"Errors and disfluencies in spoken corpora: Setting the scene"

Gilquin, Gaëtanelle ; De Cock, Sylvie

ABSTRACT

In this introduction, we wish to provide a broad overview of errors and disfluencies, showing how they are defined – and distinguished from each other – in the literature, and what impact the corpus revolution has had on the study of these phenomena. We will also demonstrate the usefulness of investigating such items by examining some of the possible applications of the study of errors and disfluencies.

CITE THIS VERSION


DIAL is an institutional repository for the deposit and dissemination of scientific documents from UCLouvain members. Usage of this document for profit or commercial purposes is strictly prohibited. User agrees to respect copyright about this document, mainly text integrity and source mention. Full content of copyright policy is available at Copyright policy.
Introduction

Errors and disfluencies in spoken corpora: Setting the scene

Gaëtanelle Gilquin and Sylvie De Cock
FNRS – Université catholique de Louvain / Université catholique de Louvain

1. Introduction

In January 2008, the Centre for English Corpus Linguistics (Université catholique de Louvain, Belgium) organised a colloquium on spoken learner corpora open to all the partners of the LINDSEI (Louvain International Database of Spoken English Interlanguage) project, a project aimed at collecting informal interviews with higher intermediate to advanced learners of English from different mother tongue backgrounds (cf. Gilquin et al. 2010). The aim of this two-day colloquium was to discuss a number of important issues that had arisen in the course of the project and also to think about possible future developments of the project. One of the questions that was brought up during the discussions had to do with the annotation of errors and disfluencies in the LINDSEI transcripts. It soon became clear that this question could not be resolved so easily and that more reflection was necessary on the theoretical relevance and practical feasibility of this type of annotation. In an attempt to explore the issue more thoroughly, a pre-conference workshop was organised at the 30th ICAME conference at Lancaster University on 27 May 2009. Next to the theoretical question of how a distinction can be drawn between errors and disfluencies, the presentations and ensuing discussions considered, among other topics, the practical problems linked to the markup and/or annotation of spoken language (especially learner speech) despite the presence of disfluencies and non-standard forms, the functions of hesitation markers and the pedagogical implications of corpus analyses of errors and disfluencies.
Several of the papers included in this special issue were presented at the ICAME workshop. The others are invited contributions from scholars who work in the field of errors and disfluencies. Together, these papers give a good illustration of how spoken corpora can be exploited to shed light on phenomena like fillers, silent pauses, speech rate or error rate, but also how these may be related to other phenomena such as syntactic complexity or perception of oral proficiency (see Section 5 for a more detailed description of the contents of each paper). In this introduction, we wish to provide a broad overview of errors and disfluencies, showing how they are defined – and distinguished from each other – in the literature (Section 2), and what impact the corpus revolution has had on the study of these phenomena (Section 3). We will also demonstrate the usefulness of investigating such items by examining some of the possible applications of the study of errors and disfluencies (Section 4).

2. Defining errors and disfluencies

Defining (and distinguishing) errors and disfluencies is not an easy task. Talking about the term ‘error’, Ringbom (1987:71) points out that it can be defined in many different ways. He adds that “[o]ne possible line of approach is to say that an error offends against the norm of the language”. The difficulty, however, is to define what the norm (or standard) is, despite the “vagueness inherent” in this concept: “[t]he norm may be fixed in some areas (e.g. spelling), but for lexis and grammar it may vary a great deal, showing, for instance, considerable differences in degree of tolerance between spoken and written language” (ibid.). This idea is also taken up by Carter & McCarthy (2006:168), who argue that “[w]hat may be considered ‘non-standard’ in writing may well be ‘standard’ in speech”. Sometimes, the notion of error is defined with reference to native language. Tenfjord et al. (2006:102) describe errors as “language usage deviating in definable ways from the way native speakers use the target language”. Such an approach would suggest that native speakers do not make errors, while in
fact “[w]e are all aware that in normal adult speech in our native language we are continually committing errors of one sort or another” (Corder 1967:166). This paradox can be resolved by introducing Corder’s (ibid.: 167) distinction between errors of performance and errors of competence. Errors of performance (or ‘mistakes’) are due to the pressure of online speech production and are shared with native speakers, whereas errors of competence (or ‘errors’ in the strict sense) result from incomplete knowledge of the language and normally display some systematicity. Errors in the strict sense are usually thought to be restricted to non-native speech (although systematic errors due to incomplete knowledge of the language can occur among some native speakers too, see Ringbom 1987), which might account for Tenfjord et al.’s (2006) definition. James (1998:78) also employs the term ‘error’ in the strict sense when he provides the definition of “[a]n instance of language that is unintentionally deviant and is not self-corrigible by its author”.

Fluency, and hence disfluency, is equally difficult to define. Chambers (1997:536), in an article entitled “What do we mean by fluency?”, observes that even if we exclude the non-technical sense of the term, i.e. the ability to speak a foreign language in a native-like manner (as in She speaks French fluently), “there is some degree of confusion as to what we mean by fluency and the professional use of the word covers several overlapping interpretations”. She refers to the communicative language teaching approach, which defines fluency as “effectiveness of language use within the constraints of limited linguistic knowledge” (ibid.), and the quantitative approach, which defines fluency in terms of quantifiable temporal variables such as speech rate, number and length of pauses, or frequency of lexical and non-lexical fillers like you know or uh (this latter approach is also called ‘pausology’, cf. O’Connell & Kowal 1980:8). Czwenar (2004:395) has shown that native teachers of English can have rather different working definitions of fluency and that some include features relating to accuracy (e.g. “correct use of grammatical structures”) and hence arguably the
absence of errors. The ‘accuracy’ element is also present in the definition of fluency given by Kormos & Dénes (2004:161) when discussing the results of a study into differences between fluent and non-fluent second language learners and native and non-native teachers’ perceptions of fluency: “[t]he results indicate that fluency is best conceived of as fast, smooth and accurate performance” (our emphasis). They further distinguish between ‘low-order fluency’ (temporal aspects of fluency) and ‘high-order fluency’ (proficiency, accuracy), a distinction that Osborne (2007) regards as useful in light of his own research into non-native fluency. He suggests that “syntactic complexity, formulaic language and semantic density are also factors in high-order fluency” (ibid.: 192). (See Fillmore 1979 for an even broader concept of fluency, including aspects such as being creative or being a good conversationalist; see also the special issue of Applied Linguistics, edited by Alex Housen and Folkert Kuiken in 2009, for additional definitions of fluency and comments about its relation to accuracy and complexity in second language acquisition.) More simply, fluency can be described as “the production of language in real time without undue pausing or hesitation” (Ellis & Barkhuizen 2005:139). By contrast, disfluency may be said to refer to lack of fluency, and disfluencies have been defined as “phenomena that interrupt the flow of speech and do not add propositional content to an utterance” (Fox Tree 1995:709). It should be pointed out, however, that some disfluencies like drawls do not really interrupt the flow of speech. In fact, silent and filled pauses, repeats or drawls are sometimes used deliberately for intelligibility or rhetorical effect (O’Connell & Kowal 2005:557).

