fsca: French syntactic complexity analyzer

Nathan Vandeweerd

Université catholique de Louvain / Vrije Universiteit Brussel

This article reports on an open-source R package for the extraction of syntactic units from dependency-parsed French texts. To evaluate the reliability of the package, syntactic units were extracted from a corpus of L2 French and were compared to units extracted manually from the same corpus. The f-score of the extracted units ranged from 0.53-0.97. Although units were not always identical between the two methods, manual and automatically-derived syntactic complexity measures were strongly and significantly correlated ($\rho = 0.62$-0.97, $p < 0.001$), suggesting that this package may be a suitable replacement for manual annotation in some cases where manual annotation is not possible but that care should be used in interpreting the measures based on these units.

Keywords: L2 French, dependency grammar, syntactic complexity, automatic annotation,

open-source R package

1. Introduction

The development of syntactic complexity is a widely attested component of L2 proficiency (Ortega, 2003), the most common operationalizations of which include measures of length (e.g. mean length of clause), the ratio of one type of unit to another (e.g. clauses per sentence) and the diversity of syntactic units (e.g. standard deviation of clause length) (De Clercq & Housen, 2017; Wolfe-Quintero, Inagaki, & Kim, 1998). In order to calculate these measures, texts must first be annotated for syntactic units at various levels of granularity (e.g. sentence, T-unit, clause, phrase), a process which is labor-intensive, especially considering the multiple annotators required to ensure reliability. Fortunately, for those researchers working on L2 English, automatic tools have been developed that can annotate and calculate syntactic complexity measures in learner texts. One such example is Lu's (2010) L2 Syntactic Complexity Analyzer (L2SCA), which uses the Stanford Parser (Klein & Manning, 2003) to segment, tokenize, part of speech (POS) tag and syntactically parse a text in terms of syntactic dependencies. A series of search expressions are then used to query the syntactic parse for syntactic units. Lu (2010) showed that there was a large degree of overlap between the manually and automatically identified units, with f-scores[1] ranging from 0.846 to 1.000. Moreover, there were strong correlations between syntactic complexity measures derived from the two identification methods. The use of such automatic tools is now common in studies of L2 English (Kyle & Crossley, 2018; e.g. Lu, 2011).

[1] F-scores represent the balance between precision (not identifying too many incorrect units) and recall (not missing too many units that were identified by the manual annotation).

For languages other than English, the availability of such automatic tools is much more limited and texts must therefore be manually annotated for syntactic units. As shown in Table 1, this has meant that previous studies of L2 French have often needed to strike a balance between sample size (and text length) on the one hand and the number of syntactic units on the other (Benevento & Storch, 2011; Bernardini & Granfeldt, 2019; De Clercq & Housen, 2017; Gyllstad, Granfeldt, Bernardini, & Källkvist, 2014; Kuiken & Vedder, 2008; Way, Joiner, & Seaman, 2000). This is unfortunate, given the importance of measuring complexity at various syntactic levels (see Norris & Ortega, 2009). What is also noticeable is that interrater reliability is only rarely reported and when it is reported (as in Benevento & Storch, 2011), very few details are provided about the process of double annotation. As such, there is currently no agreed-upon standard as to an acceptable level of reliability for syntactic annotation in L2 French.

Table 1. Previous studies of syntactic complexity in L2 French

	Number of Texts					Manually Annotated Units						
Study	EN	IT	FR	Total	Length	AS	T	C	DP	CC	NP	IRR
Benvento and Storch (2011)	-	-	45	45	2-300†		x	x				100%
Bernardini and Granfeldt (2019)	20	20	20	60	n.r.		x	x	x			n.r.
De Clercq and Housen (2017)	100	-	100	200	n.r.	x		x	x	x	x	n.r.
Gyllstad et al. (2014)	108	56	76	240	n.r.		x	x	x			n.r.
Kuiken and Vedder (2008)	-	162	246	408	>150†		x	x	x			n.r.
Way et al. (2000)	-	-	937	937	47-118‡		x					n.r.

