 = cluster D; 3 = cluster B; 4 = cluster E; 5 = cluster F; 6 = cluster C.

Comparaisons multiples								
Variable dépendante				Différence moyenne (I-J)	Erreur standard	Sig.	Intervalle de confiance à 95 %	
							Borne inférieure	Borne supérieure
factor score comp 1 Anderson Rubin	Gabriel	1	2	1.20126*	.25398	**.000**	.4220	1.9805
			3	.24488	.36806	1.000	-.8624	1.3522
			4	1.73771*	.32127	**.000**	.7549	2.7206
			5	1.44717*	.32127	**.001**	.4643	2.4300
			6	-.48765	.28585	.742	-1.3689	.3936
		2	1	-1.20126*	.25398	**.000**	-1.9805	-.4220
			3	-.95637	.35009	.087	-1.9850	.0722
			4	.53646	.30052	.650	-.3677	1.4406
			5	.24591	.30052	.999	-.6582	1.1500
			6	-1.68891*	.26231	**.000**	-2.4914	-.8864
		3	1	-.24488	.36806	1.000	-1.3522	.8624
			2	.95637	.35009	.087	-.0722	1.9850
			4	1.49283*	.40158	**.008**	.2607	2.7250
			5	1.20229	.40158	.061	-.0299	2.4344
			6	-.73254	.37385	.519	-1.8631	.3980
		4	1	-1.73771*	.32127	**.000**	-2.7206	-.7549
			2	-.53646	.30052	.650	-1.4406	.3677
			3	-1.49283*	.40158	**.008**	-2.7250	-.2607
			5	-.29054	.35919	.999	-1.3982	.8171
			6	-2.22537*	.32789	**.000**	-3.2314	-1.2193
		5	1	-1.44717*	.32127	**.001**	-2.4300	-.4643
			2	-.24591	.30052	.999	-1.1500	.6582
			3	-1.20229	.40158	.061	-2.4344	.0299
			4	.29054	.35919	.999	-.8171	1.3982
			6	-1.93482*	.32789	**.000**	-2.9409	-.9288
		6	1	.48765	.28585	.742	-.3936	1.3689
			2	1.68891*	.26231	**.000**	.8864	2.4914
			3	.73254	.37385	.519	-.3980	1.8631
			4	2.22537*	.32789	**.000**	1.2193	3.2314
			5	1.93482*	.32789	**.000**	.9288	2.9409
	Hochberg	1	2	1.20126*	.25398	**.000**	.4180	1.9845

			3	.24488	.36806	1.000	-.8902	1.3799
			4	1.73771*	.32127	**.000**	.7470	2.7285
			5	1.44717*	.32127	**.001**	.4564	2.4379
			6	-.48765	.28585	.743	-1.3692	.3939
		2	1	-1.20126*	.25398	**.000**	-1.9845	-.4180
			3	-.95637	.35009	.122	-2.0360	.1233
			4	.53646	.30052	.684	-.3903	1.4632
			5	.24591	.30052	.999	-.6808	1.1727
			6	-1.68891*	.26231	**.000**	-2.4978	-.8800
		3	1	-.24488	.36806	1.000	-1.3799	.8902
			2	.95637	.35009	.122	-.1233	2.0360
			4	1.49283*	.40158	**.008**	.2544	2.7313
			5	1.20229	.40158	.063	-.0361	2.4407
			6	-.73254	.37385	.549	-1.8854	.4204
		4	1	-1.73771*	.32127	**.000**	-2.7285	-.7470
			2	-.53646	.30052	.684	-1.4632	.3903
			3	-1.49283*	.40158	**.008**	-2.7313	-.2544
			5	-.29054	.35919	.999	-1.3982	.8171
			6	-2.22537*	.32789	**.000**	-3.2365	-1.2142
		5	1	-1.44717*	.32127	**.001**	-2.4379	-.4564
			2	-.24591	.30052	.999	-1.1727	.6808
			3	-1.20229	.40158	.063	-2.4407	.0361
			4	.29054	.35919	.999	-.8171	1.3982
			6	-1.93482*	.32789	**.000**	-2.9460	-.9237
		6	1	.48765	.28585	.743	-.3939	1.3692
			2	1.68891*	.26231	**.000**	.8800	2.4978
			3	.73254	.37385	.549	-.4204	1.8854
			4	2.22537*	.32789	**.000**	1.2142	3.2365
			5	1.93482*	.32789	**.000**	.9237	2.9460
factor score comp 2 Anderson Rubin	Gabriel	1	2	-.37146	.29173	.957	-1.2666	.5236
			3	-2.40681*	.42276	**.000**	-3.6787	-1.1349
			4	-1.44893*	.36902	**.004**	-2.5778	-.3200
			5	.30547	.36902	.999	-.8234	1.4344
			6	-.56097	.32833	.740	-1.5732	.4512
		2	1	.37146	.29173	.957	-.5236	1.2666
			3	-2.03535*	.40213	**.000**	-3.2168	-.8539
			4	-1.07747*	.34518	**.037**	-2.1160	-.0390
			5	.67693	.34518	.510	-.3616	1.7154
			6	-.18951	.30130	1.000	-1.1113	.7322
		3	1	2.40681*	.42276	**.000**	1.1349	3.6787
			2	2.03535*	.40213	**.000**	.8539	3.2168
			4	.95788	.46127	.453	-.4574	2.3732
			5	2.71227*	.46127	**.000**	1.2970	4.1276
			6	1.84584*	.42942	**.001**	.5473	3.1444
		4	1	1.44893*	.36902	**.004**	.3200	2.5778

			2	1.07747*	.34518	**.037**	.0390	2.1160
			3	-.95788	.46127	.453	-2.3732	.4574
			5	1.75440*	.41257	**.002**	.4821	3.0267
			6	.88796	.37663	.271	-.2676	2.0435
		5	1	-.30547	.36902	.999	-1.4344	.8234
			2	-.67693	.34518	.510	-1.7154	.3616
			3	-2.71227*	.46127	**.000**	-4.1276	-1.2970
			4	-1.75440*	.41257	**.002**	-3.0267	-.4821
			6	-.86644	.37663	.303	-2.0220	.2891
		6	1	.56097	.32833	.740	-.4512	1.5732
			2	.18951	.30130	1.000	-.7322	1.1113
			3	-1.84584*	.42942	**.001**	-3.1444	-.5473
			4	-.88796	.37663	.271	-2.0435	.2676
			5	.86644	.37663	.303	-.2891	2.0220
	Hochberg	1	2	-.37146	.29173	.959	-1.2711	.5282
			3	-2.40681*	.42276	**.000**	-3.7105	-1.1031
			4	-1.44893*	.36902	**.004**	-2.5869	-.3109
			5	.30547	.36902	.999	-.8325	1.4435
			6	-.56097	.32833	.741	-1.5735	.4516
		2	1	.37146	.29173	.959	-.5282	1.2711
			3	-2.03535*	.40213	**.000**	-3.2754	-.7953
			4	-1.07747*	.34518	**.045**	-2.1420	-.0130
			5	.67693	.34518	.548	-.3876	1.7414
			6	-.18951	.30130	1.000	-1.1187	.7397
		3	1	2.40681*	.42276	**.000**	1.1031	3.7105
			2	2.03535*	.40213	**.000**	.7953	3.2754
			4	.95788	.46127	.460	-.4646	2.3804
			5	2.71227*	.46127	**.000**	1.2898	4.1348
			6	1.84584*	.42942	**.001**	.5216	3.1701
		4	1	1.44893*	.36902	**.004**	.3109	2.5869
			2	1.07747*	.34518	**.045**	.0130	2.1420
			3	-.95788	.46127	.460	-2.3804	.4646
			5	1.75440*	.41257	**.002**	.4821	3.0267
			6	.88796	.37663	.277	-.2735	2.0494
		5	1	-.30547	.36902	.999	-1.4435	.8325
			2	-.67693	.34518	.548	-1.7414	.3876
			3	-2.71227*	.46127	**.000**	-4.1348	-1.2898
			4	-1.75440*	.41257	**.002**	-3.0267	-.4821
			6	-.86644	.37663	.310	-2.0279	.2950
		6	1	.56097	.32833	.741	-.4516	1.5735
			2	.18951	.30130	1.000	-.7397	1.1187
			3	-1.84584*	.42942	**.001**	-3.1701	-.5216
			4	-.88796	.37663	.277	-2.0494	.2735
			5	.86644	.37663	.310	-.2950	2.0279
	Gabriel	1	2	1.40304*	.30940	**.001**	.4537	2.3523

factor score comp 3 Anderson Rubin			3	.48082	.44836	.988	-.8681	1.8297
			4	.46640	.39136	.974	-.7309	1.6637
			5	.64588	.39136	.772	-.5514	1.8432
			6	1.93822*	.34822	**.000**	.8647	3.0117
		2	1	-1.40304*	.30940	.001	-2.3523	-.4537
			3	-.92222	.42647	.329	-2.1752	.3308
			4	-.93665	.36608	.157	-2.0380	.1647
			5	-.75716	.36608	.429	-1.8585	.3442
			6	.53517	.31954	.755	-.4424	1.5127
		3	1	-.48082	.44836	.988	-1.8297	.8681
			2	.92222	.42647	.329	-.3308	2.1752
			4	-.01442	.48920	1.000	-1.5154	1.4866
			5	.16506	.48920	1.000	-1.3359	1.6660
			6	1.45739*	.45542	**.031**	.0802	2.8346
		4	1	-.46640	.39136	.974	-1.6637	.7309
			2	.93665	.36608	.157	-.1647	2.0380
			3	.01442	.48920	1.000	-1.4866	1.5154
			5	.17948	.43755	1.000	-1.1699	1.5288
			6	1.47182*	.39943	**.009**	.2463	2.6974
		5	1	-.64588	.39136	.772	-1.8432	.5514
			2	.75716	.36608	.429	-.3442	1.8585
			3	-.16506	.48920	1.000	-1.6660	1.3359
			4	-.17948	.43755	1.000	-1.5288	1.1699
			6	1.29234*	.39943	**.032**	.0668	2.5179
		6	1	-1.93822*	.34822	**.000**	-3.0117	-.8647
			2	-.53517	.31954	.755	-1.5127	.4424
			3	-1.45739*	.45542	**.031**	-2.8346	-.0802
			4	-1.47182*	.39943	**.009**	-2.6974	-.2463
			5	-1.29234*	.39943	**.032**	-2.5179	-.0668
	Hochberg	1	2	1.40304*	.30940	**.001**	.4489	2.3572
			3	.48082	.44836	.990	-.9018	1.8635
			4	.46640	.39136	.976	-.7405	1.6733
			5	.64588	.39136	.781	-.5610	1.8528
			6	1.93822*	.34822	**.000**	.8644	3.0121
		2	1	-1.40304*	.30940	**.001**	-2.3572	-.4489
			3	-.92222	.42647	.399	-2.2374	.3930
			4	-.93665	.36608	.181	-2.0656	.1923
			5	-.75716	.36608	.466	-1.8861	.3718
			6	.53517	.31954	.765	-.4503	1.5206
		3	1	-.48082	.44836	.990	-1.8635	.9018
			2	.92222	.42647	.399	-.3930	2.2374
			4	-.01442	.48920	1.000	-1.5230	1.4942
			5	.16506	.48920	1.000	-1.3436	1.6737
			6	1.45739*	.45542	**.037**	.0529	2.8618
		4	1	-.46640	.39136	.976	-1.6733	.7405

			2	.93665	.36608	.181	-.1923	2.0656
			3	.01442	.48920	1.000	-1.4942	1.5230
			5	.17948	.43755	1.000	-1.1699	1.5288
			6	1.47182*	.39943	**.009**	.2400	2.7036
		5	1	-.64588	.39136	.781	-1.8528	.5610
			2	.75716	.36608	.466	-.3718	1.8861
			3	-.16506	.48920	1.000	-1.6737	1.3436
			4	-.17948	.43755	1.000	-1.5288	1.1699
			6	1.29234*	.39943	**.033**	.0606	2.5241
		6	1	-1.93822*	.34822	**.000**	-3.0121	-.8644
			2	-.53517	.31954	.765	-1.5206	.4503
			3	-1.45739*	.45542	**.037**	-2.8618	-.0529
			4	-1.47182*	.39943	**.009**	-2.7036	-.2400
			5	-1.29234*	.39943	**.033**	-2.5241	-.0606
factor score comp 4 Anderson Rubin	Gabriel	1	2	.02615	.41096	1.000	-1.2348	1.2871
			3	-.84323	.59553	.896	-2.6349	.9485
			4	.43161	.51983	.999	-1.1587	2.0219
			5	-.19677	.51983	1.000	-1.7871	1.3935
			6	.15285	.46252	1.000	-1.2730	1.5787
		2	1	-.02615	.41096	1.000	-1.2871	1.2348
			3	-.86938	.56647	.807	-2.5337	.7949
			4	.40546	.48625	.999	-1.0574	1.8684
			5	-.22292	.48625	1.000	-1.6858	1.2400
			6	.12670	.42444	1.000	-1.1718	1.4252
		3	1	.84323	.59553	.896	-.9485	2.6349
			2	.86938	.56647	.807	-.7949	2.5337
			4	1.27484	.64978	.540	-.7188	3.2685
			5	.64646	.64978	.995	-1.3472	2.6401
			6	.99608	.60491	.762	-.8332	2.8253
		4	1	-.43161	.51983	.999	-2.0219	1.1587
			2	-.40546	.48625	.999	-1.8684	1.0574
			3	-1.27484	.64978	.540	-3.2685	.7188
			5	-.62838	.58118	.990	-2.4207	1.1639
			6	-.27876	.53054	1.000	-1.9066	1.3491
		5	1	.19677	.51983	1.000	-1.3935	1.7871
			2	.22292	.48625	1.000	-1.2400	1.6858
			3	-.64646	.64978	.995	-2.6401	1.3472
			4	.62838	.58118	.990	-1.1639	2.4207
			6	.34962	.53054	1.000	-1.2782	1.9775
		6	1	-.15285	.46252	1.000	-1.5787	1.2730
			2	-.12670	.42444	1.000	-1.4252	1.1718
			3	-.99608	.60491	.762	-2.8253	.8332
			4	.27876	.53054	1.000	-1.3491	1.9066
			5	-.34962	.53054	1.000	-1.9775	1.2782
	Hochberg	1	2	.02615	.41096	1.000	-1.2412	1.2935