Depending on the definitions one adopts, error and disfluency may be difficult to distinguish from each other (cf. Lennon 1990:395). This can be illustrated by Stenström & Svartvik’s (1994:242) questioning:
Can features of spoken discourse really be considered ‘deviant’ and ‘irregular’ when they occur with a frequency that borders on regularity in ‘normal’ native speakers’ use of the language? If so, where do we draw the line between ‘well-formed’ and ‘ill-formed’? Is there ‘acceptable’ deviance in spoken language, i.e. irregularities that are produced by, and acceptable to (or at least unnoticeable by) native speakers, as opposed to ‘unacceptable’ deviance, i.e. irregularities produced by non-native speakers but unacceptable to (and clearly noticed by) native speakers?

A sentence like (1), for example, taken from Hopper (2001:122), contains a repetition of is in a pseudocleft sentence. Hopper explains this repetition by the fact that “the speaker has processed the pseudocleft as an unanalyzed unit to the extent that she is unaware of having said is, and feels bound to produce a further is to mark the boundary of the pseudocleft” (ibid.).

(1) But surely surely what this suggests is is that you have to have classroom community and culturally-specific notions of what masculinity (...) and femininity are.

Given that, according to Hopper (ibid.), such repetitions are common in constructions of this type and may not necessarily be noticed by native speakers (whether they be the hearer or the speaker of the utterance), we may wonder whether they should be considered as real errors, or rather as some sort of disfluency resulting from the lack of pre-planning that is typical of spontaneous conversation (see Stenström & Svartvik 1994:249). The case of inappropriate collocations (e.g. fully different instead of completely/entirely/totally different, Granger
1998:153; *make an experience* instead of *have an experience*, Nesselhauf 2005:193) is equally difficult to resolve. While some characterise them as errors (cf. Sadeghi 2009), others may want to include them in a category of “infelicities” or, as Randolph (1917) would call them, “conventional aversions”, which he opposes to “fundamental errors”. The term ‘infelicities’ can more generally be used to refer to dispreferred non-native-like forms that cannot be seen to deviate from a given norm (Dagneaux et al. 2008). An example would be *We have here another role given to prisons* or *She has a strong belief that...* (Thewissen 2009).

Terminological confusion between errors and disfluencies is such that what is usually accepted as errors are sometimes called disfluencies and vice versa. Thus, Oostdijk (2003) includes errors (as exemplified by *I didn’t see too many evidence*) in her list of “types of disfluencies”, along with categories like hesitations, self-corrections or incomplete items. On the other hand, Clark & Clark (1977:262ff) refer to phenomena like silent and filled pauses, repeats (e.g. *Turn on the heater the heater switch*) or false starts (e.g. *Turn on the stove heater switch*) as ‘speech errors’. One and the same term can also be used to cover a number of different realities: unlike Clark & Clark (ibid.), who use the term ‘speech errors’ to cover both the performance phenomena mentioned above and slips of the tongue (e.g. *Turn on the sweeter hitch*), Poulisse (1999) uses the term to refer to slips of the tongue only.

The situation becomes even more complex when we compare native and non-native speech, for the status of the speaker may influence the way an erroneous or disfluent feature is perceived and characterised. Temple (2000:289) points out that “[g]enerally the native speaker’s pausing and hesitating go unnoticed by the listener. But this is not the case for the language learner, whose disfluent speech tends to define him or her as a nonnative”. Similarly, Lennon (1990:395) notes that if a speaker is identified as non-native, “pauses, repetitions, self-corrections and the like are then more likely to be interpreted as dysfluency
than they would be if produced by a native speaker”. The same is true of errors, which may be more easily perceived and condemned in non-native speech than in native speech (cf. Lennon ibid.: 396). Non-native speakers’ errors will arguably more readily be labelled as “errors of competence” than as “errors of performance” even if they result from processing problems rather than incomplete knowledge. As pointed out by Ellis (2008:46), while non-native speakers’ errors tend to be seen as “unwanted forms” (i.e. errors of competence), the errors made by children learning their first language (native speakers) tend to be considered not as errors but as “transitional forms”, and adult native speakers’ errors are usually regarded as “slips of the tongue” (i.e. errors of performance).

As demonstrated in this section, defining errors and disfluencies is a real challenge. They are rather complex and fuzzy notions that involve numerous factors and that could be seen to (at least) partially overlap. In this paper, the terms ‘error’ and ‘disfluency’ will be used in a fairly non-committal way: errors are defined as forms that deviate from a given native-speaker norm and disfluencies cover phenomena that are generally seen to reflect speakers’ online planning and encoding difficulties. In addition, we propose that errors and disfluencies form a continuum, with some phenomena closer to the “error” end of the continuum (e.g. complementation errors, morphosyntactic errors, article errors) and others closer to the “disfluency” end of the continuum (e.g. false starts, self-corrections, truncated words). It should however be borne in mind that the exact definition of errors and disfluencies, as well as the position of specific items along the continuum, depend on the context, the type of language investigated, the purpose of the study, etc. Here, we can only present a theoretical position that will have to be adapted to the situation at hand.

3. Errors and disfluencies in the era of corpus linguistics
The “corpus revolution” (Rundell & Stock 1992) has had a great impact on the study (and conceptualisation) of errors and disfluencies. On the one hand, the availability of verbatim transcriptions of authentic speech has brought to the linguists’ attention less salient phenomena that used to be overlooked or whose importance used to be downplayed. On the other hand, the automation made possible by the use of corpora has allowed for a more systematic analysis of these phenomena in context, for example through annotation. The first point will be dealt with in Section 3.1, while the second point will be examined in Section 3.2. In Section 3.3 we will show how non-standard forms and disfluencies are slowly making their way into reference works and will give examples from the literature of insights gained through the corpus approach. In Section 3.4, finally, we will say a few words about the limitations of spoken corpora to study errors and disfluencies, and how they may be combined with other types of data.

3.1. Beyond the written bias

Up until very recently, there has been a strong tendency among linguists to focus on written language to the detriment of spoken language. This is what Linell (2005) calls the “written language bias in linguistics”. Grammars used to be exclusively based on the norms of written language, sometimes illustrating their statements with examples taken from the great authors of the past, and linguistic studies mainly examined and described phenomena found in writing, which was regarded as “the ‘grammatical’ or correct language” (ibid.: 24; emphasis original). Speech, on the other hand, was often viewed as an inferior and less prestigious form of language, and has been quite appropriately described as “a stepchild of descriptive linguistics” (Tottie 1991:255).

The situation changed with the advent of spoken corpora like the London-Lund Corpus (LLC) or the Lancaster/IBM Spoken English Corpus (SEC), which, by giving access
to large quantities of authentic spoken text, contributed to a surge of interest in speech and in features typical of speech. Thanks to corpus-based studies like Chafe (1982), Biber (1988) or Flowerdew (1993), it has now become clear that the written and spoken modes present different characteristics and follow essentially different rules, which deserve to be considered in their own right. Studies centring around speech have become more common (e.g. Altenberg 1998, Aijmer 2002, Stenström et al. 2002, Biber et al. 2003, Cheng 2007, Norrick 2008) and grammars have started to integrate aspects of speech (cf. Biber et al. 1999 or Carter & McCarthy 2006; see also Section 3.3 below, as well as Leech 2000). Biber et al.’s (1999) grammar, for example, which relies on a corpus of spoken and written English, distinguishes between four registers of language, of which one represents conversation. In this grammar, conversational English is “treated as equal in standing to written English” (Hirst 2001:132) and a whole chapter is even devoted to the grammar of conversation.

