

CBTS meets readability research: New methodological insights for the study of the simplification hypothesis

Thomas François¹ & Marie-Aude Lefer²

UCLouvain

thomas.francois@uclouvain.be¹, marie-aude.lefer@uclouvain.be²

Ever since the publication of Laviosa's (1998a, 1998b) pioneering work, the study of lexico-syntactic simplification has held centre stage in corpus translation research concerned with the typical features of translated text (see e.g. Corpas Pastor et al. 2008, Grabowski 2013, Kajzer-Wietrzny 2015, Ferraresi et al. 2018, Lv & Liang 2019). Lexico-syntactic simplification can be defined as translators' tendency to produce target texts that are less informationally dense, less lexically varied and/or sophisticated, and less syntactically elaborate than comparable texts in the same language that have been produced in *unmediated* circumstances, i.e. in situations of monolingual text production (cf. Bernardini et al. 2016: 64-65). While empirical evidence of simplification has been found in different translated languages, including non-European ones, and translation modalities (written translation, consecutive and simultaneous interpreting), methodological advancement in the area has been rather modest. To date, corpus translation scholars have mainly relied on Laviosa's linguistic operationalizations of simplification (lexical density, core vocabulary coverage, list head coverage and mean sentence length) without providing aggregate simplification profiles of translated texts. In the present paper, we propose an innovative approach to the study of simplification in translation that aims to move beyond this set of operationalizations. Specifically, the approach draws on insights from readability research, which has recently undergone major advances under the influence of machine learning and natural language processing (NLP) (cf. Benjamin 2012, Collins-Thompson 2014). This paradigm offers robust, sophisticated analytical models with which to investigate the simplicity/complexity spectrum in language. In particular, NLP-informed readability studies rely on a wide range of simplification parameters, which are

more likely to capture text dimensions that might be overlooked by shallow parameters, and make use of advanced statistical methods to aggregate these parameters.

In this study, we use the translated and original proceedings of the European Parliament as a test case. More specifically, we rely on two French subcorpora of *Europarl-direct* (Cartoni and Meyer 2012): (i) an original French subcorpus made up of 1,880 speeches delivered by 192 speakers (630,000+ tokens) and (ii) a French-translated-from-English subcorpus containing 5,257 speeches given by 237 speakers (1.5 million+ tokens). Both subcorpora have been POS-tagged with the TreeTagger (Schmid 1995). One of the reasons for choosing Europarl is its availability and multilingualism, which make it possible to enlarge the empirical foundation of the approach presented here by replicating it on other Europarl datasets. The simplification analyses are based on François's (2011) readability model, which includes both classic and NLP-enabled simplification parameters. For the present analyses, we have selected the 19 most relevant lexical, syntactic and discursive parameters from François's set (see Table 1). The selection has been guided by previous research in both CBTS and NLP-informed readability studies. We have chosen parameters that have proved useful in previous readability research and that may be meaningful in translation research, alongside more traditional simplification indicators à la Laviosa.

We apply a twofold approach in the statistical analysis. First, drawing on Laviosa's methodology, we analyze all variables separately in order to detect simplification effects at the variable level. First, we use Wilcoxon rank sum tests to compare the means of each parameter in the original French (OF) and translated French (TF) conditions. In addition, to better characterize the size of the effect of translation on simplification, we compute point-biserial correlation coefficients between the two conditions (OF and TF) and each simplification parameter. Second, we use multivariate analyses so as to consider the effect of translation on all the simplification parameters at once, offering a more comprehensive way of testing the simplification hypothesis. In particular, we apply a principal component analysis (PCA) transformation of the 19 parameters, which allows us to deal with multicollinearity issues. With

the help of the PCA, we reduce the analysis to the three most explanatory components, which together explain 61% of the variance. The first component encodes lexical frequency information, the second is a mix of lexical and syntactic information and the third corresponds to word length. In view of the fact that the three components are independent, we apply a Wilcoxon rank sum test on each component to compare the two conditions (OF and TF).

Table 1: Lexical, syntactic and discursive simplification parameters examined in the study

Parameter type	Simplification parameter	Description
<i>Lexical</i>	TTR-L	Lexical variety index 1: ratio of the number of types to the number of tokens (based on lemmas)
	NormTTR-L	Lexical variety index 2: type-token ratio, based on lemmas and normalized per 100 words
	LEX/GRAM	Lexical density index 1: ratio of lexical words to grammatical words
	LEX/ALL	Lexical density index 2: ratio of lexical words to all running words
	ConcDens	Estimate of conceptual density as defined by Kintsch et al. (1975) and computed with <i>Densidées</i> (Lee et al. 2010)
	MWL	Mean word length: average number of letters per word
	PW10	Proportion of words of 10 letters or more
	Syllper100	Number of syllables per 100 words (see François and Miltsakaki 2012)
	CVC200 CVC1000 CVC2000 CVC5000	Core vocabulary coverage: percentage of lemmas found in the top-frequency list extracted from the web-crawled frTenTen reference corpus (cf. Jakubiček et al. 2013). We calculated the variables on the basis of four different list sizes: top 200,