			3	-.84323	.59553	.911	-2.6798	.9933
			4	.43161	.51983	.999	-1.1715	2.0347
			5	-.19677	.51983	1.000	-1.7998	1.4063
			6	.15285	.46252	1.000	-1.2735	1.5792
		2	1	-.02615	.41096	1.000	-1.2935	1.2412
			3	-.86938	.56647	.853	-2.6163	.8775
			4	.40546	.48625	.999	-1.0941	1.9050
			5	-.22292	.48625	1.000	-1.7225	1.2766
			6	.12670	.42444	1.000	-1.1822	1.4356
		3	1	.84323	.59553	.911	-.9933	2.6798
			2	.86938	.56647	.853	-.8775	2.6163
			4	1.27484	.64978	.547	-.7290	3.2787
			5	.64646	.64978	.995	-1.3574	2.6503
			6	.99608	.60491	.784	-.8694	2.8615
		4	1	-.43161	.51983	.999	-2.0347	1.1715
			2	-.40546	.48625	.999	-1.9050	1.0941
			3	-1.27484	.64978	.547	-3.2787	.7290
			5	-.62838	.58118	.990	-2.4207	1.1639
			6	-.27876	.53054	1.000	-1.9149	1.3574
		5	1	.19677	.51983	1.000	-1.4063	1.7998
			2	.22292	.48625	1.000	-1.2766	1.7225
			3	-.64646	.64978	.995	-2.6503	1.3574
			4	.62838	.58118	.990	-1.1639	2.4207
			6	.34962	.53054	1.000	-1.2865	1.9857
		6	1	-.15285	.46252	1.000	-1.5792	1.2735
			2	-.12670	.42444	1.000	-1.4356	1.1822
			3	-.99608	.60491	.784	-2.8615	.8694
			4	.27876	.53054	1.000	-1.3574	1.9149
			5	-.34962	.53054	1.000	-1.9857	1.2865
factor score comp 5 Anderson Rubin	Gabriel	1	2	-.17612	.38975	1.000	-1.3720	1.0197
			3	-1.33567	.56480	.242	-3.0349	.3636
			4	.47444	.49300	.996	-1.0338	1.9827
			5	-.28796	.49300	1.000	-1.7962	1.2202
			6	-.12292	.43865	1.000	-1.4752	1.2293
		2	1	.17612	.38975	1.000	-1.0197	1.3720
			3	-1.15955	.53723	.331	-2.7380	.4189
			4	.65056	.46116	.898	-.7368	2.0380
			5	-.11185	.46116	1.000	-1.4993	1.2756
			6	.05320	.40253	1.000	-1.1783	1.2846
		3	1	1.33567	.56480	.242	-.3636	3.0349
			2	1.15955	.53723	.331	-.4189	2.7380
			4	1.81011	.61625	.070	-.0807	3.7009
			5	1.04770	.61625	.741	-.8431	2.9385
			6	1.21275	.57370	.404	-.5221	2.9476
		4	1	-.47444	.49300	.996	-1.9827	1.0338

			2	-.65056	.46116	.898	-2.0380	.7368
			3	-1.81011	.61625	.070	-3.7009	.0807
			5	-.76240	.55119	.924	-2.4622	.9374
			6	-.59736	.50316	.975	-2.1412	.9465
		5	1	.28796	.49300	1.000	-1.2202	1.7962
			2	.11185	.46116	1.000	-1.2756	1.4993
			3	-1.04770	.61625	.741	-2.9385	.8431
			4	.76240	.55119	.924	-.9374	2.4622
			6	.16504	.50316	1.000	-1.3788	1.7089
		6	1	.12292	.43865	1.000	-1.2293	1.4752
			2	-.05320	.40253	1.000	-1.2846	1.1783
			3	-1.21275	.57370	.404	-2.9476	.5221
			4	.59736	.50316	.975	-.9465	2.1412
			5	-.16504	.50316	1.000	-1.7089	1.3788
	Hochberg	1	2	-.17612	.38975	1.000	-1.3780	1.0258
			3	-1.33567	.56480	.273	-3.0774	.4061
			4	.47444	.49300	.997	-1.0459	1.9948
			5	-.28796	.49300	1.000	-1.8083	1.2324
			6	-.12292	.43865	1.000	-1.4756	1.2298
		2	1	.17612	.38975	1.000	-1.0258	1.3780
			3	-1.15955	.53723	.402	-2.8163	.4972
			4	.65056	.46116	.913	-.7716	2.0727
			5	-.11185	.46116	1.000	-1.5340	1.3103
			6	.05320	.40253	1.000	-1.1881	1.2945
		3	1	1.33567	.56480	.273	-.4061	3.0774
			2	1.15955	.53723	.402	-.4972	2.8163
			4	1.81011	.61625	.073	-.0903	3.7105
			5	1.04770	.61625	.747	-.8527	2.9481
			6	1.21275	.57370	.433	-.5564	2.9819
		4	1	-.47444	.49300	.997	-1.9948	1.0459
			2	-.65056	.46116	.913	-2.0727	.7716
			3	-1.81011	.61625	.073	-3.7105	.0903
			5	-.76240	.55119	.924	-2.4622	.9374
			6	-.59736	.50316	.976	-2.1490	.9543
		5	1	.28796	.49300	1.000	-1.2324	1.8083
			2	.11185	.46116	1.000	-1.3103	1.5340
			3	-1.04770	.61625	.747	-2.9481	.8527
			4	.76240	.55119	.924	-.9374	2.4622
			6	.16504	.50316	1.000	-1.3866	1.7167
		6	1	.12292	.43865	1.000	-1.2298	1.4756
			2	-.05320	.40253	1.000	-1.2945	1.1881
			3	-1.21275	.57370	.433	-2.9819	.5564
			4	.59736	.50316	.976	-.9543	2.1490
			5	-.16504	.50316	1.000	-1.7167	1.3866

*. La différence moyenne est significative au niveau 0.05.

Table 9-19: Post hoc comparisons (Gabriel's procedure and Hochberg GT2)

9.8.2 LOCNEC+

9.8.2.1 The make-up of the clusters

Table 9-20 shows the ID of the speakers in the two clusters from LOCNEC+. Table 9-21 shows the ID of the native speakers in the 5-cluster solution.

Cluster 1			**Cluster 2**	
EN001 EN002 EN005 EN006 EN007 EN011 EN013 EN014 EN017	EN019 EN020 EN021 EN027 EN028 EN029 EN030 EN031 EN032	EN034 EN035 EN040 EN041 EN042 EN043 EN046 EN050	EN003 EN008 EN009 EN010 EN012 EN015 EN016 EN018 EN022 EN023	EN024 EN025 EN026 EN033 EN036 EN037 EN039 EN044 EN045 EN048
n = 26			*n* = 20	

Table 9-20: The make-up of the 2 main clusters in LOCNEC+ (n=46)

Cluster A	**Cluster B**	**Cluster C**	**Cluster D**	**Cluster E**
EN001 EN005 EN006 EN017 EN042	EN002 EN011 EN013 EN019 EN020 EN040 EN041 EN046 EN050	EN007 EN014 EN021 EN027 EN028 EN029 EN030 EN031 EN032 EN034 EN035 EN043	EN003 EN010 EN012 EN023 EN024 EN025 EN033 EN036 EN039 EN044 EN045	EN008 EN009 EN015 EN016 EN018 EN022 EN026 EN037 EN048
n = 5	*n* = 9	*n* = 12	*n* = 11	*n* = 9

Table 9-21: The make-up of the 6 clusters in LOCNEC+ (n = 46)

9.8.2.2 Cluster profiles per (dis)fluency component (5-cluster solution)

Figure 9-11 to 9-15 show the cluster profiles per (dis)fluency component in the 5-cluster solution of LOCNEC+.