AS=AS-Unit; T=T-unit, C=Clause; DP=Dependent Clause; CC=Coordinated Clause; NP=Noun Phrase; IRR = interrater reliability of syntactic annotation; n.r.=not reported
†According to instructions in writing prompt
‡The range of mean text lengths reported per group

The impetus for the development of an automatic tool for L2 French was a separate project which investigated various domains of linguistic complexity in French learner

writing (Vandeweerd, Housen, & Paquot, this issue) Following the recommendation of Norris and Ortega (2009), we wanted to investigate syntactic complexity at the T-unit, clause and phrase level but given the size of the corpus, manual annotation of those units was not feasible. Fortunately, because the learner texts had already been dependency parsed, it was possible to write a script in R (R Core Team, 2019) to extract syntactic units from the grammatical dependencies. The scripts were then compiled into a package (*fsca*) which is now available to download from github.[2] This article describes the process that was used to evaluate the reliability of syntactic complexity measures based on the units that were extracted automatically with this tool. As such, it serves as an accompanying method report to Vandeweerd et al. (this issue).

2. Methodology

In order to evaluate the automatic extraction tool, a set of learner data was first annotated for syntactic units (Section 2.1). These manual annotations served as a "gold-standard" against which the automatic extraction method was compared (Section 2.2). Reliability was evaluated by calculating the precision, recall and f-score for each syntactic unit (Section 3.1) and by correlating syntactic complexity measures calculated on the basis of manual versus automatically extracted units (Section 3.2).

2.1 Manual annotation

The source of learner data was the *Leerdercorpus Frans* (Demol & Hadermann, 2008; Vanderbauwhede, 2012), a corpus of texts written by university-level Dutch learners of

[2] *https://github.com/nvandeweerd/fsca*

French in Belgium. In the context of the larger project (Vandeweerd, Housen, & Paquot, this issue), the subcorpus of argumentative essays was rated by two trained language assessors and all texts (n=251) were found to range from level B2 to C2 of the Common European Framework of Reference (Council of Europe, 2001). The original version of the corpus contains XML tags around each sentence. Three sets of segments were extracted on the basis of these tags: a set of 60 segments for training annotators, a set of 100 segments for testing interrater reliability and a final set of 400 segments for testing the automatic extraction. In order to evaluate on as diverse a sample as possible, these segments were randomly extracted from the argumentative subcorpus as a whole. The annotators were Master's students who are native speakers of French with experience in syntactic annotation from their coursework in linguistics and translation. They were trained to identify the following units: sentences,[3] clauses, coordinated clauses, dependent clauses, T-units, noun phrases and verb phrases using a set of guidelines (see definitions in Appendix). Mistakes in the identification of units were discussed with the lead investigator and the guidelines were further refined during these meetings. After training, a second set of 100 sentences was used to test the interrater reliability of the manual annotation. Table 2 lists the percentage agreement for each syntactic unit as well as Scott's (1955) π, which applies a correction for chance agreement. The interrater agreement (π) for all units except dependent clauses and coordinated clauses was found to be above the minimally acceptable

[3] Following the extraction of the segments, it became apparent that the segments which had been pre-tagged as "sentences" in the original corpus sometimes contained more than one sentence according to our definition (see Appendix). Because of this, annotators were asked to manually identify sentences within each extracted segment.

estimate of reliability for second language research (0.83) as recommended by Plonsky and Derrick (2016). The lowest levels of reliability were found for dependent clauses ($\pi = 0.77$) and coordinated clauses ($\pi = 0.57$). In most cases, disagreements on these units were due to inconsistent annotation (e.g. annotating only one of two coordinated clauses in a sentence or annotating the matrix clause separately from the dependent clause embedded within) rather than lack of knowledge about the segments themselves. Following the calculation of interrater reliability, all cases of disagreement were discussed and resolved and the guidelines were further refined, with a particular focus on dependent and coordinated clauses as they had been the largest source of disagreement.

Table 2. Interrater agreement for manual annotation of syntactic units

Unit	Total	Agree	Disagree	%	π
Sentences	102	102	0	1.00	1.00
Clauses	203	194	9	0.96	0.94
Dependent Clauses	65	54	11	0.83	0.77
Coordinated Clauses	38	26	12	0.68	0.57
T-units	139	128	11	0.92	0.89
Noun Phrases	458	417	41	0.91	0.88
Verb Phrases	278	256	22	0.92	0.89

The annotators then worked separately to annotate the final set of 400 sentences. All cases of disagreement were resolved in meetings with the lead investigator, who also checked the final list for errors. The combined set of 500 segments (interrater reliability set plus the 400 set) was then used as a gold standard to evaluate the R package described in Section 2.2. All annotations were carried out using Dexter Coder (Version 0.6.4, Garretson, 2011), a program in which texts are annotated by highlighting segments in different colors. The highlighted segments were then extracted as an XML file for comparison to the automatically extracted units.