		1000, 2000 and 5000 most frequent lemmas in frTenTen
	GMLF	Geometric mean of lemma frequencies
	75LF	75 th percentile of the probability distribution of lemmas per speech
	90LF	90 th percentile of the probability distribution of lemmas per speech
<i>Syntactic</i>	MSL	Mean sentence length: average number of words per sentence
	%LongSent	Percentage of sentences that are longer than 30 words (cf. Daoust et al. 1996)
<i>Discursive</i>	PRO/NAM	Ratio of pronouns to proper names
	PRO/NOM+NAM	Ratio of pronouns to nouns (common nouns and proper names)

The results show that the simplification hypothesis is largely confirmed: translated texts are found to be simpler, both lexically and syntactically, than original texts. Lexically, for instance, translations are less dense (LEX/GRAM, LEX/ALL), contain fewer words of 10+ letters (PW10) and rely more on high-frequency words (all CVC variables, with CVC5000 being the most powerful parameter). As regards syntax, we find that sentences are shorter in translations (MSL) and that there are fewer sentences of 30+ words in translations (%LongSent). Interestingly, the two pronoun-noun ratios we used point to discursive complexification in translation. This trend might be linked to a well-known cross-linguistic contrast between French and English, French being more nominal than English. In the presentation, we will zoom in on a few key methodological takeaways for the study of simplification in CBTS.

References

- Benjamin, R.G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review* 24(1): 63-88.
- Bernardini, S., Ferraresi, A., & Milicevic, M. (2016). From EPIC to EPTIC - Exploring simplification in interpreting and translation from an intermodal perspective. *Target* 28: 61-86.

- Cartoni, B. & Meyer, Th. (2012). Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, May 2012, Istanbul, Turkey.
- Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics* 165(2): 97-135.
- Corpas Pastor, G., Mitkov, R., Naveed, A., et al. (2008). Translation universals: do they exist? A corpus-based NLP study of convergence and simplification. In *MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*. (AMTA2008: The Eighth Conference of the Association for Machine Translation in the Americas, Waikiki, 21-25 October 2008). Stroudsburg: Association for Machine Translation in the Americas, 75-81.
- Daoust, F., Laroche, L., & Ouellet, L. (1996). SATO-CALIBRAGE: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue Québécoise de Linguistique* 25(1): 205-234.
- Ferraresi, A., Bernardini, S., Milicevic Petrovic, M., & Lefer, M.-A. (2018). Simplified or not simplified? The different guises of mediated English at the European Parliament. *Meta* 63(3): 717-737.
- François, Th. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Doctoral dissertation, unpublished. Louvain-la-Neuve: Université catholique de Louvain.
- François, Th. & Mitsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? In: *NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations* (PITR 2012) (Montreal, Canada, June 7, 2012), 49-57.
- Grabowski, Ł. (2013). Interfacing corpus linguistics and computational stylistics. Translation universals in translational literary Polish. *International Journal of Corpus Linguistics* 18(2): 254-280.
- Jakubicek, M., Kilgarrieff, A., Kovar, V., Rychly, P., & Suchomel, V. (2013). The TenTen corpus family. In: *7th International Corpus Linguistics Conference*, 125-127.
- Kajzer-Wietrzny, M. (2015). Simplification in interpreting and translation. *Across Languages and Cultures* 16(2): 233-255.
- Kintsch, W., Kozminsky, E., Streby, W., McKoon, G., & Keenan, J. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior* 14(2): 196-214.
- Laviosa, S. (1998a). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43(4): 557-570.
- Laviosa, S. (1998b). The English Comparable Corpus: a Resource and a Methodology. In L. Bowker, M. Cronin, D. Kenny & J. Pearson (eds.) *Unity in Diversity? Current Trends in Translation Studies*. Manchester: St. Jerome Publishing.
- Lee, H., Gambette, Ph., Maille, E., & Thuillier, C. (2010). *Densidées : calcul automatique de la densité des idées dans un corpus oral*. In *RECITAL'2010 : 12ième Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*. (Montréal, Canada, 19-22 July 2010).

- Lv, Q. & Liang, J. (2019). Is consecutive interpreting easier than simultaneous interpreting? – a corpus-based study of lexical simplification in interpretation. *Perspectives* 27(1):91-106.
- Schmid, H.L. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*. (Dublin, Ireland).