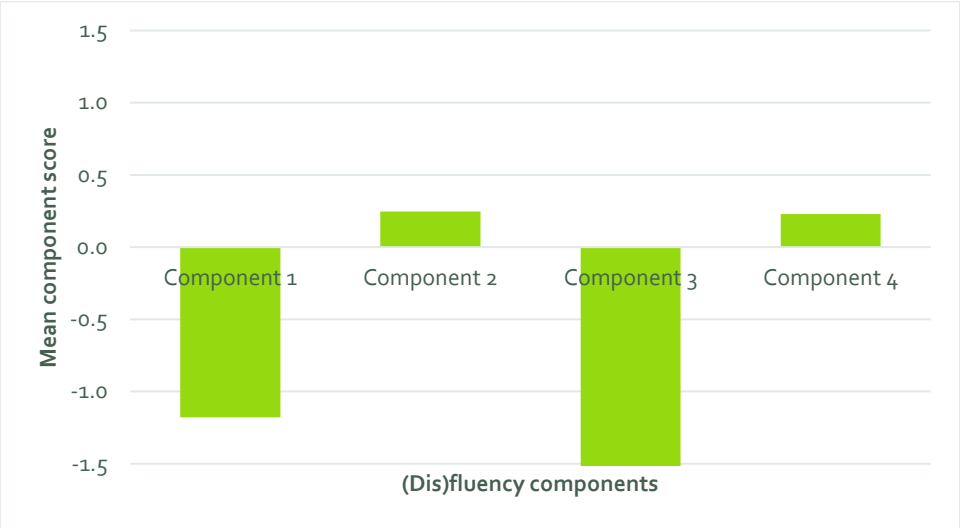


Figure 9-11: Cluster A profile per (dis)fluency components in LOCNEC+

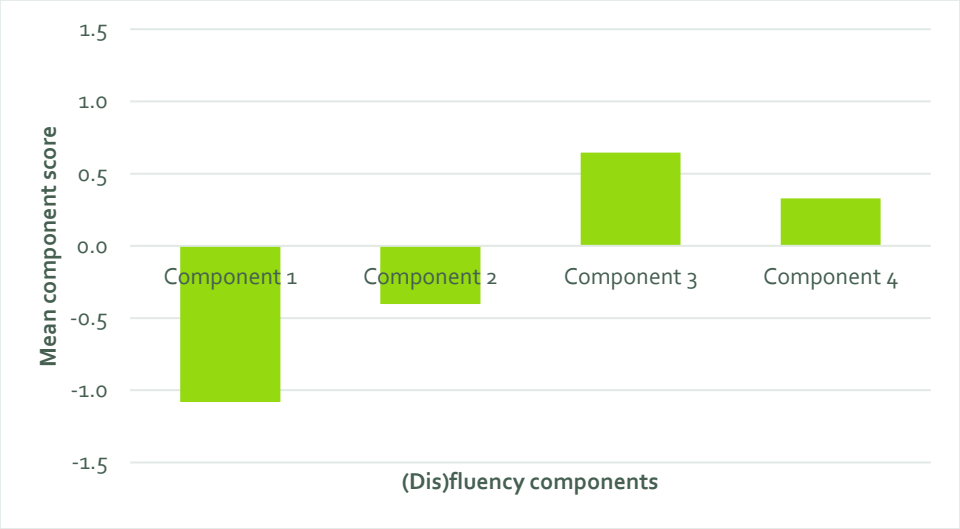


Figure 9-12: Cluster B profile per (dis)fluency components in LOCNEC+

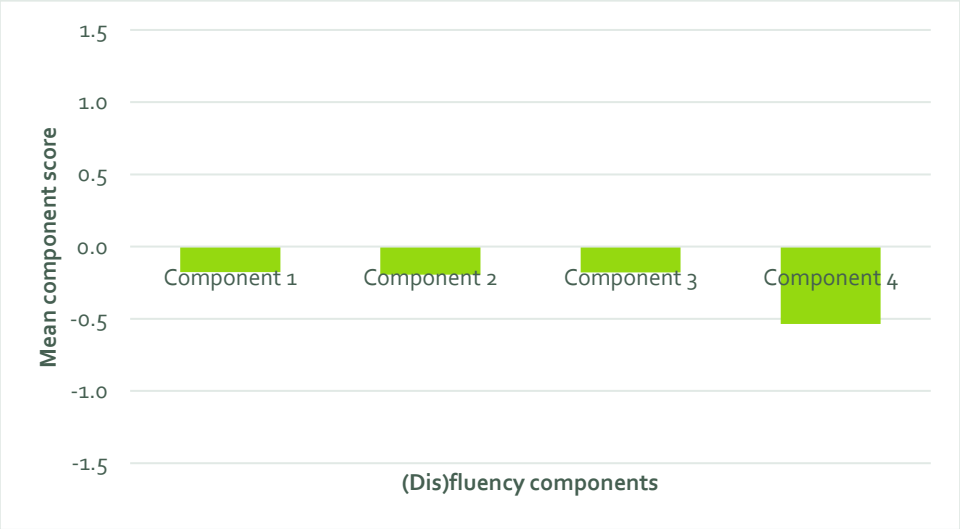


Figure 9-13: Cluster C profile per (dis)fluency components in LOCNEC+

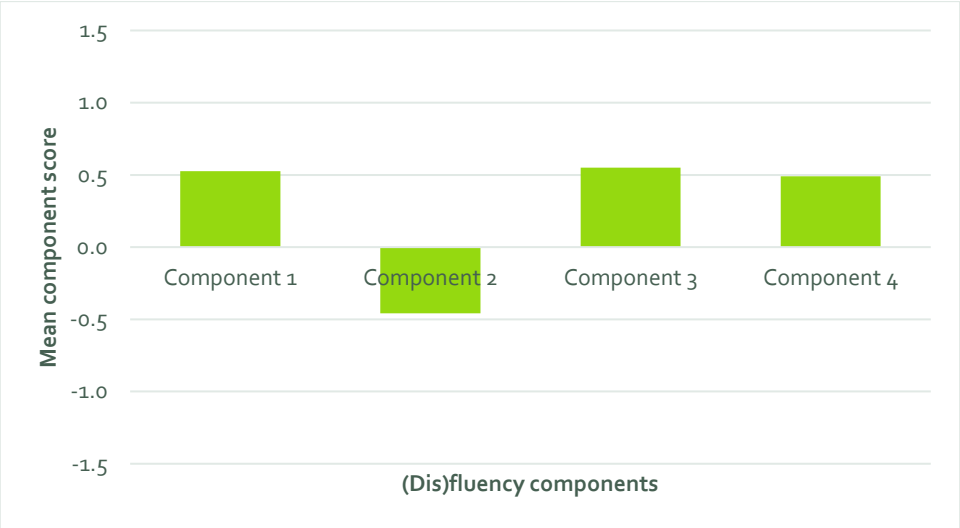


Figure 9-14: Cluster D profile per (dis)fluency components in LOCNEC+

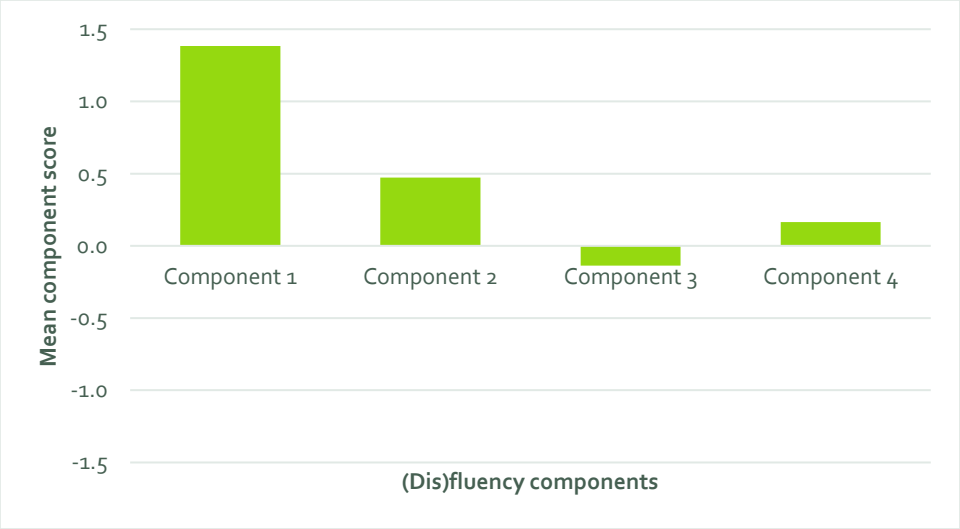


Figure 9-15: Cluster E profile per (dis)fluency components in LOCNEC+

9.8.2.3 ANOVA results (14 (dis)fluency variables)

Table 9-22 shows the results of Levene's test of homogeneity of variances. Table 9-23 displays the results of the ANOVA test.

Note: significant results are shown in bold font.

Test d'homogénéité des variances				
	Statistique de Levene	ddl1	ddl2	Sig.
Score Z(C_phw)	1.970	4	41	.117
Score Z(DM_phw)	2.050	4	41	.105
Score Z(FP_phw)	3.076	4	41	**.026**
Score Z(FS_phw)	1.932	4	41	.123
Score Z(L_phw)	1.306	4	41	.284

Score Z(Mean_length_of_runs)	1.495	4	41	.222
Score Z(Mean_UP_length_sec)	3.036	4	41	**.028**
Score Z(Phonation_time_ratio)	1.962	4	41	.119
Score Z(Rep_phw)	3.429	4	41	**.017**
Score Z(RS_phw)	.448	4	41	.773
Score Z(Speech_rate_wpm)	1.351	4	41	.268
Score Z(T_phw)	2.800	4	41	**.038**
Score Z(UP_phw)	1.771	4	41	.153
Score Z(W_phw)	11.435	4	41	**.000**

Table 9-22: Levene's test of homogeneity of variances

ANOVA						
		Somme des carrés	ddl	Carré moyen	F	Sig.
Score Z(C_phw)	Inter-groupes	6.070	4	1.518	2.232	.082
	Intragroupes	27.881	41	.680		
	Total	33.951	45			
Score Z(DM_phw)	Inter-groupes	10.259	4	2.565	3.234	**.021**
	Intragroupes	32.514	41	.793		
	Total	42.773	45			
Score Z(FP_phw)	*Inter-groupes*	*27.959*	*4*	*6.990*	*15.374*	*.000*
	Intragroupes	*18.641*	*41*	*.455*		
	Total	*46.600*	*45*			
Score Z(FS_phw)	Inter-groupes	11.577	4	2.894	4.034	**.008**
	Intragroupes	29.419	41	.718		
	Total	40.996	45			
Score Z(L_phw)	Inter-groupes	5.442	4	1.361	1.641	.182
	Intragroupes	33.988	41	.829		
	Total	39.431	45			
Score Z(Mean_length_of_runs)	Inter-groupes	31.004	4	7.751	18.351	**.000**
	Intragroupes	17.318	41	.422		
	Total	48.322	45			
Score Z(Mean_UP_length_sec)	*Inter-groupes*	*18.926*	*4*	*4.732*	*8.162*	*.000*
	Intragroupes	*23.769*	*41*	*.580*		
	Total	*42.696*	*45*			
Score Z(Phonation_time_ratio)	Inter-groupes	29.970	4	7.493	19.244	**.000**
	Intragroupes	15.963	41	.389		
	Total	45.934	45			
Score Z(Rep_phw)	*Inter-groupes*	*18.187*	*4*	*4.547*	*7.696*	*.000*
	Intragroupes	*24.224*	*41*	*.591*		
	Total	*42.412*	*45*			
Score Z(RS_phw)	Inter-groupes	2.163	4	.541	.918	.463
	Intragroupes	24.149	41	.589		
	Total	26.312	45			

Score Z(Speech_rate_wpm)	Inter-groupes	26.293	4	6.573	14.373	**.000**
	Intragroupes	18.751	41	.457		
	Total	45.045	45			
Score Z(T_phw)	*Inter-groupes*	*.725*	*4*	*.181*	*.494*	*.740*
	Intragroupes	*15.031*	*41*	*.367*		
	Total	*15.756*	*45*			
Score Z(UP_phw)	Inter-groupes	33.706	4	8.427	27.362	**.000**
	Intragroupes	12.627	41	.308		
	Total	46.333	45			
Score Z(W_phw)	*Inter-groupes*	*3.257*	*4*	*.814*	*1.565*	*.202*
	Intragroupes	*21.333*	*41*	*.520*		
	Total	*24.590*	*45*			

Table 9-23: ANOVA results

Table 9-24 displays the results of the Welch's test.

Tests robustes d'égalité des moyennes					
		Statistiques[a]	ddl1	ddl2	Sig.
Score Z(C_phw)	*Welch*	*3.144*	*4*	*19.801*	*.037*
Score Z(DM_phw)	*Welch*	*8.005*	*4*	*18.532*	*.001*
Score Z(FP_phw)	Welch	5.745	4	15.922	**.005**
Score Z(FS_phw)	*Welch*	*4.062*	*4*	*16.263*	*.018*
Score Z(L_phw)	*Welch*	*1.441*	*4*	*16.359*	*.265*
Score Z(Mean_length_of_runs)	*Welch*	*11.333*	*4*	*17.494*	*.000*
Score Z(Mean_UP_length_sec)	Welch	12.900	4	16.035	**.000**
Score Z(Phonation_time_ratio)	*Welch*	*18.209*	*4*	*16.446*	*.000*
Score Z(Rep_phw)	Welch	3.318	4	15.558	**.038**
Score Z(RS_phw)	*Welch*	*.931*	*4*	*16.656*	*.470*
Score Z(Speech_rate_wpm)	*Welch*	*15.471*	*4*	*16.879*	*.000*
Score Z(T_phw)	Welch	.732	4	17.554	.582
Score Z(UP_phw)	*Welch*	*26.807*	*4*	*15.223*	*.000*
Score Z(W_phw)	Welch	3.767	4	18.144	**.021**
a. F distribué asymptotiquement					

Table 9-24: Welch's F (for FP, Rep, T, W phw & mean length of UP)

Table 9-25 displays the results of the pairwise comparisons using Gabriel and Hochberg's procedures.

Note: 1 = cluster A; 2 = cluster B; 3 = cluster D; 4 = cluster C; 5 = cluster E.

Comparaisons multiples :				
Variable dépendante			Sig.	Intervalle de confiance à 95 %

				Différence moyenne (I-J)	Erreur standard		Borne inférieure	Borne supérieure
Score Z(C_phw)	Gabriel	1	2	-.538	.460	.926	-1.880	.804
			3	-.820	.445	.480	-2.107	.467
			4	.124	.439	1.000	-1.141	1.389
			5	-.390	.460	.991	-1.732	.952
		2	1	.538	.460	.926	-.804	1.880
			3	-.282	.371	.996	-1.374	.809
			4	.662	.364	.519	-.407	1.731
			5	.148	.389	1.000	-.998	1.294
		3	1	.820	.445	.480	-.467	2.107
			2	.282	.371	.996	-.809	1.374
			4	.944	.344	.084	-.071	1.959
			5	.430	.371	.933	-.662	1.521
		4	1	-.124	.439	1.000	-1.389	1.141
			2	-.662	.364	.519	-1.731	.407
			3	-.944	.344	.084	-1.959	.071
			5	-.514	.364	.810	-1.583	.555
		5	1	.390	.460	.991	-.952	1.732
			2	-.148	.389	1.000	-1.294	.998
			3	-.430	.371	.933	-1.521	.662
			4	.514	.364	.810	-.555	1.583
	Hochberg	1	2	-.538	.460	.930	-1.894	.818
			3	-.820	.445	.505	-2.131	.491
			4	.124	.439	1.000	-1.170	1.418
			5	-.390	.460	.992	-1.746	.966
		2	1	.538	.460	.930	-.818	1.894
			3	-.282	.371	.996	-1.375	.811
			4	.662	.364	.522	-.410	1.734
			5	.148	.389	1.000	-.998	1.294
		3	1	.820	.445	.505	-.491	2.131
			2	.282	.371	.996	-.811	1.375
			4	.944	.344	.084	-.071	1.959
			5	.430	.371	.934	-.663	1.523
		4	1	-.124	.439	1.000	-1.418	1.170
			2	-.662	.364	.522	-1.734	.410
			3	-.944	.344	.084	-1.959	.071
			5	-.514	.364	.813	-1.586	.558
		5	1	.390	.460	.992	-.966	1.746
			2	-.148	.389	1.000	-1.294	.998
			3	-.430	.371	.934	-1.523	.663
			4	.514	.364	.813	-.558	1.586
Score Z(DM_phw)	Gabriel	1	2	-1.406	.497	.063	-2.855	.043
			3	-1,435*	.480	**.039**	-2.825	-.045