It is in this context of revaluation of spoken language that research into disfluency should be situated. Although disfluencies had been investigated before electronic corpora became widespread (e.g. Maclay & Osgood 1959), the use of spoken corpora may be said to have been an eye-opener to both the presence and the functions of fillers, pauses and other related phenomena. Regarding their presence, it is a fact that “normal”, non-pathological disfluencies often go unnoticed in everyday conversations (except perhaps when they are produced by language learners, see Temple’s quote above) as listeners tend to listen selectively (Stubbs 1983). In an experiment, Lindsay & O’Connell (1995) asked four undergraduate volunteers to transcribe an audio-taped interview. They observed that discourse markers (e.g. well, now) and hesitation phenomena (filled pauses, false starts, repetitions, etc.) were among the most likely items to be missing from the transcriptions. This is because, in Rieger’s (2003:43) words, “addressees tend to tune [them] out”. Not only the addressees, in fact, tend to tune them out, but sometimes the speakers themselves. Thus, Watts (1989:216)
explains how, in a family gathering, a person criticising the use of well to start an utterance prefaced his response by well without realising it. The same is true, to a certain extent, of errors. We seem to have, as Hill (1973:205) puts it, “a sort of internal editing mechanism which enables us to correct a speaker’s error without noticing that an error has occurred”. By providing linguists with verbatim transcriptions of authentic speech, corpora have given more prominence to less conspicuous “errors” like atypical pseudocleft sentences (Hopper 2001, see above) or unexpected forms of question tags (e.g. I won’t even talk to myself am I, Axelsson 2009). Spoken corpora have also made it possible to cast light on the recurrent and widespread use of some of these non-standard forms, thereby also questioning their erroneous status in speech. McCarthy & Carter (2002:64) argue that, if what they call “unpleasant anomalies” (e.g. the use of would in conditional clauses) are shown to be recurrent and “distributed across a wide range of speakers and contexts in a corpus designed to reflect a broad demographic and social spectrum” (ibid.: 65), they should be included in the grammar “even though [they] may still be deemed unacceptable in more formal contexts or in writing” (ibid.: 66). In the same vein, although the words “performance error” and “erroneously” crop up in Biber et al. (1999:1063-1066), the authors later highlight the “unclarity about what makes a well-formed grammatical construction in speech. Some types of apparent blend are relatively systematic, and it could be argued that they are not performance errors: that in speech, the language tolerates a freedom of syntactic structure that would generally be regarded as unacceptable in writing” (ibid.: 1065).

The use of spoken corpora has also highlighted the numerous functions of fillers and discourse markers that used to be dismissed as “throwaways” (Erard 2004), “exasperating expressions” (Stubbe & Holmes 1995) or even “pollution” (Boula De Mareüil et al. 2005:27), and hence has helped portray disfluencies and related phenomena in a more positive light. The very term ‘disfluency’ (or ‘dysfluency’), as demonstrated by Rühlemann (2006:400),
implies a negative evaluation of the phenomenon, with the prefix *dys-* alluding to a pathological condition. Similarly, the term ‘false start’ is pejorative, suggesting as it does that false starts “invariably impact detrimentally on comprehension and interaction” (ibid.: 399). Reference to hesitation is also negatively loaded, and describing filled pauses as “hesitation markers” is, according to Kjellmer (2003:171), not justified, for “hesitation may not always be involved”. The detailed study of disfluencies in context, facilitated by the availability of spoken corpora, has shown that most of them, rather than being “a nuisance, (...) a kind of debris lying in the way of an ordered exposition” (ibid.: 170), actually fulfill important functions in discourse (see also Rühlemann et al. this issue). Mukherjee (2000:582), for instance, distinguishes between two types of intended pauses at tone unit boundaries, each of which has a specific function: segmentation pauses “emphasize the demarcation of a syntactic or a textual unit” and anticipation pauses “enhance the hearer’s attention as to the following information”. Drawls on the (that is, the pronunciation of the as “thiy” rather than “thuh”) are said to prepare the hearer for an immediate suspension of speech due to a major problem in production (Fox Tree & Clark 1997). The discourse markers *sort of* and *you know* can be seen to offer the hearer guidance as to how the message should be interpreted (Aijmer 2002, Hasselgren 2002; see also Aijmer’s description of the functions of *well* in this issue). More generally, disfluencies “can be seen as adaptations to the needs arising from the interactive nature of real-time conversation” (Rühlemann 2006:402). (Romero Trillo 1997, Swerts 1998 and Kjellmer 2003 are other examples of studies dealing with the functions of disfluencies.) In addition, it has become increasingly clear that the investigation of disfluencies has interesting applications in different fields (see Section 4), which reinforces their perception as useful devices. As a result of these findings, disfluencies have gradually moved from being viewed as “nonwords” (Lennon 1990:406), “communicatively disturbing” phenomena (Möhle 1984:36), to being recognised as “legitimate parts of the English language” (O’Connell &
Kowal 2004:467). This shift is reflected in terminology, with the appearance of (more positively connoted) terms like “speech management phenomena” (Allwood et al. 1990) or “planners” (Tottie this issue).

3.2. Annotation and computer-aided analysis

Modern corpus linguistics offers the advantage of total or partial automation. It makes it possible, among other things, to retrieve all the occurrences of a particular item automatically, for example all the occurrences of the filler *erm* or of the combination *she don’t*. The range of possibilities is even wider if the corpus is annotated. If the corpus is annotated with parts of speech (POS-tagged), one can for instance look for *there’s*/*there is* followed by a plural noun (the tag “NN2” in the CLAWS7 tagset) to retrieve constructions like *there’s people* (see Biber et al. 1999:186 on this type of construction). With a parsed corpus, the annotation of syntactic functions and structures would allow the user to search for, e.g., cases where a verb is (incorrectly) separated from its direct object by an adverb (as in *He speaks well English*, see Osborne 2008 on the use of such structures in written learner English). A semantically annotated corpus, finally, could make it possible to extract, say, occurrences of *which* preceded by an animate noun (e.g. *the man which*...).

It should be borne in mind, however, that while these types of automatic annotation (with the exception of semantic annotation, probably) have become quite commonplace for written corpora, with success rates that are, if not perfect, at least perfectly reasonable, the situation is quite different for spoken corpora, especially when the corpus data were produced by non-native speakers of the language (‘learner corpora’). This is due to the fact that most of the software programs used to annotate corpus data were designed to deal with standard written data and may therefore experience difficulties when confronted with features typical of (native or non-native) speech (among the few exceptions is SOUP, a parser for
spontaneous speech, see Gavaldà 2004; in addition, COALA and COMOLA were designed to parse interlanguage data – not specifically spoken interlanguage data, though – see Pienemann 1992 and Jagtman & Bongaerts 1994, respectively). As pointed out by Rahman & Sampson (2000:309), who discuss the challenges of annotating spontaneous spoken English within the framework of the CHRISTINE project, “in annotating speech, whose special structural features have had little influence on the analytic tradition, ambiguities of classification constantly arise that cut across traditional category schemes”. For example, they report that deciding whether particular sequences of words are repairs or well-formed constructions can sometimes be far from straightforward. Tagging sequences with unclear passages and distinguishing between performance errors and non-standard dialect are equally challenging for annotators.