2.2 Automatic extraction of syntactic units

In the context of the larger study (Vandeweerd, Housen, & Paquot, this issue), the learner texts were POS tagged with the MElt Tagger (Denis & Sagot, 2012) and parsed with Malt Parser (Nivre, Hall, & Nilsson, 2006) trained on the French Tree Bank (Abeillé & Barrier, 2004), a parsing method which has been shown to have a high level of accuracy (87% labeled attachment) on L1 data (Candito, Nivre, Denis, & Anguiano, 2010). The choice of these tools was necessitated by the requirement to use the same processing chain as was used for the reference corpus in that study. The output of Malt Parser is a text file containing a list of tables, each corresponding to one sentence. Each table lists the tokens, the lemmas, the part of speech, the position of the word in the sentence, the type of dependency relationship and the dependency relationships between the words in the sentence (Table 3). Each of the sentences segmented by Malt parser were manually aligned with the manually identified segments by assigning the corresponding code from the XML manual annotation file.

Table 3 . Example of a Dependency Parsed Sentence

Token	Lemma	Part of Speech	Position	Dependency Relation	Dependent On
C'	ce	PRO:DEM	1	suj	2
est	être	VER:pres	2	root	0
un	un	DET:ART	3	det	4
point	point	NOM	4	ats	2
très	très	ADV	5	advmod_ADJ	6
important	important	ADJ	6	amod	4
.	.	SENT	7	ponct	2

The *getUnits()* function from the *fsca* package extracts syntactic units by first getting a list of the relevant node words for a given unit (e.g. nouns for noun phrases) and then extracting all of the dependencies on each of the node words. Additional cleaning

steps are then performed (e.g. ensuring there is a finite verb, removing punctuation etc.) The output of the function is a list of syntactic units including the number of units, the length of each unit and the tokens belonging to each unit. The following syntactic units are extracted by the function: sentences, clauses, dependent clauses, coordinated clauses, T-units, noun phrases and verb phrases. An example output of the function is provided below. Space does not permit a more detailed explanation but the reader can consult the package documentation for more information about the extraction of specific units.

```
#Example sentence
manual.sents[["b.208.1"]]
```

[1] "Ils prétendent qu'il est impossible de rééduquer un tel jeune criminel."

```
#Extraction of dependent clauses
getUnits(test.sents[["b.208.1"]],
         what = "tokens",
         units = c("DEP_CLAUSES"),
         paste.tokens = TRUE)
```

$DEP_CLAUSES
$DEP_CLAUSES[[1]]
[1] "qu' il est impossible de rééduquer un tel jeune criminel"

3. Results

3.1 Precision and recall of automatically identified units

To evaluate the reliability of the automatic method, precision, recall and f-scores for each syntactic unit were calculated according to the formulas used in Lu (2010: 486). As shown in Table 4, although the number of units identified by each method was quite similar, the number of identical units identified by the two methods varied across the unit types. Lower f-scores were found for dependent clauses (0.60) and coordinated clauses (0.53) and higher

f-scores were found for clauses (0.74), T-units (0.83), sentences (0.97), noun phrases (0.84) and verb phrases (0.78).

Table 4 . Precision and recall of automatically extracted units

Unit	Manual	Automatic	Identical	Precision	Recall	F-score
Sentences	517	519	500	0.97	0.96	0.97
Clauses	903	871	652	0.72	0.75	0.74
Dependent Clauses	325	295	187	0.58	0.63	0.60
Coordinated Clauses	153	144	78	0.51	0.54	0.53
T-units	581	576	478	0.82	0.83	0.83
Noun Phrases	2,277	2,278	1,903	0.84	0.84	0.84
Verb Phrases	1,328	1,317	1,030	0.78	0.78	0.78

3.2 Correlation between manual and automatic methods

Eleven syntactic complexity measures were computed on each of these segments using both the manual and automatically identified units. These measures target complexity at the sentence, T-unit, clause and phrase-level in terms of mean length, number of embedded units and standard deviation (see Vandeweerd, Housen, & Paquot, this issue). Pearson's r was calculated for all normally distributed measures (tested using Shapiro Wilks tests) and Spearman's rho (ρ) was calculated for all non-normally distributed measures. These are provided in Table 5. All correlations were found to be significant ($p < 0.001$) and are considered large according to Plonsky and Oswald's reccomended guidelines for L2 research (Plonsky & Oswald, 2014).