			4	-.925	.474	.398	-2.291	.441
			5	-1,567*	.497	**.027**	-3.016	-.118
		2	1	1.406	.497	.063	-.043	2.855
			3	-.029	.400	1.000	-1.208	1.149
			4	.481	.393	.908	-.674	1.636
			5	-.161	.420	1.000	-1.399	1.076
		3	1	1,435*	.480	**.039**	.045	2.825
			2	.029	.400	1.000	-1.149	1.208
			4	.510	.372	.837	-.585	1.606
			5	-.132	.400	1.000	-1.311	1.047
		4	1	.925	.474	.398	-.441	2.291
			2	-.481	.393	.908	-1.636	.674
			3	-.510	.372	.837	-1.606	.585
			5	-.642	.393	.657	-1.797	.512
		5	1	1,567*	.497	**.027**	.118	3.016
			2	.161	.420	1.000	-1.076	1.399
			3	.132	.400	1.000	-1.047	1.311
			4	.642	.393	.657	-.512	1.797
	Hochberg	1	2	-1.406	.497	.067	-2.870	.059
			3	-1,435*	.480	**.045**	-2.851	-.019
			4	-.925	.474	.428	-2.322	.473
			5	-1,567*	.497	**.029**	-3.032	-.103
		2	1	1.406	.497	.067	-.059	2.870
			3	-.029	.400	1.000	-1.209	1.151
			4	.481	.393	.910	-.677	1.639
			5	-.161	.420	1.000	-1.399	1.076
		3	1	1,435*	.480	**.045**	.019	2.851
			2	.029	.400	1.000	-1.151	1.209
			4	.510	.372	.837	-.586	1.606
			5	-.132	.400	1.000	-1.312	1.048
		4	1	.925	.474	.428	-.473	2.322
			2	-.481	.393	.910	-1.639	.677
			3	-.510	.372	.837	-1.606	.586
			5	-.642	.393	.660	-1.800	.515
		5	1	1,567*	.497	**.029**	.103	3.032
			2	.161	.420	1.000	-1.076	1.399
			3	.132	.400	1.000	-1.048	1.312
			4	.642	.393	.660	-.515	1.800
Score Z(FS_phw)	Gabriel	1	2	-.975	.472	.341	-2.354	.403
			3	-.961	.457	.307	-2.284	.361
			4	.209	.451	1.000	-1.091	1.508
			5	-.437	.472	.984	-1.815	.941
		2	1	.975	.472	.341	-.403	2.354
			3	.014	.381	1.000	-1.107	1.135
			4	1,183*	.374	**.027**	.085	2.282

			5	.538	.399	.851	-.639	1.716
		3	1	.961	.457	.307	-.361	2.284
			2	-.014	.381	1.000	-1.135	1.107
			4	1,169*	.354	**.019**	.128	2.212
			5	.524	.381	.833	-.597	1.645
		4	1	-.209	.451	1.000	-1.508	1.091
			2	-1,183*	.374	**.027**	-2.282	-.085
			3	-1,169*	.354	**.019**	-2.212	-.128
			5	-.646	.374	.588	-1.744	.453
		5	1	.437	.472	.984	-.941	1.815
			2	-.538	.399	.851	-1.716	.639
			3	-.524	.381	.833	-1.645	.597
			4	.646	.374	.588	-.453	1.744
	Hochberg	1	2	-.975	.472	.354	-2.368	.418
			3	-.961	.457	.330	-2.308	.386
			4	.209	.451	1.000	-1.121	1.538
			5	-.437	.472	.985	-1.830	.956
		2	1	.975	.472	.354	-.418	2.368
			3	.014	.381	1.000	-1.109	1.136
			4	1,183*	.374	**.028**	.083	2.285
			5	.538	.399	.851	-.639	1.716
		3	1	.961	.457	.330	-.386	2.308
			2	-.014	.381	1.000	-1.136	1.109
			4	1,169*	.354	**.019**	.127	2.212
			5	.524	.381	.834	-.598	1.647
		4	1	-.209	.451	1.000	-1.538	1.121
			2	-1,183*	.374	**.028**	-2.285	-.083
			3	-1,169*	.354	**.019**	-2.212	-.127
			5	-.646	.374	.591	-1.747	.456
		5	1	.437	.472	.985	-.956	1.830
			2	-.538	.399	.851	-1.716	.639
			3	-.524	.381	.834	-1.647	.598
			4	.646	.374	.591	-.456	1.747
Score Z(L_phw)	Gabriel	1	2	.648	.508	.881	-.834	2.129
			3	.662	.491	.837	-.759	2.083
			4	.255	.485	1.000	-1.142	1.651
			5	-.221	.508	1.000	-1.702	1.261
		2	1	-.648	.508	.881	-2.129	.834
			3	.014	.409	1.000	-1.191	1.219
			4	-.393	.401	.977	-1.574	.788
			5	-.869	.429	.380	-2.134	.397
		3	1	-.662	.491	.837	-2.083	.759
			2	-.014	.409	1.000	-1.219	1.191
			4	-.407	.380	.959	-1.528	.713
			5	-.883	.409	.298	-2.088	.322

		4	1	-.255	.485	1.000	-1.651	1.142
			2	.393	.401	.977	-.788	1.574
			3	.407	.380	.959	-.713	1.528
			5	-.475	.401	.924	-1.656	.705
		5	1	.221	.508	1.000	-1.261	1.702
			2	.869	.429	.380	-.397	2.134
			3	.883	.409	.298	-.322	2.088
			4	.475	.401	.924	-.705	1.656
	Hochberg	1	2	.648	.508	.887	-.850	2.145
			3	.662	.491	.851	-.786	2.110
			4	.255	.485	1.000	-1.174	1.683
			5	-.221	.508	1.000	-1.718	1.276
		2	1	-.648	.508	.887	-2.145	.850
			3	.014	.409	1.000	-1.192	1.221
			4	-.393	.401	.977	-1.577	.791
			5	-.869	.429	.380	-2.134	.397
		3	1	-.662	.491	.851	-2.110	.786
			2	-.014	.409	1.000	-1.221	1.192
			4	-.407	.380	.959	-1.528	.713
			5	-.883	.409	.299	-2.089	.324
		4	1	-.255	.485	1.000	-1.683	1.174
			2	.393	.401	.977	-.791	1.577
			3	.407	.380	.959	-.713	1.528
			5	-.475	.401	.925	-1.659	.708
		5	1	.221	.508	1.000	-1.276	1.718
			2	.869	.429	.380	-.397	2.134
			3	.883	.409	.299	-.324	2.089
			4	.475	.401	.925	-.708	1.659
Score Z(Mean_length_of_runs)	Gabriel	1	2	-.128	.363	1.000	-1.186	.929
			3	-1.007	.351	.053	-2.022	.007
			4	-.472	.346	.824	-1.469	.525
			5	-2,367*	.363	**.000**	-3.425	-1.310
		2	1	.128	.363	1.000	-.929	1.186
			3	-,879*	.292	**.042**	-1.739	-.019
			4	-.343	.287	.919	-1.186	.499
			5	-2,238*	.306	**.000**	-3.142	-1.336
		3	1	1.007	.351	.053	-.007	2.022
			2	,8790*	.292	**.042**	.019	1.739
			4	.536	.271	.412	-.264	1.335
			5	-1,359*	.292	**.000**	-2.220	-.500
		4	1	.472	.346	.824	-.525	1.469
			2	.343	.287	.919	-.499	1.186
			3	-.536	.271	.412	-1.335	.264
			5	-1,895*	.287	**.000**	-2.738	-1.053
		5	1	2,367*	.363	**.000**	1.310	3.425

			2	2,238*	.306	**.000**	1.336	3.142
			3	1,359*	.292	**.000**	.500	2.220
			4	1,895*	.287	**.000**	1.053	2.738
	Hochberg	1	2	-.128	.363	1.000	-1.197	.940
			3	-1.007	.351	.060	-2.041	.026
			4	-.472	.346	.842	-1.492	.548
			5	-2,367*	.363	**.000**	-3.436	-1.298
		2	1	.128	.363	1.000	-.940	1.197
			3	-,879*	.292	**.043**	-1.740	-.018
			4	-.343	.287	.920	-1.188	.501
			5	-2,238*	.306	**.000**	-3.142	-1.336
		3	1	1.007	.351	.060	-.026	2.041
			2	,879*	.292	**.043**	.018	1.740
			4	.536	.271	.413	-.264	1.335
			5	-1,359*	.292	**.000**	-2.221	-.499
		4	1	.472	.346	.842	-.548	1.492
			2	.343	.287	.920	-.501	1.188
			3	-.536	.271	.413	-1.335	.264
			5	-1,895*	.287	**.000**	-2.740	-1.050
		5	1	2,367*	.363	**.000**	1.298	3.436
			2	2,238*	.306	**.000**	1.336	3.142
			3	1,359*	.292	**.000**	.499	2.221
			4	1,895*	.287	**.000**	1.050	2.740
Score Z(Phonation_time_ratio)	Gabriel	1	2	.132	.348	1.000	-.883	1.147
			3	-1,645*	.337	**.000**	-2.620	-.672
			4	-.477	.332	.778	-1.434	.481
			5	-1,852*	.348	**.000**	-2.868	-.837
		2	1	-.132	.348	1.000	-1.147	.883
			3	-1,777*	.280	**.000**	-2.604	-.952
			4	-.609	.275	.267	-1.418	.201
			5	-1,984*	.294	**.000**	-2.851	-1.117
		3	1	1,645*	.337	**.000**	.672	2.620
			2	1,777*	.280	**.000**	.952	2.604
			4	1,169*	.260	**.001**	.401	1.937
			5	-.206	.280	.997	-1.032	.619
		4	1	.477	.332	.778	-.481	1.434
			2	.609	.275	.267	-.201	1.418
			3	-1,169*	.260	**.001**	-1.937	-.401
			5	-1,375*	.275	**.000**	-2.185	-.567
		5	1	1,852*	.348	**.000**	.837	2.868
			2	1,984*	.294	**.000**	1.117	2.851
			3	.206	.280	.997	-.619	1.032
			4	1,375*	.275	**.000**	.567	2.185
	Hochberg	1	2	.132	.348	1.000	-.894	1.158
			3	-1,645*	.337	**.000**	-2.638	-.654

			4	-.477	.332	.800	-1.456	.503
			5	-1,852*	.348	**.000**	-2.878	-.826
		2	1	-.132	.348	1.000	-1.158	.894
			3	-1,777*	.280	**.000**	-2.605	-.951
			4	-.609	.275	.269	-1.420	.203
			5	-1,984*	.294	**.000**	-2.851	-1.117
		3	1	1,645*	.337	**.000**	.654	2.638
			2	1,777*	.280	**.000**	.951	2.605
			4	1,169*	.260	**.001**	.401	1.937
			5	-.206	.280	.997	-1.033	.620
		4	1	.477	.332	.800	-.503	1.456
			2	.609	.275	.269	-.203	1.420
			3	-1,169*	.260	**.001**	-1.937	-.401
			5	-1,375*	.275	**.000**	-2.187	-.564
		5	1	1,852*	.348	**.000**	.826	2.878
			2	1,984*	.294	**.000**	1.117	2.851
			3	.206	.280	.997	-.620	1.033
			4	1,375*	.275	**.000**	.564	2.187
Score Z(RS_phw)	Gabriel	1	2	.058	.428	1.000	-1.190	1.307
			3	-.065	.414	1.000	-1.263	1.133
			4	-.169	.409	1.000	-1.347	1.008
			5	-.569	.428	.853	-1.818	.680
		2	1	-.058	.428	1.000	-1.307	1.190
			3	-.123	.345	1.000	-1.139	.893
			4	-.228	.338	.999	-1.223	.767
			5	-.627	.362	.587	-1.694	.439
		3	1	.065	.414	1.000	-1.133	1.263
			2	.123	.345	1.000	-.893	1.139
			4	-.105	.320	1.000	-1.049	.840
			5	-.504	.345	.781	-1.520	.512
		4	1	.169	.409	1.000	-1.008	1.347
			2	.228	.338	.999	-.767	1.223
			3	.105	.320	1.000	-.840	1.049
			5	-.400	.338	.925	-1.395	.596
		5	1	.569	.428	.853	-.680	1.818
			2	.627	.362	.587	-.439	1.694
			3	.504	.345	.781	-.512	1.520
			4	.400	.338	.925	-.596	1.395
	Hochberg	1	2	.058	.428	1.000	-1.204	1.320
			3	-.065	.414	1.000	-1.285	1.156
			4	-.169	.409	1.000	-1.374	1.035
			5	-.569	.428	.861	-1.831	.693
		2	1	-.058	.428	1.000	-1.320	1.204
			3	-.123	.345	1.000	-1.140	.894
			4	-.228	.338	.999	-1.225	.770