Oostdijk (2003) aptly illustrates the problems one may encounter when parsing native spoken corpus data (in this particular case, the spoken part of ICE-GB, the British component of the International Corpus of English). The author shows how the parsing of ICE-GB was preceded by a phase of normalisation, aimed at dealing with disfluencies and involving three main processes. First, phenomena like repetitions, false starts, self-corrections and hesitations were marked up for normative deletion and excluded from the input to the parser (in the sentence I’m bi ... bilingual, for example, the repetition of the “bi” syllable was deleted). Second, normative insertion took place when words had been omitted (e.g. in the sentence you see it as another dance form within own right, “its” was inserted before “own right” to form the phrase “within its own right”). Finally, some phenomena necessitated both deletion and insertion, for example with incomplete words (e.g. in that is hilar, “hilar” was deleted and “hilarious” was inserted) or sound errors (e.g. inagdequate was replaced by inadequate). In addition, Oostdijk points out that lexical and grammatical errors in ICE-GB were normalised so as to facilitate the task of the parser (e.g. too many evidence became too much evidence,
and you was became you were). It must be underlined that the approach described by Oostdijk for the parsing of ICE-GB is largely manual. By contrast, Dister’s (2007) pre-treatment of the data in an attempt to facilitate the POS tagging of French spoken corpora is fully automatic. It takes care of disfluencies by marking some elements as material to be ignored (e.g. repetitions, self-corrections and filled pauses). Other disfluencies are marked up but not disregarded in the tagging process, for instance silent pauses and truncated words. Using this system, Dister is able to obtain a relatively good accuracy rate of 77.23% for the POS tagging of her data.

Annotating spoken data produced by learners is even more complex, as such data are likely to include a higher number of errors and disfluencies than native corpora, and also a larger variety of them (see Ringbom 1987:73 about errors), perhaps more confusing for a tagger or parser. While some attempts have been made to POS tag written learner corpora (cf. Aarts & Granger 1998, Granger & Rayson 1998, Meunier & de Mönnink 2001, Van Rooy & Schäfer 2002), and with quite good results (a pilot study carried out on the POS-tagged version of the International Corpus of Learner English, ICLEv2, revealed accuracy rates ranging between 95 and 99.1%, cf. Granger et al. 2009:16), work on spoken learner corpora has hardly started. Jendryczka-Wierszycka et al. (2009) first applied the standard version of CLAWS4 (Garside & Smith 1997) designed for written English on samples from LINDSEI with an accuracy rate of 88.3%. As well as identifying various problems linked to the transcription conventions in LINDSEI, the authors described a mismatch between the written version of the tagger and spoken nature of the corpus (e.g. uhu tagged as a singular noun or the discourse marker like tagged as an infinitive in stairs will like I will fall) and discussed learner-related issues (e.g. the noun flat tagged as an adjective in we’re gonna rent flat or apartment, probably due to the absence of a determiner). Using several repair mechanisms (e.g. adapting some of the transcription conventions to those that were recognised by the POS
tagger, adopting a spoken version of the tagger and including an idiomlist corresponding to some common structures typical of speech), they were able to improve the accuracy rate up to 98.5%, a score which appears to be as high as those achieved when POS tagging written learner corpora.

We saw above that spoken corpus data may be annotated despite the presence of errors and disfluencies. But errors and disfluencies may also themselves be the target of the annotation. The process known as ‘error tagging’ consists in marking the errors in a corpus and, usually, providing a corrected form for each error. The term is normally used with reference to learner corpora, with error tagging being “specially designed to cater for the anomalous nature of learner language” (Granger 2002:18). Error tagging lies at the basis of a method called “computer-aided error analysis” (CEA), which builds on traditional error analysis but represents a notable improvement over it, as described in Dagneaux et al. (1998).

While it considerably facilitates the analysis of errors in corpora, it must be said that error tagging is a complex process, as it involves drawing a line between erroneous and non-erroneous forms (although the system proposed in Dagneaux et al. 2008 also includes a category for infelicitous forms) and, in the case of erroneous forms, interpreting what the learner meant to say in an attempt to propose a correct alternative (although some types of annotation make it possible to propose more than one correction, cf. the system used to annotate the corpus of written learner German Falko, Fehlerannotiertes Lernerkorpus, Lüdeling et al. 2005). This twofold task may turn out to be quite tricky, since errors tend to be both difficult to identify (Lennon 1994:89, Hasbún Hasbún 2007:117) and difficult to reconstruct – and hence correct (Hamid 2007). Error tagging is also, to some extent at least, subjective, as appears from the fact that “the number of errors detected by two different correctors annotating the same texts can differ drastically” (Dagneaux et al. 2008:7) – although, as noted by Andreu-Andres et al. (2010), the degree of inter-rater reliability is to
some extent also dependent on the type of errors (morphosyntactic errors, for instance, tend to reach a higher degree of inter-rater reliability than lexical errors). Moreover, it is a time-consuming process, as it is generally performed by hand (but see Izumi et al. 2004 for an attempt at automatic detection of learners’ errors), sometimes with the help of an error editor which can speed up the error tagging process somewhat by the menu-driven insertion of the error tag and the corrected form (cf. UCLEE, the Université Catholique de Louvain Error Editor, Dagneaux et al. 1998). Up to now, error tagging has essentially concerned written learner corpora, making it possible to automate the retrieval of errors in learners’ writing and allowing their analysis in context. Recently, however, error tagging of spoken learner corpora has been undertaken. Kämmerer (2009) explains how the UCL Error Tagging Manual (Dagneaux et al. 2008), which targets written learner errors, was applied to the German component of LINDSEI, and how it was partly adapted for use on spoken data, for example when dealing with self-corrections or error repetitions. The NICT JLE (Japanese Learner English) corpus is another spoken corpus that has been (partly) error tagged, and whose preliminary analysis (e.g. Izumi & Isahara 2004, Tanimura et al. 2004) has yielded promising results. The practice of annotating disfluencies offers the same sorts of advantages as error tagging in that it facilitates the retrieval of disfluencies and their study in context. The CHAT transcription format (MacWhinney 2010), developed within the framework of the CHILDES (Child Language Data Exchange System) project, makes it possible to represent many types of disfluencies, including incomplete and omitted words, filled and unfilled pauses, false starts, repetitions with and without correction, but also errors. It has been used to annotate the PAROLE (Parallèle, Oral, en Langue Etrangère) corpus, a learner corpus containing oral productions from learners of English, French and Italian (Hilton et al. 2008), which has subsequently been exploited to investigate fluency in non-native speech (e.g. Osborne 2007), as well as its links with other phenomena like overall proficiency (Osborne in press) or
informational content (Osborne this issue). Fluency annotation schemes are sometimes
developed within the framework of specific research projects. For example, Dahlmann &
Adolphs (2009) report on how the delicate and precise annotation of pauses (making use of
audio recording and wave forms) in the English Native Speaker Interview Corpus (ENSIC)
has been used to assess the psycholinguistic validity of some multi-word expressions (see
Section 4). Note that error tagging and the annotation of disfluencies may be conveniently
combined with each other, and also with POS tagging and/or parsing, by using a multi-level
standoff architecture, as described in Lüdeling et al. (2005) or Mukherjee (2007), where each
level corresponds to one type of annotation (e.g. one line for POS tags, one for disfluencies,
one for error identification, one or several for correction(s), etc.).