Table 5 . Correlation between manual- and automatic-based measures

Measure	Measure Description	ρ	r	*p*
MLS	Mean length of sentence	0.972	-	***
DIVS	Standard deviation of sentence length	-	0.961	***
T_S	T-units per sentence	0.645	-	***
MLT	Mean length of T-unit	0.871	-	***
DIVT	Standard deviation of T-unit length	0.620	-	***
C_T	Clauses per T-unit	0.823	-	***
MLC	Mean length of clause	0.887	-	***
DIVC	Standard deviation of clause length	0.759	-	***
MLNP	Mean length of noun phrase	0.720	-	***
DIVNP	Standard deviation of noun phrase length	0.713	-	***

Measure	Measure Description	ρ	r	*p*
NP_C	Noun phrases per clause	0.892	-	***

****p < 0.001*

3.3 Sources of error

Each case of disagreement between the manual and automatic methods was annotated as being due to: incorrect segmentation (Seg.), an erroneous part of speech tag (POS), an incorrect dependency label or incorrect dependency relation (Dep.), a learner error (Learner) or an error due to scripts in the package (Fun.). The tabulation of these errors is provided in Table 6.

Table 6 . Errors in automatic annotation

	Preprocessing					
Unit	Seg.	POS	Dep.	Learner	Fun.	Total
Sentences	13	-	-	3	-	16
Dependent Clauses	1	3	81	9	52	146
Coordinated Clauses	5	1	31	3	21	61
Noun Phrases	1	23	331	12	74	441
Verb Phrases	1	9	143	11	98	262
Total	21	36	586	38	245	926

N.B. Errors in T-units and clauses were not annotated separately as they necessarily include the errors in the embedded units.

As shown in Table 6, the majority of the errors occured during pre-processing (segmentation, POS tagging and dependency parsing). At the segmentation stage, some sentences were split at non-sentence boundaries (e.g. semi-colons, ellipses) which caused otherwise contiguous segments to be analyzed separately. POS-tagging errors also caused problems because the root nodes, which form the basis of the identification of syntactic units, could not be found. For example, mis-tagging the noun *jeune* (‘young/teenager’) as an adjective, meant that noun phrases which were dependent on *jeune* were not identified. This also caused problems for other units in which the noun phrase was embedded. The most common type of error was due to incorrect assignment of dependency relations

(e.g. prepositional phrases that were analyzed by the parser as being dependent on the immediately preceding noun instead of the verb that they modify). This often caused a cascade of problems which meant that several units were misidentified or unidentified.

In addition, although the learner texts in the corpus are fairly advanced (B2-C2), there were still cases where learner errors caused problems for the parser. These included: punctuation errors (e.g. lack of space after a period), spelling errors (e.g. *son* ('his/her') for *sont*, 'be-3PL.PRES') and the omission of obligatory elements (e.g. lack of a finite verb in a dependent clause). Each of these errors caused POS-tagging and parsing errors which led to the misidentification of syntactic units by the function.

About a quarter of the errors were due to the current limitations of the function itself. Some particular structures that cause issues for the function include: verb phrases with adverbs between the auxiliary verb and the main verb; coordinated phrases; dependent clauses directly dependent on prepositions or adverbial interrogatives; non-nouns preceded by a determiner including superlatives and the non-functional determiner *l'* in *que l'on*; coordinated clauses with a pronoun as the subject of the verb; noun phrases modified by both a prepositional phrase and a relative clause; imperative clauses; exclamatory clauses headed by a subordinating conjunction and inverted comparatives (e.g. *Aussi divergeant que les affirmations sont les approches possibles.*). It is important to note however that not all cases of disagreement between the manual and automatic methods are necessarily equally problematic. For example, Malt Parser analyzes both words in multi-word conjunctions (e.g. *tandis que*; 'while') as modifiers of the main verb and therefore only *que* is captured as part of the dependent clause. As a result, multi-word conjunctions are often

analysed by the function as belonging to the main (or in the case of coordinated clauses, left) clause by the function instead of the dependent (or right) clause. In these cases, the only difference between the units identified by the two methods is the location of the conjunction. This may explain why complexity measures such as the mean length of clause are strongly correlated between the manual and automatic methods despite the low f-scores found for coordinated and dependent clauses.