			5	-.627	.362	.587	-1.694	.439
		3	1	.065	.414	1.000	-1.156	1.285
			2	.123	.345	1.000	-.894	1.140
			4	-.105	.320	1.000	-1.049	.840
			5	-.504	.345	.783	-1.521	.513
		4	1	.169	.409	1.000	-1.035	1.374
			2	.228	.338	.999	-.770	1.225
			3	.105	.320	1.000	-.840	1.049
			5	-.400	.338	.926	-1.397	.598
		5	1	.569	.428	.861	-.693	1.831
			2	.627	.362	.587	-.439	1.694
			3	.504	.345	.783	-.513	1.521
			4	.400	.338	.926	-.598	1.397
Score Z(Speech_rate_wpm)	Gabriel	1	2	-.440	.377	.927	-1.541	.660
			3	-1,275*	.365	**.009**	-2.331	-.220
			4	-.834	.360	.195	-1.872	.203
			5	-2,398*	.377	**.000**	-3.498	-1.298
		2	1	.440	.377	.927	-.660	1.541
			3	-.835	.304	.082	-1.730	.060
			4	-.394	.298	.863	-1.271	.483
			5	-1,957*	.319	**.000**	-2.897	-1.018
		3	1	1,275*	.365	**.009**	.220	2.331
			2	.835	.304	.082	-.060	1.730
			4	.441	.282	.713	-.391	1.274
			5	-1,122*	.304	.006	-2.017	-.227
		4	1	.834	.360	.195	-.203	1.872
			2	.394	.298	.863	-.483	1.271
			3	-.441	.282	.713	-1.274	.391
			5	-1,563*	.298	**.000**	-2.440	-.687
		5	1	2,398*	.377	**.000**	1.298	3.498
			2	1,957*	.319	**.000**	1.018	2.897
			3	1,122*	.304	**.006**	.227	2.017
			4	1,563*	.298	**.000**	.687	2.440
	Hochberg	1	2	-.440	.377	.931	-1.553	.672
			3	-1,275*	.365	.011	-2.351	-.201
			4	-.834	.360	.218	-1.896	.227
			5	-2,398*	.377	**.000**	-3.510	-1.286
		2	1	.440	.377	.931	-.672	1.553
			3	-.835	.304	.082	-1.732	.061
			4	-.394	.298	.865	-1.273	.485
			5	-1,957*	.319	**.000**	-2.897	-1.018
		3	1	1,275*	.365	**.011**	.201	2.351
			2	.835	.304	.082	-.061	1.732
			4	.441	.282	.713	-.391	1.274
			5	-1,122*	.304	**.006**	-2.018	-.226

		4	1	.834	.360	.218	-.227	1.896
			2	.394	.298	.865	-.485	1.273
			3	-.441	.282	.713	-1.274	.391
			5	-1,563*	.298	**.000**	-2.443	-.684
		5	1	2,398*	.377	**.000**	1.286	3.510
			2	1,957*	.319	**.000**	1.018	2.897
			3	1,122*	.304	**.006**	.226	2.018
			4	1,563*	.298	**.000**	.684	2.443
Score Z(UP_phw)	Gabriel	1	2	-.208	.310	.999	-1.112	.695
			3	1,878*	.299	**.000**	1.012	2.744
			4	,948*	.295	**.020**	.097	1.800
			5	1,886*	.310	**.000**	.984	2.790
		2	1	.208	.310	.999	-.695	1.112
			3	2,086*	.249	**.000**	1.352	2.821
			4	1,156*	.245	**.000**	.437	1.876
			5	2,095*	.262	**.000**	1.324	2.866
		3	1	-1,878*	.299	**.000**	-2.744	-1.012
			2	-2,086*	.249	**.000**	-2.821	-1.352
			4	-,929*	.232	**.002**	-1.613	-.247
			5	.009	.249	1.000	-.726	.743
		4	1	-,948*	.295	**.020**	-1.800	-.097
			2	-1,156*	.245	**.000**	-1.876	-.437
			3	,929*	.232	**.002**	.247	1.613
			5	,938*	.245	**.004**	.219	1.658
		5	1	-1,886*	.310	**.000**	-2.790	-.984
			2	-2,095*	.262	**.000**	-2.866	-1.324
			3	-.009	.249	1.000	-.743	.726
			4	-,938*	.245	**.004**	-1.658	-.219
	Hochberg	1	2	-.208	.310	.999	-1.121	.704
			3	1,878*	.299	**.000**	.996	2.761
			4	,948*	.295	**.025**	.077	1.819
			5	1,886*	.310	**.000**	.974	2.799
		2	1	.208	.310	.999	-.704	1.121
			3	2,086*	.249	**.000**	1.351	2.822
			4	1,156*	.245	**.000**	.435	1.878
			5	2,095*	.262	**.000**	1.324	2.866
		3	1	-1,878*	.299	**.000**	-2.761	-.996
			2	-2,086*	.249	**.000**	-2.822	-1.351
			4	-,929*	.232	**.002**	-1.613	-.247
			5	.009	.249	1.000	-.727	.744
		4	1	-,948*	.295	**.025**	-1.819	-.077
			2	-1,156*	.245	**.000**	-1.878	-.435
			3	,929*	.232	**.002**	.247	1.613
			5	,938*	.245	**.004**	.217	1.660
		5	1	-1,886*	.310	**.000**	-2.799	-.974

			2	-2,095*	.262	**.000**	-2.866	-1.324
			3	-.009	.249	1.000	-.744	.727
			4	-,938*	.245	**.004**	-1.660	-.217
*. La différence moyenne est significative au niveau 0.05.								

Table 9-25: Pairwise comparisons with Gabriel's procedure and Hochberg GT2

Table 9-26 shows the results of pairwise comparisons using the Games-Howell procedure.

Comparaisons multiples :							
Games-Howell							
Variable dépendante			Différence moyenne (I-J)	Erreur standard	Sig.	Intervalle de confiance à 95 %	
						Borne inférieure	Borne supérieure
Score Z(Mean_UP_length_sec)	1	2	.471	.611	.930	-1.862	2.805
		3	.757	.571	.692	-1.657	3.172
		4	-.140	.622	.999	-2.460	2.180
		5	1.664	.570	.162	-.755	4.084
	2	1	-.471	.611	.930	-2.805	1.862
		3	.286	.282	.845	-.614	1.186
		4	-.611	.375	.496	-1.738	.516
		5	1,193*	.280	**.008**	.296	2.090
	3	1	-.757	.571	.692	-3.172	1.657
		2	-.286	.282	.845	-1.186	.614
		4	-.897	.305	.065	-1.837	.043
		5	,907*	.176	**.001**	.374	1.441
	4	1	.140	.622	.999	-2.180	2.460
		2	.611	.375	.496	-.516	1.738
		3	.897	.305	.065	-.043	1.837
		5	1,804*	.303	**.000**	.869	2.740
	5	1	-1.664	.570	.162	-4.084	.755
		2	-1,193*	.280	**.008**	-2.090	-.296
		3	-,907*	.176	**.001**	-1.441	-.374
		4	-1,804*	.303	**.000**	-2.740	-.869
Score Z(FP_phw)	1	2	1.889	.612	.124	-.565	4.343
		3	2,621*	.595	**.042**	.125	5.117
		4	2.395	.596	.058	-.098	4.887
		5	2,506*	.622	**.045**	.071	4.942
	2	1	-1.889	.612	.124	-4.343	.565
		3	.732	.248	.064	-.033	1.497
		4	.505	.251	.303	-.264	1.275
		5	.617	.306	.304	-.324	1.558
	3	1	-2,621*	.595	**.042**	-5.117	-.125
		2	-.732	.248	.064	-1.497	.033
		4	-.226	.207	.808	-.844	.391

		5	-.115	.272	.993	-.964	.735
	4	1	-2.395	.596	.058	-4.887	.098
		2	-.505	.251	.303	-1.275	.264
		3	.226	.207	.808	-.391	.844
		5	.112	.275	.994	-.742	.965
	5	1	-2,506*	.622	**.045**	-4.942	-.071
		2	-.617	.306	.304	-1.558	.324
		3	.115	.272	.993	-.735	.964
		4	-.112	.275	.994	-.965	.742
Score Z(Rep_phw)	1	2	1.835	.745	.246	-1.263	4.933
		3	2.132	.728	.166	-1.027	5.291
		4	1.924	.740	.218	-1.187	5.036
		5	1.348	.772	.482	-1.689	4.386
	2	1	-1.835	.745	.246	-4.933	1.263
		3	.297	.224	.682	-.410	1.004
		4	.089	.263	.997	-.706	.885
		5	-.487	.342	.623	-1.549	.575
	3	1	-2.132	.728	.166	-5.291	1.027
		2	-.297	.224	.682	-1.004	.410
		4	-.208	.209	.854	-.837	.422
		5	-.784	.302	.141	-1.769	.201
	4	1	-1.924	.740	.218	-5.036	1.187
		2	-.089	.263	.997	-.885	.706
		3	.208	.209	.854	-.422	.837
		5	-.576	.332	.444	-1.610	.457
	5	1	-1.348	.772	.482	-4.386	1.689
		2	.487	.342	.623	-.575	1.549
		3	.784	.302	.141	-.201	1.769
		4	.576	.332	.444	-.457	1.610
Score Z(T_phw)	1	2	.211	.266	.926	-.716	1.137
		3	.205	.307	.960	-.785	1.195
		4	.203	.307	.961	-.784	1.189
		5	-.096	.258	.995	-1.016	.824
	2	1	-.211	.266	.926	-1.137	.716
		3	-.006	.262	1.000	-.802	.790
		4	-.008	.262	1.000	-.799	.782
		5	-.307	.202	.565	-.925	.312
	3	1	-.205	.307	.960	-1.195	.785
		2	.006	.262	1.000	-.790	.802
		4	-.002	.303	1.000	-.906	.902
		5	-.300	.253	.758	-1.074	.473
	4	1	-.203	.307	.961	-1.189	.784
		2	.008	.262	1.000	-.782	.799
		3	.002	.303	1.000	-.902	.906
		5	-.298	.253	.764	-1.066	.470

	5	1	.096	.258	.995	-.824	1.016
		2	.307	.202	.565	-.312	.925
		3	.300	.253	.758	-.473	1.074
		4	.298	.253	.764	-.470	1.066
Score Z(W_phw)	1	2	-.281	.187	.587	-.927	.364
		3	-.207	.090	.220	-.502	.088
		4	-.780	.361	.262	-1.946	.386
		5	-.218	.103	.301	-.574	.138
	2	1	.281	.187	.587	-.364	.927
		3	.074	.207	.996	-.590	.738
		4	-.499	.406	.736	-1.742	.745
		5	.063	.213	.998	-.613	.740
	3	1	.207	.090	.220	-.088	.502
		2	-.074	.207	.996	-.738	.590
		4	-.573	.372	.556	-1.752	.606
		5	-.011	.137	1.000	-.427	.405
	4	1	.780	.361	.262	-.386	1.946
		2	.499	.406	.736	-.745	1.742
		3	.573	.372	.556	-.606	1.752
		5	.562	.375	.581	-.622	1.747
	5	1	.218	.103	.301	-.138	.574
		2	-.063	.213	.998	-.740	.613
		3	.011	.137	1.000	-.405	.427
		4	-.562	.375	.581	-1.747	.622
*. La différence moyenne est significative au niveau 0.05.							

Table 9-26: ANOVA post hoc test results: Pairwise comparisons with Games-Howell procedure

9.8.2.4 ANOVA results (4 (dis)fluency components)

Table 9-27 shows the results of Levene's test of homogeneity of variances. Table 9-28 displays the results of the ANOVA test.

Note: significant results are shown in bold font.

	Statistique de Levene	**ddl1**	**ddl2**	**Sig.**
Component 1 score (Anderson Rubin)	.561	4	41	.692
Component 2 score (Anderson Rubin)	1.018	4	41	.409
Component 3 score (Anderson Rubin)	1.702	4	41	.168
Component 4 score (Anderson Rubin)	1.641	4	41	.182

Table 9-27: Levene's test of homogeneity of variance

ANOVA						
		Somme des carrés	ddl	Carré moyen	F	Sig.
	Inter-groupes	38.105	4	9.526	42.074	**.000**

Component 1 score Anderson Rubin	Intragroupes	9.283	41	.226		
	Total	47.388	45			
Component 2 score Anderson Rubin	Inter-groupes	5.868	4	1.467	3.529	**.015**
	Intragroupes	17.045	41	.416		
	Total	22.913	45			
Component 3 score Anderson Rubin	Inter-groupes	20.268	4	5.067	7.884	**.000**
	Intragroupes	26.349	41	.643		
	Total	46.617	45			
Component 4 score Anderson Rubin	Inter-groupes	7.116	4	1.779	2.833	**.037**
	Intragroupes	25.749	41	.628		
	Total	32.866	45			

Table 9-28: ANOVA test results

Table 9-29 shows the results of pairwise comparisons using Gabriel and Hochberg's procedures.