3.3. The study of errors and disfluencies

Until fairly recently, speech-related phenomena like errors and disfluencies did not figure
prominently in descriptive linguistics (which of course is to be linked to the general bias in
favour of written, rather than spoken language in linguistics; see Section 3.1). That this
situation is slowly changing is reflected by the fact that such phenomena have started to make
their appearance in reference works like grammars or dictionaries. While disfluencies used to
be either “not accounted for in grammars” (Stenström & Svartvik 1994:242) or “summarily
dismissed” (Kjellmer 2003:170), they have now found their way into corpus-informed or
corpus-based grammars like Biber et al.’s (1999) Longman Grammar of Spoken and Written
English (based on the Longman Spoken and Written English Corpus) and Carter &
Corpus), together with other phenomena typical of spoken language. Long sections are
specifically devoted to spoken language, with particular emphasis on errors and disfluencies:
“The grammar of conversation” in Biber et al. (1999:1037-1125) and “Spoken language” in
Carter & McCarthy (2006:163-240). As far as dictionaries are concerned, filled pauses (usually labelled as interjections) seem to be increasingly considered as worthy of an individual entry, as testified by O’Connell & Kowal’s (2004:464) comparative list of dictionaries from the 1930s to the year 2003 (see also Tottie this issue for more recent dictionaries, including online dictionaries). While this trend contributes to the recognition of filled pauses as words, it must be acknowledged with O’Connell & Kowal (2004:469) that “the legitimation of _uh_ and _um_ as words is far from accomplished”.

What may contribute to an even better recognition of speech-related phenomena like errors and disfluencies in the future is a joint effort, among linguists, to pursue their research into these phenomena. The growing literature on the subject over the last few years suggests that things are shaping up, and as an illustration of this, we would like to briefly present, in what follows, four studies that have exploited spoken corpora to shed light on errors and disfluencies in native and non-native speech.

The first study centres around disfluencies in native speech. In a paper entitled “Hesitation. In defence of _er_ and _erm_”, Kjellmer (2003) sets out to investigate the use of the fillers _er_ and _erm_ in _Cobuild Direct_, a corpus which is part of the Bank of English. The focus is on frequency of occurrence and more specifically on frequent co-occurrence patterns and functions. Kjellmer presents a very detailed overview not only of the most frequent collocates of the filled pauses (e.g. _and_, _er_, _I_) but also, using the POS tagging in the corpus, of their various preferred co-occurrence patterns when introducing thought units at word, phrase and clause level. When it comes to function, it is suggested that, rather than dismissing _er_ and _erm_ as haphazard and automatic, they “should be looked upon in most cases as task-performing elements, employed to bring about certain effects” (ibid.: 181). Beside “hesitation proper” (presented as “not the whole story”, ibid.: 183), the following functions are highlighted and amply illustrated using data from the corpus: sign-posting speaker turns (turn-taking, turn-
holding, turn-yielding), attracting attention, highlighting significant elements in the utterance and correcting part of the utterance. In view of these functions, which can to some extent be carried out simultaneously, *er* and *erm* are presented as beneficial to both the speaker and the listener.

The focus of the next study is also on disfluencies but in non-native speech. Götz (2007) investigates selected performance phenomena in the German component of LINDSEI and in the *Louvain Corpus of Native English Conversation* (LOCNEC): the use of filled and unfilled pauses, repetitions of function words, and repetitions involving pronoun subject-verb contractions (e.g. *it’s it’s*). She shows that the German-speaking learners tend to overuse filled and unfilled pauses, as well as some repetitions. On the other hand, she finds that the learners display a marked tendency to underuse repetitions of pronoun subject-verb contractions (compared with the native speakers) and that, more generally, contracted forms are underrepresented in the learners’ spoken productions. Götz (ibid.: 80) suggests that the marked underuse of these typical features of native speech may well make learner speech less natural than native speech. Qualitatively, the study also uncovers differing positions of disfluency phenomena in learner and native speech, which can be seen to reflect the considerable planning pressure learners experience throughout syntactic units (not predominantly at the beginning of clauses or utterances like native speakers).

Corpus-based studies of errors in native and learner speech are comparatively less frequent. An illustration for native speech is Stromswold (1989, 1994), who investigates the acquisition of lexical vs. functional morphemes and of auxiliaries by young children in spontaneous speech transcriptions from the CHILDES corpus. The children are shown to use erroneously inflected lexical words (e.g. *eated* for *ate*) thirty times more frequently than erroneously inflected function words (e.g. *somes, thems*), which Stromswold sees as evidence that “children have innate knowledge of the existence of finite categories and infinite
categories of words” (1994:8). As far as auxiliary verbs are concerned, most error types are infrequent except perhaps for inversion errors, which occur consistently in the data. In addition, the data include no clear instances of (inflectional, combination or word order) error types the children would have made if they had mixed up auxiliaries and lexical verbs. Stromswold concludes that children tend to acquire the auxiliary system conservatively in that they do not tend to generalise their knowledge of the behaviour of one auxiliary or lexical verb to another auxiliary. More generally, Stromswold’s research highlights the contribution of spoken corpora to the study of errors in first language acquisition (compared with experimental data or collected lists of errors): researchers have access to the (in)frequency of occurrence of the errors investigated in naturally-occurring speech.

An analysis of learner speech can be found in Izumi & Isahara (2004), who use data from the NICT JLE corpus to test two hypotheses about the acquisition order of eight major grammatical morphemes. The first hypothesis is that the acquisition of these morphemes follows a “natural” order which is shared by learners of different ages, from different mother tongue backgrounds, different learning environments, etc. The second hypothesis is that learners’ mother tongues may affect the order of acquisition and that Japanese learners, for example, may display a sequence that does not necessarily correspond to the purportedly natural sequence. The basis for their analysis is a set of error-tagged samples from the NICT JLE corpus, representing spoken English produced by learners of different proficiency levels (as determined through an oral proficiency test called the Standard Speaking Test). By examining the erroneous use of the eight grammatical morphemes, but also their omission and correct occurrence in obligatory context, the authors establish a ranking of morphemes that differs from the “natural” sequence described in the literature. In particular, Japanese learners appear to acquire articles and plural -s at a later stage, which could be due to the fact that in Japanese there are no relevant markers for articles and plural forms. These findings suggest
that the mother tongue has a role to play in the order of acquisition of grammatical morphemes.

We could have cited many other studies (e.g. Saxton 2000, Clark & Fox Tree 2002, Campione & Véronis 2004, De Cock 2004, Müller 2004, Abe 2007, Tono 2007, Gilquin 2008), all of which rely on spoken (native or learner) corpora to provide a better understanding of the contexts of use of errors and disfluencies in speech, as well as the functions and likely effects of such uses. And while some types of studies are less frequent than others (studies of errors in native and non-native spoken corpora, in particular, still remain relatively rare), a review of the literature shows that errors and disfluencies in speech are no longer dismissed as a marginal and negligible aspect of language.