4. Discussion and conclusion

The aim of this article was to determine the reliability of syntactic complexity measures calculated on the basis of syntactic units extracted using the *fsca* package. To answer this question, a sample of segments was extracted from a learner corpus of French and was manually annotated for the presence of six syntactic units. The manually annotated units were compared to a list of automatically extracted units. The f-score, which represents the balance between precision and recall, was found to range from 0.53-0.97 and correlations between manually- and automatically-based syntactic complexity measures were found to range from 0.62 to 0.97. In other words, the units extracted by the *fsca* package did not always match up identically with the units identified manually but the measures calculated on the basis of the units were still strongly correlated with manually-based measures. When compared to Lu's (2010) tool for L2 English (L2SCA), the reliability reported here is noticeably lower (Lu reports f-scores of 0.83-1.00 and correlations of 0.83-0.94). The question is whether this tool can be considered reliable, given the strength of these correlations.

While standard thresholds have been suggested for other types of reliability such as inter-item, interrater and intrarater (Brown, 2014; see Landis & Koch, 1977; Plonsky & Derrick, 2016; Shrout, 1998), there is currently no agreed-upon threshold for the reliability of automatic annotation in the field of learner corpus research. Among studies of syntactic complexity in L2 French more specifically, only one study has reported a coefficient of reliability (and only interrater reliability) for this type of annotation. Benevento and Storch (2011), report 100% interrater agreement on the annotation of T-units and clauses, two units which were also found to have a high level of interrater reliability (π=0.89; π=0.94) as well as high f-scores in the automatic extraction (0.83 and 0.74) in this study. However, the authors do not report how many texts were double annotated and so it is difficult to directly compare these results. In addition, whether such high levels of agreement would also be found for other units such as dependent and coordinated clauses is unknown but perhaps doubtful given the low level of agreement obtained by the annotators in this study. It is also important to note that manual annotation is not necessarily unproblematic either. The results presented here suggest that even high levels of interrater reliability may be difficult to achieve for some types of syntactic units (especially dependent and coordinated clauses). This is especially troubling considering the fact that few studies report coefficients of interrater reliability for syntactic annotation.

In the absence of general thresholds of reliability for automatic annotation or comparable studies of syntactic annotation in L2 French, it may be more useful to consider each of the measures as containing more or less noise due to machine-induced error. Whether to use such measures ultimately depends on the amount of error that a researcher

is willing to accept. In the current study, measures with the highest correlations (above .90) such as MLS and DIVS involve less noise while measures with lower correlations such as DIVT (ρ = 0.620) and T_S (ρ = 0.645) involve more noise and therefore more caution should be used when interpreting results based on these measures.

In order to reduce such measurement error, it is vital that we continue to refine the accuracy of natural language processing tools used in Learner Corpus Research. This study serves as a reminder that there is an element of error at each step of processing a learner text. While individual errors may not be detrimental on their own, as our processing techniques become more advanced, we need to ensure more than ever that the tools we use are as accurate as possible (see the recent special issue on this topic). This study revealed that even commonly used techniques such as sentence segmentation and POS tagging still require fine-tuning. It may be beneficial therefore to test the reliability of newer, state of the art natural language processing tools (e.g. spaCy 2, Honnibal & Montani, 2017) on learner data. However, as noted by an anonymous reviewer, it is also important to remember that these tools need to be tested on a variety of situations (e.g. L1/L2 data, corrected/uncorrected texts, lower/higher proficiency learners etc.) in order to obtain a clear picture of their accuracy for a given purpose. While that is beyond the scope of this article, this recommendation would undoubtedly serve the field. That being said, automatic annotation tools for languages other than English are sorely needed and this open-source package is shared with the hope that it will be further refined through collaboration between researchers willing to broaden the scope of L2 French learner corpus research beyond what is possible with manual annotation alone.