Comparaisons multiples :								
Variable dépendante				Différence moyenne (I-J)	Erreur standard	Sig.	Intervalle de confiance à 95 %	
							Borne inférieure	Borne supérieure
Component 1 score Anderson Rubin	Gabriel	1	2	-.096	.265	1.000	-.870	.678
			3	-1.704*	.257	**.000**	-2.447	-.961
			4	-1.001*	.253	**.002**	-1.732	-.272
			5	-2.561*	.265	**.000**	-3.335	-1.787
		2	1	.096	.265	1.000	-.678	.870
			3	-1.608*	.214	**.000**	-2.238	-.978
			4	-.906*	.210	**.001**	-1.523	-.289
			5	-2.465*	.224	**.000**	-3.127	-1.804
		3	1	1.704*	.257	**.000**	.961	2.447
			2	1.608*	.214	**.000**	.978	2.238
			4	.702*	.199	**.010**	.117	1.288
			5	-.857*	.214	**.002**	-1.487	-.227
		4	1	1.001*	.253	**.002**	.272	1.732
			2	.906*	.210	**.001**	.289	1.523
			3	-.702*	.199	**.010**	-1.288	-.117
			5	-1.559*	.210	**.000**	-2.176	-.942
		5	1	2.561*	.265	**.000**	1.787	3.335
			2	2.465*	.224	**.000**	1.804	3.127
			3	.857*	.214	**.002**	.227	1.487
			4	1.559*	.210	**.000**	.942	2.176
	Hochberg	1	2	-.096	.265	1.000	-.878	.686
			3	-1.704*	.257	**.000**	-2.461	-.947
			4	-1.001*	.253	**.003**	-1.749	-.255
			5	-2.561*	.265	**.000**	-3.344	-1.779
		2	1	.096	.265	1.000	-.686	.878
			3	-1.608*	.214	**.000**	-2.239	-.978

			4	-.906*	.210	**.001**	-1.525	-.287
			5	-2.465*	.224	**.000**	-3.127	-1.804
		3	1	1.704*	.257	**.000**	.947	2.461
			2	1.608*	.214	**.000**	.978	2.239
			4	.702*	.199	**.010**	.117	1.288
			5	-.857*	.214	**.003**	-1.488	-.227
		4	1	1.001*	.253	**.003**	.255	1.749
			2	.906*	.210	**.001**	.287	1.525
			3	-.702*	.199	**.010**	-1.288	-.117
			5	-1.559*	.210	**.000**	-2.178	-.941
		5	1	2.561*	.265	**.000**	1.779	3.344
			2	2.465*	.224	**.000**	1.804	3.127
			3	.857*	.214	**.003**	.227	1.488
			4	1.559*	.210	**.000**	.941	2.178
Component 2 score Anderson Rubin	Gabriel	1	2	.649	.360	.519	-.400	1.699
			3	.704	.348	.356	-.303	1.710
			4	.441	.343	.869	-.548	1.430
			5	-.226	.360	.999	-1.275	.823
		2	1	-.649	.360	.519	-1.699	.400
			3	.054	.290	1.000	-.799	.908
			4	-.209	.284	.997	-1.045	.628
			5	-.875	.304	.059	-1.771	.021
		3	1	-.704	.348	.356	-1.710	.303
			2	-.054	.290	1.000	-.908	.799
			4	-.263	.269	.978	-1.056	.530
			5	-.929*	.290	**.025**	-1.783	-.076
		4	1	-.441	.343	.869	-1.430	.548
			2	.209	.284	.997	-.628	1.045
			3	.263	.269	.978	-.530	1.056
			5	-.667	.284	.203	-1.503	.169
		5	1	.226	.360	.999	-.823	1.275
			2	.875	.304	.059	-.021	1.771
			3	.929*	.290	**.025**	.076	1.783
			4	.667	.284	.203	-.169	1.503
	Hochberg	1	2	.649	.360	.533	-.411	1.710
			3	.704	.348	.380	-.322	1.729
			4	.441	.343	.883	-.571	1.453
			5	-.226	.360	.999	-1.286	.834
		2	1	-.649	.360	.533	-1.710	.411
			3	.054	.290	1.000	-.800	.909
			4	-.209	.284	.997	-1.047	.630
			5	-.875	.304	.059	-1.771	.021
		3	1	-.704	.348	.380	-1.729	.322
			2	-.054	.290	1.000	-.909	.800
			4	-.263	.269	.978	-1.056	.531

			5	-.929*	.290	**.025**	-1.784	-.075
		4	1	-.441	.343	.883	-1.453	.571
			2	.209	.284	.997	-.630	1.047
			3	.263	.269	.978	-.531	1.056
			5	-.667	.284	.206	-1.505	.171
		5	1	.226	.360	.999	-.834	1.286
			2	.875	.304	.059	-.021	1.771
			3	.929*	.290	**.025**	.075	1.784
			4	.667	.284	.206	-.171	1.505
Component 3 score Anderson Rubin	Gabriel	1	2	-2.235*	.447	**.000**	-3.540	-.931
			3	-2.140*	.432	**.000**	-3.392	-.889
			4	-1.410*	.427	**.016**	-2.640	-.181
			5	-1.454*	.447	**.020**	-2.759	-.150
		2	1	2.235*	.447	**.000**	.931	3.540
			3	.094	.360	1.000	-.967	1.155
			4	.825	.353	.209	-.215	1.864
			5	.781	.378	.353	-.333	1.895
		3	1	2.140*	.432	**.000**	.889	3.392
			2	-.094	.360	1.000	-1.155	.967
			4	.730	.335	.285	-.256	1.717
			5	.686	.360	.459	-.375	1.747
		4	1	1.410*	.427	**.016**	.181	2.640
			2	-.825	.353	.209	-1.864	.215
			3	-.730	.335	.285	-1.717	.256
			5	-.044	.353	1.000	-1.083	.996
		5	1	1.454*	.447	**.020**	.150	2.759
			2	-.781	.378	.353	-1.895	.333
			3	-.686	.360	.459	-1.747	.375
			4	.044	.353	1.000	-.996	1.083
	Hochberg	1	2	-2.235*	.447	**.000**	-3.553	-.917
			3	-2.140*	.432	**.000**	-3.415	-.866
			4	-1.410*	.427	**.019**	-2.668	-.152
			5	-1.454*	.447	**.022**	-2.772	-.136
		2	1	2.235*	.447	**.000**	.917	3.553
			3	.094	.360	1.000	-.968	1.157
			4	.825	.353	.211	-.217	1.867
			5	.781	.378	.353	-.333	1.895
		3	1	2.140*	.432	**.000**	.866	3.415
			2	-.094	.360	1.000	-1.157	.968
			4	.730	.335	.285	-.256	1.717
			5	.686	.360	.461	-.376	1.749
		4	1	1.410*	.427	**.019**	.152	2.668
			2	-.825	.353	.211	-1.867	.217
			3	-.730	.335	.285	-1.717	.256
			5	-.044	.353	1.000	-1.086	.998

		5	1	1.454*	.447	**.022**	.136	2.772
			2	-.781	.378	.353	-1.895	.333
			3	-.686	.360	.461	-1.749	.376
			4	.044	.353	1.000	-.998	1.086
Component 4 score Anderson Rubin	Gabriel	1	2	-.102	.442	1.000	-1.391	1.188
			3	-.262	.427	.999	-1.499	.975
			4	.764	.422	.499	-.452	1.980
			5	.064	.442	1.000	-1.225	1.354
		2	1	.102	.442	1.000	-1.188	1.391
			3	-.160	.356	1.000	-1.209	.888
			4	.865	.349	.153	-.162	1.893
			5	.166	.374	1.000	-.936	1.267
		3	1	.262	.427	.999	-.975	1.499
			2	.160	.356	1.000	-.888	1.209
			4	1.025*	.331	**.033**	.051	2.001
			5	.326	.356	.986	-.723	1.375
		4	1	-.764	.422	.499	-1.980	.452
			2	-.865	.349	.153	-1.893	.162
			3	-1.025*	.331	**.033**	-2.001	-.051
			5	-.700	.349	.391	-1.727	.328
		5	1	-.064	.442	1.000	-1.354	1.225
			2	-.166	.374	1.000	-1.267	.936
			3	-.326	.356	.986	-1.375	.723
			4	.700	.349	.391	-.328	1.727
	Hochberg	1	2	-.102	.442	1.000	-1.405	1.202
			3	-.262	.427	.999	-1.522	.998
			4	.764	.422	.529	-.480	2.007
			5	.064	.442	1.000	-1.239	1.367
		2	1	.102	.442	1.000	-1.202	1.405
			3	-.160	.356	1.000	-1.211	.890
			4	.865	.349	.155	-.165	1.896
			5	.166	.374	1.000	-.936	1.267
		3	1	.262	.427	.999	-.998	1.522
			2	.160	.356	1.000	-.890	1.211
			4	1.025*	.331	**.034**	.050	2.001
			5	.326	.356	.986	-.724	1.376
		4	1	-.764	.422	.529	-2.007	.480
			2	-.865	.349	.155	-1.896	.165
			3	-1.025*	.331	**.034**	-2.001	-.050
			5	-.700	.349	.394	-1.730	.331
		5	1	-.064	.442	1.000	-1.367	1.239
			2	-.166	.374	1.000	-1.267	.936
			3	-.326	.356	.986	-1.376	.724
			4	.700	.349	.394	-.331	1.730

*. La différence moyenne est significative au niveau 0.05.

Table 9-29: Pairwise comparisons with Gabriel's procedure and Hochberg's GT2

9.9 POST-HOC TEST RESULTS FOR REPEATED MEASURES ONE-WAY ANOVAS

Table 9-30 displays the results of the post-hoc tests (pairwise comparisons with Bonferroni correction) for the repeated measures one-way ANOVAs comparing the 14 (dis)fluency measures in the CEFR rated excerpt, the free discussion and the interview.

		Mean difference	Standard error	Significance	95% confidence interval	
					Upper bound	Lower bound
Conjunctions						
1	2	,181	,180	,957	-,265	,627
	3	-,384	,205	,199	-,892	,123
2	1	-,181	,180	,957	-,627	,265
	3	-,565*	,103	**,000**	-,820	-,311
3	1	,384	,205	,199	-,123	,892
	2	,565*	,103	**,000**	,311	,820
Discourse markers						
1	2	,006	,069	1,000	-,164	,177
	3	,198	,092	,108	-,030	,425
2	1	-,006	,069	1,000	-,177	,164
	3	,191	,078	,054	-,003	,385
3	1	-,198	,092	,108	-,425	,030
	2	-,191	,078	,054	-,385	,003
Unfilled pauses						
1	2	,527*	,129	**,000**	,209	,846
	3	-,330	,202	,324	-,829	,170
2	1	-,527*	,129	**,000**	-,846	-,209
	3	-,857*	,179	**,000**	-1,301	-,414
3	1	,330	,202	,324	-,170	,829
	2	,857*	,179	**,000**	,414	1,301
Speech rate						
1	2	-,825	,626	,582	-2,377	,728
	3	6,008*	1,147	**,000**	3,165	8,851
2	1	,825	,626	,582	-,728	2,377
	3	6,833*	,995	**,000**	4,366	9,299
3	1	-6,008*	1,147	**,000**	-8,851	-3,165
	2	-6,833*	,995	**,000**	-9,299	-4,366
Mean length of runs						
1	2	,143*	,047	**,011**	,026	,260
	3	-,087	,070	,668	-,261	,087
2	1	-,143*	,047	**,011**	-,260	-,026
	3	-,230*	,064	**,002**	-,388	-,072
3	1	,087	,070	,668	-,087	,261

	2	,230*	,064	**,002**	,072	,388
Mean length of unfilled pauses						
1	2	-,011	,005	,054	-,023	,000
	3	-,018	,008	,061	-,037	,001
2	1	,011	,005	,054	,000	,023
	3	-,007	,008	1,000	-,025	,012
3	1	,018	,008	,061	-,001	,037
	2	,007	,008	1,000	-,012	,025
Phonation time ratio						
1	2	-,270	,175	,386	-,704	,163
	3	,530	,328	,337	-,282	1,343
2	1	,270	,175	,386	-,163	,704
	3	,800*	,257	**,009**	,163	1,438
3	1	-,530	,328	,337	-1,343	,282
	2	-,800*	,257	**,009**	-1,438	-,163

Table 9-30: Results of ANOVA post-hoc tests with Bonferroni correction

Notes: (1) 1 = rated excerpt; 2 = free discussion task; 3 = interview; (2) significant results are shown in bold font.

9.10 B2 AND C1 LEARNERS

9.10.1 Descriptive statistics of B2 and C1 learners in the rated excerpt and in the interview

Table 9-31 shows the means for B2 and C1 learners for each (dis)fluency variable and component in the rated excerpt ("CEFR") and the whole interview ("int").