3.4. The limitations of spoken corpora – and other types of data

As demonstrated above, the availability of large quantities of naturally-occurring spoken data in the form of well-balanced and well-designed corpora has opened up new possibilities for the study of errors and disfluencies. However, we should not forget that spoken corpora may also present some problems when they serve as a basis for investigating such phenomena. For reasons outlined earlier, some erroneous or disfluent forms may be overlooked during the transcription process. Just like Lindsay & O’Connell’s (1995) undergraduate volunteers referred to in Section 3.1, professional transcribers may delete some elements, add others, or substitute one for another accidentally, thus correcting certain errors or introducing new ones (see Randolph 1917:323 for the example of a stenographer, in the pre-corpus era, who transcribed “they wuz” as *there was* a significant number of times, despite her belief that “her reports were absolutely faithful”; see also Mollin 2007 for another example, as well as Stubbs’s 1983: 228 discussion of “auditory hallucinations”). In addition, transcribing speech may involve a certain degree of subjectivity. For instance, the length of silent pauses or of
syllable prolongation, if measured perceptually rather than instrumentally, may not be a reliable indicator (cf. O’Connell & Kowal 1990). Subjectivity is also involved in an example like the following, taken from LINDSEI, where the fillers *er* and *eh* might as well have been interpreted as repetitions of the article *a*:

(2) at one moment there was a rock 'n' roll and *(er)* *(eh)* a friend of mine an Irish friend of mine and me we decided to to dance rock 'n' roll with her (LINDSEI-FR005)

Transcribing errors and disfluencies undoubtedly enhances the value of a corpus (Leech et al. 1995) and facilitates their quantitative and qualitative analysis, but it should be borne in mind that transcription is theory (Ochs 1979), necessarily influenced by the transcriber’s own biases. And the fact that the audio files of spoken corpora are usually not publicly accessible makes any verification against the original recordings almost impossible (see Tottie this issue for a similar remark). Besides, working exclusively with transcripts exposes researchers to what Stubbs (1983:228) calls the “estrangement effect” of transcription: spoken discourse “looks odd, incoherent and broken when seen in the written medium – but it does not sound odd to those taking part in it” (ibid.; emphasis original).

When using corpus data, one question worth asking is how prototypical one’s corpus is. Prototypical corpora are characterised by the fact that they have been produced in a natural communicative setting, which sets them apart from more experimental data like acceptability judgements, word association tests or measurements of reaction times (Gilquin & Gries 2009:6). In this respect, so-called “corpora of speech errors” (Schattuck-Hufnagel & Klatt 1980, James 1998, Stemberger 2009), “corpora of slips of the tongue” (Poulisse 1999) or “corpora of syntactic blends” (Coppock 2010), which consist in lists of speech errors, slips of the tongue or syntactic blends do not qualify as corpora in the corpus linguistic sense. But
even among those collections of oral texts that do qualify as corpora, some are more prototypical than others. The LLC, for example, is a prototypical spoken corpus in that it contains texts that were produced in natural communicative settings: face-to-face conversations, telephone conversations, interviews, public orations, etc. The core part of LINDSEI (i.e. the whole database with the exception of a picture description task) is slightly less prototypical because the interviews it is made up of were not produced for real communicative purposes, but for classroom (and corpus collection) purposes. A corpus of picture-based storytelling (e.g. Lennon 1994, Ash et al. 2010), on the other hand, is a peripheral type of spoken corpus, in which the data were not produced fully naturally, but were elicited by means of a picture or series of pictures. A review of the literature suggests that peripheral corpora are actually quite common in corpus-based studies of errors and disfluencies. Kormos’s (2000:366) list of tasks used in the investigation of self-repairs, for example, mostly points to data that are closer to peripheral corpora: picture description, spatial description, storytelling, information-gap activity. This is not to say that non-prototypical corpora cannot be used to study errors and disfluencies in speech. In fact, there can be sound arguments in favour of using more elicited data. Kormos & Dénes (2004:151), for example, explain that the reliance on picture-based narratives in studies of (dis)fluency enables researchers to eliminate the content variable, a variable that may be of paramount importance in some contexts, since “having to produce different types of content places different cognitive load on speakers, which, in turn, influences the fluency of production”. However, one should be aware, when using such non-prototypical corpora, that the type of language they include is likely to differ from that found in prototypical corpora, and that, consequently, the nature and frequency of the errors and disfluencies observed do not necessarily correspond to their nature and frequency in spontaneous, naturally-occurring speech.
The same is true of other types of data that can be relied on to study errors and disfluencies in spoken language. While their use may be justified by the particular purpose of one’s study, they should be employed in full awareness of their limitations. Thus, some linguists rely on the pen-and-paper method to study errors (e.g. Ferber 1991), a method that consists in writing relevant items along as they are heard by the linguist (Poulisse 2000:138). Because of a phenomenon known as “perceptual bias” (Ferber 1991) or “perceptual confusion” (Cutler 1982:16ff), however, some items are more easily perceived than others and hence more likely to be recorded (cf. Poulisse 2000:141 for an illustration of how this may have influenced the results of a study on slips of the tongue). Other studies are based on (quasi-)authentic spoken data, but a very small set of them, often painstakingly transcribed by the linguist him/herself and analysed manually, i.e. without the tools and methods that characterise corpus linguistics. Larsen-Freeman (2006), for example, examines fluency and accuracy as evidenced by the production of five Chinese learners of English (see also the literature review in García-Amaya 2009:69-70). While certainly valid for the particular speakers under investigation, the results obtained in such studies may not be representative of a larger class of speakers, and one must therefore acknowledge that the conclusions drawn are only tentative, being constrained by the limited scope of the data. Alternatively, data can be collected in an experimental setting, for example when attempting to induce idiomatic blends (Cutting & Bock 1997) or predictable involuntary speech errors using the so-called phonological bias technique (see Baars 1980), or when setting out to assess the effects of false starts on comprehension (Fox Tree 1995). Studies of errors and disfluencies in pathological subjects, in particular, appear to frequently rely on experimental data (e.g. Martin et al. 1996, Jacobson & Livert 2010). Such studies show that experimental data can provide valuable insights into errors and disfluencies. Moreover, working in an experimental setting makes it possible to systematically control for more variables than is usually the case in corpus studies.
However, because experimental data lack the spontaneous character of spoken corpus data, they may miss important aspects of the use of errors and disfluencies.

The above makes it clear that corpora and other types of data have both advantages and disadvantages. This probably explains why some scholars have chosen to combine corpus data and other types of data, especially experimental data, to analyse errors and/or disfluencies. While the corpus data make it possible to examine how these items are used in context and in authentic speech, the experimental data can, for instance, help assess their perception by hearers or the influence of specific linguistic contexts on their production. Kormos & Dénes (2004) nicely illustrate how a corpus-based study of temporal variables in non-native speakers’ oral productions can be combined with native and non-native teachers’ perception of their fluency (see Brand & Götz this issue for another example of a study combining corpus analysis and perception assessment).

4. Applications

The corpus-based study of errors and disfluencies in spoken language has possible applications in various fields. The most obvious one is probably Foreign Language Teaching (FLT), but there are others, including language testing, psycholinguistics, clinical linguistics and Natural Language Processing (NLP). In this section, we give a brief overview of some of these applications.