5. Disclosures

The *fsca* package described in this article, as well as the helper functions required to use it are available at *https://github.com/nvandeweerd/fsca*. The function was created using R Studio (Version 1.2.1335, RStudio Team, 2018) and R (Version 3.6.1, R Core Team, 2019) as well as the following packages: *igraph* (Csardi & Nepusz, 2006) and *purrr* (Henry & Wickham, 2020).

6. Acknowledgements

This work is indebted to Hubert Naets (Centre de traitement automatique du langage, Université catholique de Louvain) for his help in parsing the learner corpus as well as to Héloïse Copin and Anthonya Delfosse for their help manually annotating the syntactic units. In addition to the anonymous reviewers, I would also like to thank Alex Housen and Magali Paquot for their constructive comments and suggestions at various stages of this research. This work was supported by the Fonds de la Recherche Scientifique (FNRS) under Grant n° T.0086.18.

7. References

Abeillé, A., & Barrier, N. (2004). Enriching a French treebank. *LREC*, 2233–2236.

Benevento, C., & Storch, N. (2011). Investigating writing development in secondary school learners of French. *Assessing Writing*, *16*(2), 97–110. *https://doi.org/10.1016/j.asw.2011.02.001*

Bernardini, P., & Granfeldt, J. (2019). On cross-linguistic variation and measures of linguistic complexity in learner texts : Italian, French and English. *International Journal of Applied Linguistics*, *29*(2), 211–232. *https://doi.org/10.1111/ijal.12257*

Brown, J. D. (2014). Classical theory reliability. In A. J. Kunnen (Ed.), *The companion to language assessment* (pp. 1165–1181). Oxford: Wiley-Blackwell.

Candito, M., Nivre, J., Denis, P., & Anguiano, E. H. (2010). Benchmarking of statistical dependency parsers for French. *COLING 2010: Poster volume*, 108–116. Beijing.

Council of Europe. (2001). *The common european framework of reference for languages : Learning, teaching, assessment*. Cambridge: Council of Europe; Cambridge University Press.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal (Complex Systems)*.

De Clercq, B., & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *Modern Language Journal*, *101*(2), 315–334. *https://doi.org/10.1111/modl.12396*

Demol, A., & Hadermann, P. (2008). An exploratory study of discourse organisation in French L1, Dutch L1, French L2 and Dutch L2 written narratives. In *Linking up contrastive and learner corpus research* (pp. 255–282). *https://doi.org/10.1163/9789401206204_011*

Denis, P., & Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, *46*(4), 721–736. *https://doi.org/10.1007/s10579-012-9193-0*

Garretson, G. (2011). *Dexter Coder*. Retrieved from *http://www.dextercoder.org/*

Gyllstad, H., Granfeldt, J., Bernardini, P., & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook*, *14*, 1–30. *https://doi.org/10.1075/eurosla.14.01gyl*

Henry, L., & Wickham, H. (2020). *purrr: Functional programming tools*. Retrieved from *https://cran.r-project.org/package=purrr*

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Klein, D., & Manning, C. (2003). Fast exact inference with a factored model for natural language parsing. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 3–10). Cambridge MA: MIT Press.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, *17*(1), 48–60. *https://doi.org/10.1016/j.jslw.2007.08.003*

Kyle, K., & Crossley, S. A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *Modern Language Journal*, *102*(2), 333–349. *https://doi.org/10.1111/modl.12468*

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. *https://doi.org/10.2307/2529310*

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. *https://doi.org/10.1075/ijcl.15.4.02lu*

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62. *https://doi.org/10.5054/tq.2011.240859*

Nivre, J., Hall, J., & Nilsson, J. (2006). MaltParser : A data-driven parser-generator for dependency parsing. *LREC 2006*, 2216–2219.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. *https://doi.org/10.1093/applin/amp044*

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college level L2 writing. *Applied Linguistics*, *24*(4), 492–518.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, *100*(2), 538–553. *https://doi.org/10.1111/modl.12335*

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. *https://doi.org/10.1111/lang.12079*

R Core Team. (2019). *R: A language and environment for statistical computing*. Retrieved from *https://www.r-project.org/*

RStudio Team. (2018). *RStudio: Integrated Development Environment for R*. Retrieved from *http://www.rstudio.com/*

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly*, *19*(3), 321–325.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*(3), 301–317. *https://doi.org/10.1191/096228098672090967*

Vanderbauwhede, G. (2012). *Le déterminant démonstratif en français et en néerlandais à travers les corpus : Théorie, description, acquisition [The demonstrative determiner in French and Dutch corpora: Theory, description and acquisition]* (PhD thesis). (Unpublished Doctoral Dissertation). Katholieke Universiteit Leuven; Université Paris Ouest Nanterre La Défense.