(Dis)fluency measures		N	Mean	Std. deviation
Speech rate (wpm) - int	B2	22	156.37	12.64
	C1	26	166.84	16.52
Speech rate (wpm) - CEFR	B2	22	161.28	15.35
	C1	26	173.03	16.83
Mean UP length (sec) - int	B2	22	0.53	0.09
	C1	26	0.49	0.09
Mean UP length (sec) - CEFR	B2	22	0.51	0.11
	C1	26	0.47	0.09
Mean length of runs - int	B2	22	5.40	0.91
	C1	26	5.79	1.11
Mean length of runs - CEFR	B2	22	5.23	1.01
	C1	26	5.76	1.20
PTR - int	B2	22	81.46	4.68
	C1	26	83.93	4.26
PTR - CEFR	B2	22	81.76	6.06
	C1	26	84.64	4.24
C (phw) - int	B2	22	5.37	1.53
	C1	26	4.82	0.91
C (phw) - CEFR	B2	22	4.69	1.46
	C1	26	4.36	1.27
DM (phw) - int	B2	22	1.43	1.12
	C1	26	2.55	1.57
DM (phw) - CEFR	B2	22	1.61	1.13
	C1	26	2.66	1.73
FP (phw) - int	B2	22	8.25	2.57
	C1	26	7.64	2.96
FP (phw) - CEFR	B2	22	8.13	2.80
	C1	26	7.50	3.13
FS (phw) - int	B2	22	0.68	0.37
	C1	26	0.70	0.28
FS (phw) - CEFR	B2	22	0.78	0.52
	C1	26	0.60	0.40
L (phw) - int	B2	22	2.93	0.89

	C1	26	3.36	1.07
L (phw) - CEFR	B2	22	2.95	0.84
	C1	26	2.92	1.24
Rep (phw) - int	B2	22	4.02	1.21
	C1	26	3.87	1.60
Rep (phw) - CEFR	B2	22	4.00	1.49
	C1	26	3.76	1.79
RS (phw) - int	B2	22	2.02	0.61
	C1	26	1.71	0.62
RS (phw) - CEFR	B2	22	2.01	0.87
	C1	26	1.67	0.71
T (phw) - int	B2	22	1.76	0.73
	C1	26	1.57	0.66
T (phw) - CEFR	B2	22	1.71	0.77
	C1	26	1.42	0.68
UP (phw) - int	B2	22	13.64	2.82
	C1	26	11.98	2.62
UP (phw) - CEFR	B2	22	13.45	3.84
	C1	26	11.52	2.72
W (phw) - int	B2	22	0.55	0.48
	C1	26	0.44	0.58
W (phw) - CEFR	B2	22	0.63	0.60
	C1	26	0.47	0.77
Component 1	B2	22	-0.34	0.94
	C1	26	0.26	1.02
Component 2	B2	22	0.16	0.86
	C1	26	-0.10	1.13
Component 3	B2	22	-0.38	0.89
	C1	26	0.25	1.02
Component 4	B2	22	0.15	1.27
	C1	26	-0.20	0.65
Component 5	B2	22	0.11	0.93
	C1	26	-0.09	1.08

Table 9-31: Descriptive statistics of B2 and C1 learners in the rated excerpt ("CEFR") and in the whole interview ("int")

9.10.2 Levene's tests and t-tests for comparisons of means

Table 9-32 and 9-33 show the results of Levene's test of homogeneity of variances between the group of B2 and C1 learners for the 14 (dis)fluency variables and the 5 (dis)fluency components, respectively.

(Dis)fluency variables	**Levene's test of equality of variances**	**T-test**

Conjunctions (phw)	F = 3.323; *p* = .075	T = 1.533; *p* = .132
Discourse markers (phw)	F = 1.857; *p* = .180	**T = -2.792; *p* = .008; *d* = 0.820**
False starts (phw)	F = .319; *p* = .575	T = -.111; *p* = .912
Filled pauses (phw)	F = 1.350; *p* = .251	T = .754; *p* = .455
Foreign words (phw)	F = .006; *p* = .940	T = .683; *p* = .498
Lengthenings (phw)	F = .240; *p* = .626	T = -1.489; *p* = .143
Mean length of runs	F = .939; *p* = .338	T = -1.341; *p* = .187
Mean UP length (sec)	F = .041; *p* = .841	T = 1.505; *p* = .139
Phonation-time ratio	F = .273; *p* = .604	T = -1.918; *p* = .061
Repetitions (phw)	F = 2.052; *p* = .159	T = .365; *p* = .716
Restarts (phw)	F = .122; *p* = .729	T = 1.745; *p* = .088
Speech rate[179] (wpm)	F = 2.373; *p* = .130	**T = -2.432; *p* = .019; *d* = 0.712**
Truncations (phw)	F = .113; *p* = .738	T = .924; *p* = .360
Unfilled pauses (phw)	F = .086; *p* = .771	**T = 2.117; *p* = .040; *d* = 0.611**

Table 9-32: Independent-samples t-test results for the 14 (dis)fluency variables in B2 and C1 learner speech

(Dis)fluency variables	Levene's test of equality of variances	T-test
Comp. 1 – temporal (dis)fluency	F = .426; *p* = .517	**T = -2.082; *p* = .043; *d* = 0.605**
Comp. 2 – repair (dis)fluency	F = .556; *p* = .460	T = .853; *p* = .398
Comp. 3 – pragmatic (dis)fluency	F = .040; *p* = .841	**T = -2.265; *p* = .028; *d* = 0.660**
Comp. 4 – cohesion	**F = 4.751; *p* = .034**	T = 1.164; *p* = .254
Comp. 5 – lexico-grammatical (dis)fluency	F = .171; *p* = .681	T = .678; *p* = .501

Table 9-33: Independent-samples t-test results for the 5 (dis)fluency components in B2 and C1 learner speech

[179] In the rated excerpt: T = -2.509; *p* = .016 (non-significant Levene's test).

9.11 CORRELATION ANALYSIS BETWEEN (DIS)FLUENCY MEASURES AND CEFR FLUENCY RATINGS

Figure 9-16 to Figure 9-24 show the correlations between (dis)fluency measures and CEFR fluency ratings (i.e. not significant correlations).

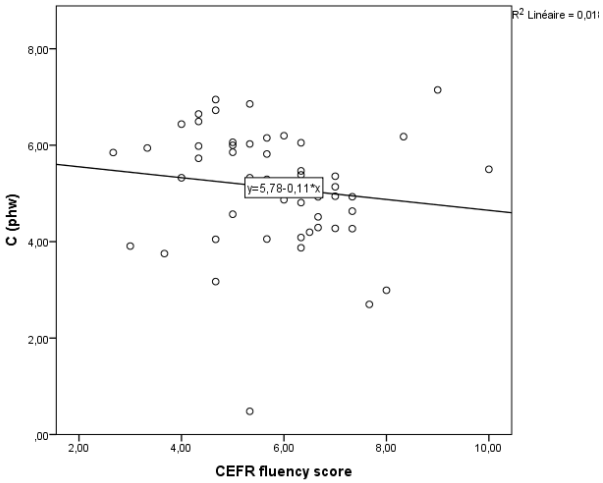


Figure 9-16: The relationship between connectors and CEFR fluency score

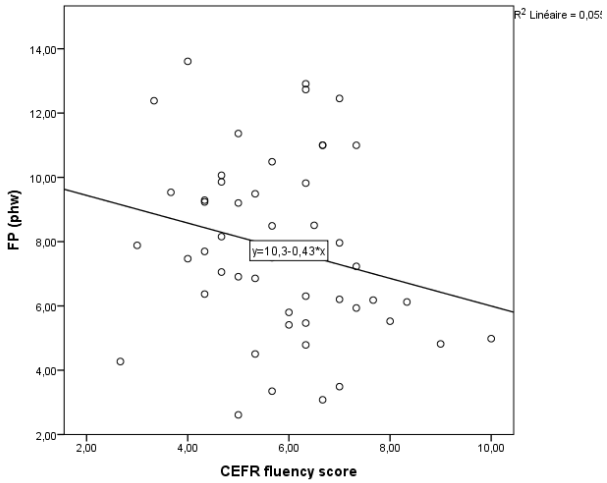


Figure 9-17: The relationship between filled pauses and CEFR fluency score

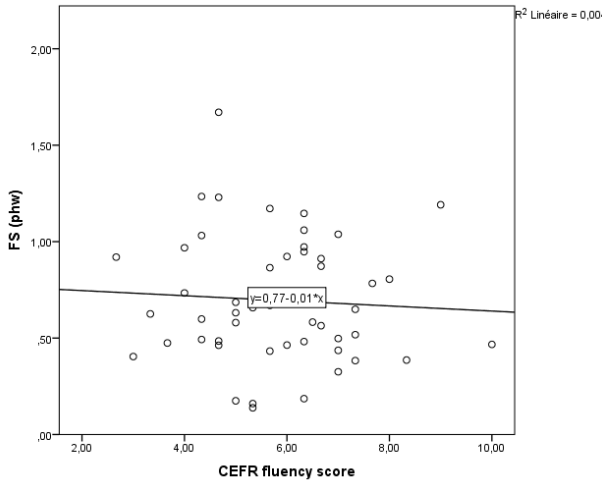


Figure 9-18: The relationship between false starts and CEF fluency score

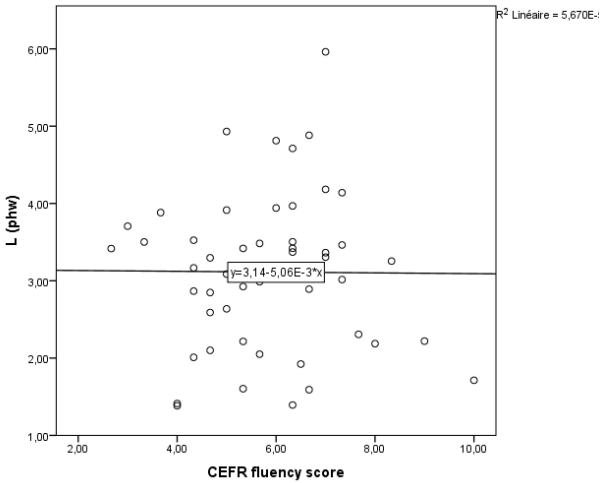


Figure 9-19: The relationship between lengthenings and CEFR fluency score

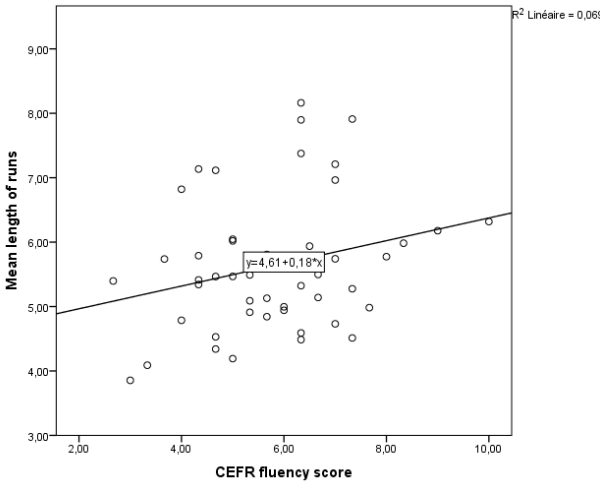


Figure 9-20: The relationship between mean length of runs and CEFR fluency score

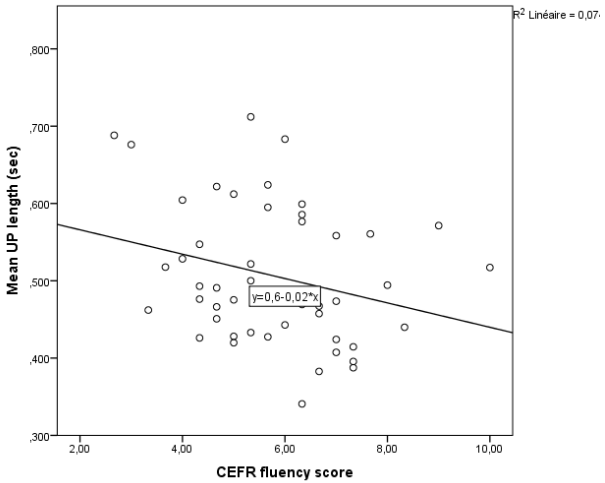


Figure 9-21: The relationship between mean length of unfilled pauses and CEFR fluency score

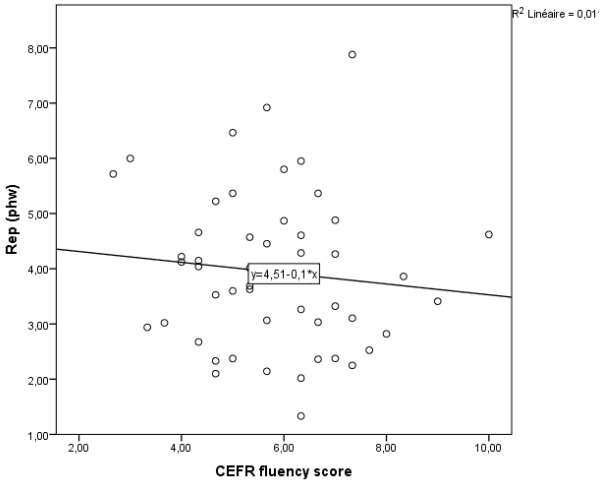


Figure 9-22: The relationship between repetitions and CEFR fluency score

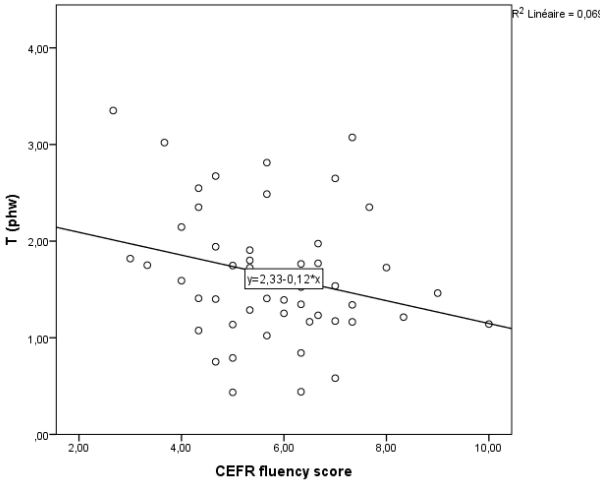


Figure 9-23: The relationship between truncations and CEFR fluency score

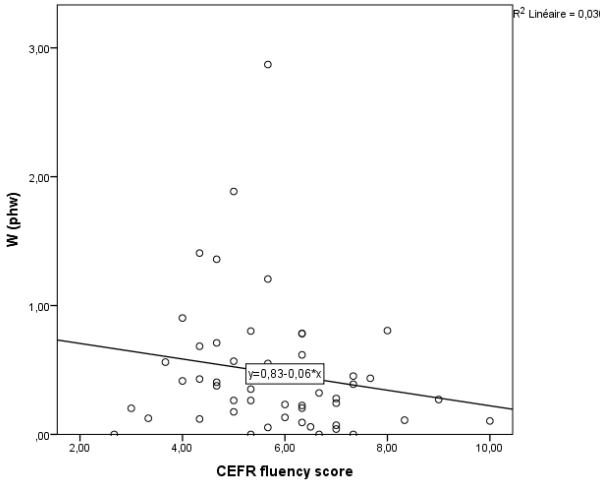


Figure 9-24: The relationship between foreign words and CEFR fluency score

9.12 MULTIPLE LINEAR REGRESSION

Table 9-34 shows the summary of the regression model. Table 9-35 displays the ANOVA test results.