In FLT, a good knowledge of the types of errors typically made by learners can lead to more efficient teaching and teaching materials. Such an approach seems to have been mostly applied to writing. The second edition of the *Longman Dictionary of Common Errors* (Turton & Heaton 1996), for example, is a good illustration of how the error analysis of a written learner corpus (in this case, the *Longman Learners’ Corpus*), used in combination with a native corpus (the *British National Corpus*), can help describe recurrent errors and propose
appropriate corrections. The second edition of the *Macmillan English Dictionary for Advanced Learners* (Rundell 2007) goes even further than this by including more extended ‘Get-it-Right’ boxes under some entries, and incorporating a thirty-page academic writing section highlighting certain difficulties often encountered by learners in the production of academic essays, for instance phraseological infelicities, semantic misuse or register confusion (cf. Gilquin et al. 2007). Similar work still remains to be done for speech, however. It would also be interesting, as Albrechtsen et al. (1980:394) propose, to examine the degree of intelligibility of specific types of errors (in Albrechtsen et al.’s study, for example, errors at the discourse level turned out to be especially difficult to understand). This would enable teachers to give priority to errors that directly affect intelligibility.

As regards disfluencies, one must recognise that “hesitation strategies” like the use of pauses or discourse markers are hardly ever taught in the foreign language classroom (Rieger 2003:41). Yet, it has been claimed that “a native-like use of pauses, fillers and repeats is a quick way for foreign learners to improve their English language proficiency” (Stenström & Svartvik 1994:252). Hasselgren (2002), for example, shows how the use of smallwords such as *you know* or *I mean* gives an impression of fluency. Similarly, Derwing et al. (2004:675) note the potentially beneficial effect of teaching “phrases to buy planning time” or “appropriate pause placement”. It thus appears important to incorporate hesitation strategies into the foreign language curriculum, since they make it possible to increase fluency and help deal with planning pressure. This is all the more important since, as rightly emphasised by Götz (2007), planning pressure is higher in a foreign language than in one’s native language. Learners should also be taught to recognise hesitation phenomena in native speech, for, as demonstrated by Reed (2000), such phenomena tend to pose perception problems for non-native speakers. As Reed (ibid.: 89) explains, learners often “do not filter them out, but rather attempt to assign meaning to a speaker’s faulty output or to discourse markers such a (sic)
‘say’ or ‘like’” (for example, transcribing and uhm then a boy as and one then a boy; ibid.: 85). Future teachers, in particular, should be taught to use hesitation phenomena adequately, not only because they are important target models for learners in classroom contexts (Mukherjee 2009:225), but also because “hesitation phenomena are a possible source of massively incomprehensible input and the use of non-grammatical pauses in teacher-talk is a potential source of confusion for the non-native learner” (Griffiths 1991:347).

The investigation of errors and disfluencies is also relevant for the related field of language testing and assessment. While based on written data, the study carried out by Granger & Thewissen (2005) offers a perfect illustration of how an error-tagged learner corpus can be exploited to flesh out the descriptors found in the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001), with a view to establishing levels of proficiency more accurately. Götz (2007) also seeks to improve the CEFR descriptors, but this time using disfluencies. She highlights the benefits of studying performance phenomena like filled and unfilled pauses or repeats to improve the vague descriptors provided for fluency (see also Hasselgren 2002, as well as Osborne 2007 and this issue for similar attempts at refining descriptors of fluency for testing on the basis of native and/or non-native spoken corpora).

Next, errors and disfluencies can help psycholinguists fine-tune models of speech production (such as Levelt’s 1989 model) or of language development. For example, Coppock (2010) shows how her analysis of both a set of syntactic blends collected from spontaneous speech and automatically generated unattested pseudo syntactic blends lends support to the single-buffer hypothesis, according to which “multiple formulations of the same message are developed in the same memory buffer, and the developing syntactic representations may interact and compete with one another during grammatical encoding” (ibid.: 48). Sabin et al. (1979:50) suggest that “a methodologically rigorous analysis of temporal speech patterns and
hesitation phenomena provides a viable and essential psycholinguistic approach to understanding language development”, and Kormos’s research on monitoring in second language speech (see e.g. Kormos 1999, 2000) is a good illustration of this. Another phenomenon that can be investigated on the basis of disfluencies is the holistic storage (or otherwise) of multi-word expressions. Dahlmann & Adolphs (2009), starting from the claim that pauses are “indirect indicators of prefabricated language and holistic storage” (ibid.: 126), investigate the pause patterns around the recurrent phrase I think and discover that the phrase is very rarely interrupted by pauses, which suggests that it may be stored holistically in speakers’ minds.

Applications of studies on errors and disfluencies can also be found in the field of clinical linguistics, and more precisely ‘pathological speech’ (e.g. stuttering, aphasia, Alzheimer, Down’s syndrome). Roberts et al. (2009:422) suggest that studies of disfluencies in non-stuttering adults in spontaneous spoken productions may be useful both when diagnosing stuttering, as “the breakdown of types of disfluencies provides clinicians a way to judge whether a client’s speech contains too many of these disfluencies”, and when treating stuttering, to provide adults who stutter with a model of normal disfluencies. Knibb et al. (2009) argue that quantitative data relating to, among others, the number of grammatical and speech sound errors in the productions of patients with progressive non-fluent aphasia are needed to improve clinical descriptions of their conversational speech.

In the field of NLP, the study of disfluencies may be useful for speech recognition and speech synthesis. Stolcke & Shriberg (1996) and Duchateau et al. (2003) explore to what extent removing (some) disfluencies from speech may help improve the results of automatic speech recognition. But other studies actually take advantage of disfluencies to enhance the performance of automatic speech recognisers. Swerts et al. (1996), for example, maintain that filled pauses act as markers of discourse structure, and that speech recognisers may therefore
make use of disfluencies to detect discourse structure (turns introducing a new topic, for instance, appear to contain comparatively more disfluencies). Stenström (1986), on the other hand, is interested in speech synthesis, and how corpora may be used to construct a “plausible model for pause assignment” (ibid.: 203). Relying on the analysis of a short corpus sample, she ends up with a set of preliminary rules that predict the position and duration of silent pauses in synthesised speech.

Errors and disfluencies may also be investigated in various other fields, such as sociolinguistics (e.g. Bortfeld et al. 2001), politeness theory (cf. Stubbe & Holmes 1995:64) or psychology (e.g. Brennan & Williams 1995). Generally speaking, they may therefore be said to “constitute a rich source of data for several disciplines” (Kirsner et al. 2003:13).