Vandeweerd, N., Housen, A., & Paquot, M. (this issue). Applying phraseological complexity measures to L2 French: A partial replication study. *International Journal of Learner Corpus Research*.

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *Modern Language Journal*, *84*(2), 171–184. *https://doi.org/10.1111/0026-7902.00060*

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Honolulu: Second Language Teaching & Curriculum Center.

Appendix: Definitions of Syntactic Units

1. Sentences

Following Lu (2010: 481) we defined a sentence as "a group of words delimited by one of the following punctuation marks that signal the end of a sentence: period, question mark, exclamation mark, quotation mark or ellipsis."

2. Clauses

Clauses are defined as structures with a subject and a finite verb (Hunt, 1965). This includes all T-units (see Section 3) as well as all dependent clauses (see Section 2.1) embedded within the T-units in a given sentence.

2.1 Dependent clauses

Dependent clauses are clauses which are semantically and/or structurally dependent on a super-ordinate syntactic structure. They include nominal clauses, adverbial clauses and adjectival clauses (Hunt, 1965; Lu, 2010). They must contain a finite verb and a subject.

2.1.1 Special cases

- Clauses of the type 'il y a' are considered dependent clauses only if *a* is directly dependent on a finite verb or if there is a finite verb dependent on *a*. This means that adverbial clauses which function like 'ago' in English (e.g. *il y a deux ans…*; 'two years ago…') are considered dependent clauses but simple declaratives (e.g. *il y a une maison*; 'there is a house') are not.
- Direct interrogatives in the form *est-ce que* are not considered the head of subordinate clauses. Rather, the head of a clause is the finite verb dominated by *est* as in interrogatives formed by inversion.
- Citations or reported speech enclosed with French guillmets («»), single or double quotation marks are also considered subordinate clauses.

2.2 Coordinated clauses

Coordinated clauses are clauses which are not semantically and/or structurally dependent on a super-ordinate syntactic structure but are conjoined to one or more clauses of syntactically equal status. They may be joined by a coordinating conjunction (e.g. *et*; 'and'), punctuation (e.g. semi-colon, colon, comma) or by juxtaposition and must contain both a subject and a finite verb.

3. T-units

We use Hunt's (1970: 199) definition of a T-unit as "one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it." Identifying T-units therefore depends on the identification of coordinated clauses since a sentence can only contain multiple T-units if it contains multiple coordinated clauses. A sentence that does not contain any coordinated clauses simply has one T-unit, provided it has at least one finite verb. Consistent with Hunt (1965), we do not classify sentence fragments (clauses without a finite verb) as T-units. Therefore, if a sentence has no coordinated clauses (and one finite verb) it has one T-unit.

4. Noun phrases

We use Lu's (2010) definition of a noun phrase as a *complex nominal* (see Cooper, 1976) which includes: nouns plus adjective(s), possessive(s), prepositional phrase(s), relative clause(s), participle(s), or appositive(s), nominal clause(s). We also include words (nouns, adverbs and pronouns) that have a determiner (e.g. *une maison* 'a house'; *cet autre* 'this other') in our definition. Following Lu (2010) and Cooper (1976), we also include gerunds and infinitives in subject position.

5. Verb phrases

As in Lu (2010) we count both finite and non-finite verb phrases. Auxiliary verbs do not constitute their own verb phrase but are considered part of the main verb they modify. However, verb phrases with modal verbs are considered separate verb phrases. When two verbs are coordinated they are also considered a singular verb phrase (e.g. *ne sont pas d'accord et présentent de solutions différents*; 'do not agree and present different solutions').

6. References

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *Journal of Educational Research*, *69*(5), 176–183. *https://doi.org/10.1080/00220671.1976.10884868*

Hunt, K. (1965). *Grammatical structures written at three grade levels*. Champaign, IL: NCTE.

Hunt, K. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly1*, *4*(3), 195–202.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. *https://doi.org/10.1075/ijcl.15.4.02lu*