Model	R	R^2	Adjusted R^2	Std. Error of the estimate	Change statistics					Durbin-Watson
					R^2 change	F Change	df1	df2	Sig. F change	
1	,478[a]	.228	.212	1.34181	.228	14.197	1	48	.000	
2	,577[b]	.333	.305	1.26071	.105	7.374	1	47	.009	2.253
a. Predictors: (Constant), RS*UP										
b. Predictors: (Constant), RS*UP, DM*SR										
c. Dependent variable: CEFR_fluency_score										

Table 9-34: Summary of the models

Model		Sum of squares	df	Mean square	F	Sig.
1	Regression	25.561	1	25.561	14.197	,000[b]
	Residual	86.422	48	1.800		
	Total	111.983	49			
2	Regression	37.282	2	18.641	11.728	,000[c]
	Residual	74.701	47	1.589		
	Total	111.983	49			
a. Dependent Variable: CEFR_fluency_score						
b. Predictors : (Constant), RS*UP						
c. Predictors: (Constant), RS*UP, DM*SR						

Table 9-35: ANOVA test results

Table 9-36 shows the model parameters, and the collinearity statistics are displayed in Table 9-37.

Model		Unstandardized Coefficients		Std. Coefficients	t	Sig.	95% confidence interval for B		Correlations		
		B	Std. Error	Beta			Lower bound	Upper bound	Zero-order	Partial	Part
1	(Constant)	7.471	.482		15.515	.000	6.503	8.439			
	RS*UP	-.071	.019	-.478	-3.768	.000	-.109	-.033	-.478	-.478	-.478
2	(Constant)	6.739	.527		12.799	.000	5.680	7.799			

	RS*UP	-.067	.018	-.451	-3.776	.000	-.103	-.031	-.478	-.482	-.450
	DM*SR	.002	.001	.325	2.716	.009	.000	.003	.361	.368	.324
a. Dependent Variable: CEFR_fluency_score											

Table 9-36: Model parameters

Model		Eigenvalue	Condition index	Variance proportions		
				(Constant)	RS*UP	DM*SR
1	1	1.919	1.000	.04	.04	
	2	.081	4.870	.96	.96	
2	1	2.621	1.000	.02	.02	.04
	2	.309	2.914	.02	.14	.78
	3	.070	6.130	.96	.84	.18
a. Dependent Variable: CEFR_fluency_score						

Table 9-37: Collinearity diagnostics

The casewise summary is displayed in Table 9-38.

	Mahalanobis Distance	Cook's Distance	Centered Leverage Value	Standardized DFFIT	Standardized DFBETA Intercept	Standardized DFBETA DM*SR	Standardized DFBETA RS*UP
1	.17765	.00701	.00363	-.14486	-.05448	-.04149	.03521
2	10.57141	.06667	**.21574**	.44554	-.22113	-.11526	.39960
3	.84175	.05022	.01718	-.40101	-.03792	-.25770	.06758
4	.09378	.05587	.00191	.44172	.17242	-.12341	.03237
5	3.61273	.00565	.07373	.12898	-.04133	-.04325	.10203
6	1.26696	.07586	.02586	-.49767	-.33401	.37140	.07151
7	1.83129	.00085	.03737	.05009	.04765	-.01686	-.03799
8	9.30257	.00012	.18985	-.01883	.01465	-.00712	-.01696
9	.92556	.00401	.01889	.10886	.09521	-.02265	-.07401
10	.76131	.00224	.01554	-.08125	-.05694	.05236	.01626
11	2.86027	.00220	.05837	.08035	-.01460	-.03604	.05611
12	.43071	.00991	.00879	.17247	.08698	-.09445	.00496
13	.80799	.00776	.01649	-.15196	-.08240	.10168	-.00146
14	.49935	.02137	.01019	.25618	.16599	-.14658	-.03768
15	.47012	.00359	.00959	.10311	.06887	-.05703	-.01851
16	.24215	.00006	.00494	.01281	.00600	.00318	-.00446
17	1.19363	.00431	.02436	-.11277	-.07223	.08342	.01173
18	.19819	.00077	.00404	-.04771	-.00615	-.01949	.00016
19	.54674	.00047	.01116	-.03698	-.03039	.01565	.01687
20	.58866	.00200	.01201	.07682	.06137	-.03891	-.02955
21	3.19768	.02998	.06526	.29980	-.11353	.25954	.05888
22	.18364	.00000	.00375	.00297	.00203	-.00056	-.00108

23		.09858	.00917	.00201	-.16628	-.08855	.04865	.01658
24		3.46527	.03042	.07072	-.30180	-.29859	.13784	.23851
25		.70385	.00026	.01436	.02775	-.00342	.01785	.00329
26		.53466	.00031	.01091	-.03024	-.02441	.00496	.01762
27		2.72263	.00431	.05556	-.11272	-.02308	-.07907	.04897
28		3.73240	.05624	.07617	.41338	-.25322	.30715	.22682
29		2.44905	.02506	.04998	.27420	.18559	.07349	-.21306
30		.27174	.00122	.00555	-.06003	.00634	-.02301	-.01772
31		8.96356	.03797	**.18293**	-.33548	.06068	-.30881	.05267
32		.60954	.02466	.01244	.27562	.23214	-.07081	-.16054
33		.47449	.00041	.00968	-.03462	.00898	-.01113	-.01719
34		1.49743	.11909	.03056	.63867	.44264	.13724	-.46444
35		.59874	.01928	.01222	.24246	.12751	-.14839	.00444
36		1.80364	.00211	.03681	-.07871	-.01778	-.05201	.03183
37		.26819	.00589	.00547	-.13249	-.02550	.04172	-.04153
38		2.07064	.00002	.04226	-.00826	.00449	-.00532	-.00466
39		1.83070	.01312	.03736	-.19765	-.01231	.11593	-.09977
40		11.73176	.00609	**.23942**	-.13376	.06208	-.12797	-.02201
41		1.47138	.05574	.03003	-.41894	.20846	-.11334	-.31235
42		1.33280	.00071	.02720	-.04579	-.03888	.00279	.03476
43		1.41063	.03065	.02879	-.30588	-.28598	.10457	.21822
44		1.00548	.02070	.02052	.25046	.21634	-.14374	-.11672
45		.26249	.00456	.00536	-.11635	-.08452	.03653	.04190
46		1.11727	.00021	.02280	-.02464	-.02115	.00275	.01794
47		5.11274	.00938	.10434	-.16627	.05445	.06328	-.13294
48		.18656	.00010	.00381	.01702	-.00058	.00110	.00678
49		1.00979	.00543	.02061	-.12676	-.10687	.07686	.05349
50		.65987	.00005	.01347	.01160	.00499	-.00705	.00152
Total	N	50	50	50	50	50	50	50

Table 9-38: Casewise diagnostics for the multiple linear regression

The examination of casewise diagnostics in Table 9-38 shows no cause for concern. No case has a Cook's distance greater than 1, which means that none of the cases is having an undue influence on the model. The average leverage values of three cases cf. bold font) could raise some concern, but given the Mahalanobis distance associated with those cases, there is probably little cause for alarm. Based on those indices, the model appears to be fairly reliable, and not unduly influenced by any case.

Figure 9-25, which plots the standardized residuals against the standardized predicted values of the dependant variable, indicates that the assumptions of linearity and homoscedasticity have been met as this graph contains randomly and evenly dispersed dots. Lastly, the histogram (Figure 9-26) shows that for our data, the distribution is roughly normal (although there is a slight deficiency of residuals around 2). The dots in the P-P plot (Figure 9-27) do not lie as close to the diagonal as ideal, but seem to tend towards the normality of residuals.

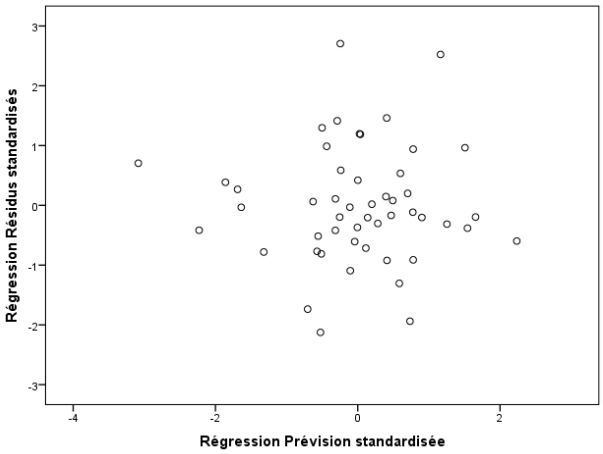


Figure 9-25: Scatterplot of standardized residuals against standardized predicted values of the dependant variable CEFR fluency ratings

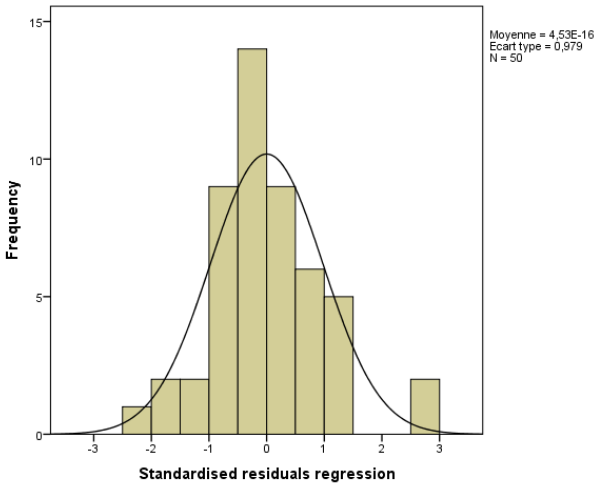


Figure 9-26: Histogram of standardized residuals

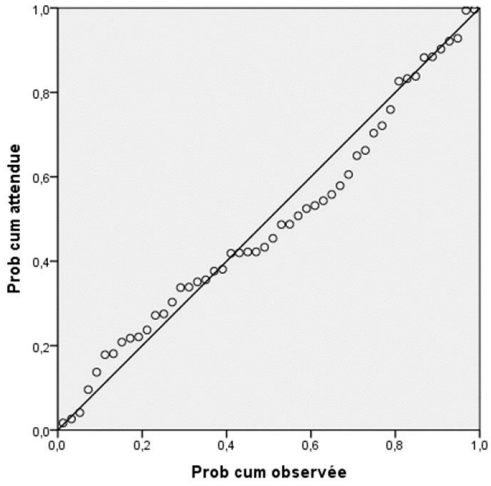


Figure 9-27: P-P plot of standardized residuals

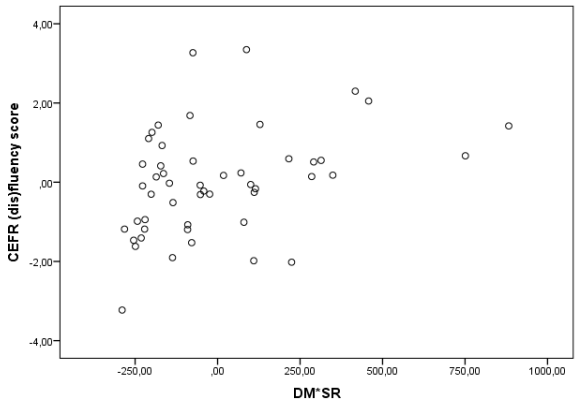


*Figure 9-28: Partial regression of DM*SR and CEFR fluency score*

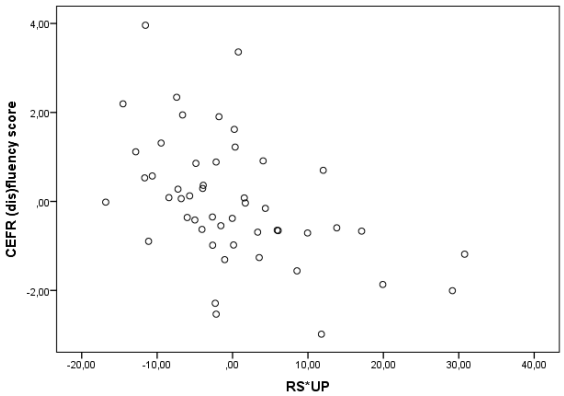


*Figure 9-29: Partial regression of RS*UP and CEFR fluency score*