5. The papers in this issue

The five papers brought together in this special issue rely on (native and/or learner) corpora to tackle various aspects of errors and disfluencies. The first two papers focus on the use of pauses in native speech. Gunnel Tottie investigates the frequency of use of the so-called filled pauses (or fillers) er/uh and erm/um in spoken British English using the spoken component of the British National Corpus (BNC). The study makes use of the sociolinguistic annotation available in the BNC (gender, age, socio-economic class) to uncover the differences of frequency between the demographic part of the spoken BNC, which is made up of informal conversations, and the context-governed part, which consists of spoken language collected in specific business, education, leisure and public/institutional contexts. The possible sociolinguistic conditioning of the use of the nasalised erm/um and non-nasalised er/uh variants is also examined. The analysis reveals that the fillers under study are used more often by men, older speakers (aged over 60) and more educated speakers, and that the nasalised erm/um variant is used more frequently by women, younger speakers and more educated
speakers. In other words, the fillers can be seen to function as sociolinguistic markers. The author also argues that the sociolinguistic factors investigated can to a certain degree account for the highly frequent use of fillers in another spoken corpus, namely the *London-Lund Corpus*. The paper is rounded off by making a strong case for portraying the items studied in a positive light; the term ‘planner’ is put forward to refer to them, thereby highlighting their essential role in the management of spoken discourse.

In their paper, Christoph Rühlemann, Andrej Bagoutdinov and Matthew Brook O’Donnell also provide evidence against the treatment of pauses as mere disfluencies in spoken discourse. The study examines the functions of silent and filled pauses on the basis of the recently created *Narrative Corpus* (NC), a corpus of conversational narratives extracted from the demographic (conversational) part of the spoken component of the *British National Corpus*. The NC, which contains both narratives and some context, has been annotated at discourse level (e.g. narrative components like ‘pre-narrative components’ or ‘narrative-initial utterances’, quotatives, direct vs. indirect reporting modes), which makes it possible for the authors to carry out a fine-grained comparative analysis of the lexical and discoursal association patterns of pauses in conversational narrative and general conversation. The results confirm the authors’ hypothesis that pauses are used more frequently in narratives than in general conversation. The analysis reveals that this can be accounted for by three main factors: (1) narrators’ need, in narrative-initial utterances, to secure the right to a series of turns and to orient listeners to the situation in which the events evolved; (2) narrators’ use of coordinated sequences of narrative clauses that each match a specific story event; and (3) narrators’ preference for the cognitively demanding direct-mode discourse presentation.

The next paper, by Karin Aijmer, combines native and non-native corpus data and focuses on the pragmatic marker *well*. Using the Swedish component of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI) and the *Louvain Corpus*...
of Native English Conversation (LOCNEC), the author demonstrates that, far from being simply a marker of disfluency, *well* is used to fulfil a number of important functions in spoken discourse. These functions can be grouped into two major categories: speech management functions (e.g. planning, reformulation, introducing a quotation in direct speech) and attitudinal functions (e.g. modifying an opinion, disagreeing). The results of the study reveal that the Swedish learners tend to use *well* in speech management functions, and particularly to plan ahead, significantly more frequently than the native speakers in a similar informal interview situation, which reflects their encoding difficulties. On the other hand, the learners are shown to significantly underuse the pragmatic marker in attitudinal functions, which seems to suggest that they may be largely unaware of the important interpersonal functions of *well* in interactions.

The last two papers of this special issue both deal with fluency and how it correlates with phenomena like accuracy or complexity. Christiane Brand and Sandra Götz investigate the possible correlation between the fluency (as measured through speech rate, as well as frequency and position of filled and unfilled pauses) and the accuracy (i.e. number of errors) of advanced German learners of English, using data from an error-tagged version of the German component of LINDSEI, as well as LOCNEC as a native control corpus. Their quantitative and qualitative analyses are combined with a study into English native speakers’ perception of the overall proficiency of some of these learners. While the results do not reveal any clear correlation between fluency and accuracy, they show that an average performance in both fluency and accuracy leads to a high level of perceived oral proficiency. From a methodological point of view, the paper makes an important contribution in that it presents a method aimed at jointly investigating a number of features of spoken interlanguage.

In the final paper, John Osborne triangulates temporal fluency, syntactic complexity and informational content (informativeness) in English and French spoken data from the
PAROLE corpus. His data represent three groups of speakers: low-fluency learners, high-fluency learners and native speakers. Each of the three phenomena under study is investigated through a combination of measures: speech rate, proportion of hesitation time and length of runs for fluency, and rate, density, condensation and subordination rate/granularity for syntax and informativeness. The analyses suggest that more fluent speakers can deliver content more rapidly (as is to be expected), but also more efficiently, by choosing syntactic units and lexical items that allow for a more economical packaging of the information. Less fluent speakers, by contrast, sometimes use long and syntactically complex structures in which the information becomes diluted (e.g. circumlocutions, “support” clauses or gradually elaborated constructions). They also tend to use fewer framing units, i.e. units that summarise a macro-event or serve as an introduction or conclusion to it, and they hesitate longer before these units.

Together, these papers show that corpus linguistics has a lot to contribute to the study of errors and disfluencies in (native and non-native) speech. Not only does it allow for a better description of the use – and usefulness – of such items, but it also makes it possible to investigate the correlation of fluency and/or accuracy with other phenomena, with a view to better characterising spoken language in general. There is no doubt that, with the growing availability of well-designed and well-transcribed corpora of naturally-occurring speech, errors and disfluencies, once thought of as mere “unsuccessful bits of language” (James 1998:1), will continue to arouse linguists’ curiosity and will eventually come to occupy the place they deserve.
Acknowledgements

We wish to thank Sebastian Hoffmann, chair of the ICAME 30 organising committee, for welcoming the idea of a pre-conference workshop on errors and disfluencies in spoken corpora and for helping us with the practical organisation of the workshop. Thanks are also due to the British Council, the CGRI-DRI and the FNRS for their financial support through the *Partnership Programme in Science*, as well as to the audience at the workshop for stimulating discussion. Finally, we are grateful to the reviewers who donated their valuable time to make this special issue possible, as well as to Sylviane Granger and Jennifer Thewissen for their useful feedback on this introductory paper.

References


Cheng, W. 2007. “‘Sorry to interrupt, but…’: Pedagogical implications of a spoken corpus”. In M. C. Campoy & M. J. Luzón (Eds.), Spoken Corpora in Applied Linguistics. Bern: Peter Lang, 199-215.


García-Amaya, L. 2009. “New findings on fluency measures across three different learning contexts”. In J. Collentine, M. García, B. Lafford & F. Marcos Marin (Eds.), *Selected


Corpus PAROLE. Available at: http://talkbank.org/BilingBank/ (accessed September 2010).


Computational Linguistics, 27 (1), 132-139.


30th annual conference of the *International Computer Archive of Modern and Medieval English (ICAME 30), Lancaster University, 27-31 May 2009.*


Mukherjee, J. 2000. “Speech is silver, but silence is golden: Some remarks on the function(s) of pauses”. *Anglia*, 118 (4), 571-584.


Müller, S. 2004. “‘Well you know that type of person’: Functions of *well* in the speech of American and German students”. *Journal of Pragmatics*, 36 (6), 1157-1182.


Editors' addresses

Gaëtanelle Gilquin
FNRS – Université catholique de Louvain
Centre for English Corpus Linguistics
Collège Erasme
Place Blaise Pascal, 1
B-1348 Louvain-la-Neuve
Belgium
gaetanelle.gilquin@uclouvain.be

Sylvie De Cock
Université catholique de Louvain
Centre for English Corpus Linguistics
Collège Erasme
Place Blaise Pascal, 1
B-1348 Louvain-la-Neuve
Belgium
sylvie.decock@uclouvain.be