Meta
Journal des traducteurs
Translators' Journal

META

Revisiting simplification in corpus-based translation studies: Insights from readability research

Thomas François et Marie-Aude Lefer

Volume 67, numéro 1, avril–mai 2022

Pour de nouvelles méthodes en traductologie quantitative
Exploring New Methods in Quantitative Translation Studies

URI : https://id.erudit.org/iderudit/1092190ar
DOI : https://doi.org/10.7202/1092190ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

François, T. & Lefer, M.-A. (2022). Revisiting simplification in corpus-based translation studies: Insights from readability research. *Meta*, *67*(1), 50–70. https://doi.org/10.7202/1092190ar

Résumé de l'article

Depuis les travaux de Laviosa (1998a ; 1998b), les études de corpus se sont souvent penchées sur le phénomène de simplification lexico-syntaxique, afin de déterminer si les textes traduits sont plus simples que les textes non traduits. Laviosa (1998a ; 1998b) traite également de l'hypothèse de convergence, selon laquelle les textes traduits sont plus homogènes que les textes non traduits. Jusqu'à présent, cette question a cependant suscité moins d'intérêt que la simplification. En traductologie de corpus, la simplification a été opérationnalisée à l'aide de cinq paramètres principaux. Sur la base de ceux-ci, les études ont montré que la simplification varie en fonction des modalités de traduction, des paires de langues et des registres analysés. Notre article a pour objectif de revisiter ce type de recherche à travers le prisme des études de lisibilité. En particulier, nous utilisons un ensemble plus fourni de paramètres de simplification et avons recours à des statistiques multivariées. Nos analyses portent sur des données en français tirées du corpus *Europarl* (français traduit de l'anglais et français non traduit) et montrent que le français traduit est plus simple, d'un point de vue lexical et syntaxique, que le français non traduit. Une tendance à la convergence lexicale est également mise au jour en français traduit, ce qui semble indiquer que les traducteurs lissent les différences lexicales entre les orateurs de la langue source.

Revisiting simplification in corpus-based translation studies: insights from readability research

THOMAS FRANÇOIS
Université catholique de Louvain, Louvain-la-Neuve, Belgium
thomas.françois@uclouvain.be

MARIE-AUDE LEFER
Université catholique de Louvain, Louvain-la-Neuve, Belgium
marie-aude.lefer@uclouvain.be

RÉSUMÉ

Depuis les travaux de Laviosa (1998a ; 1998b), les études de corpus se sont souvent penchées sur le phénomène de simplification lexico-syntaxique, afin de déterminer si les textes traduits sont plus simples que les textes non traduits. Laviosa (1998a ; 1998b) traite également de l'hypothèse de convergence, selon laquelle les textes traduits sont plus homogènes que les textes non traduits. Jusqu'à présent, cette question a cependant suscité moins d'intérêt que la simplification. En traductologie de corpus, la simplification a été opérationnalisée à l'aide de cinq paramètres principaux. Sur la base de ceux-ci, les études ont montré que la simplification varie en fonction des modalités de traduction, des paires de langues et des registres analysés. Notre article a pour objectif de revisiter ce type de recherche à travers le prisme des études de lisibilité. En particulier, nous utilisons un ensemble plus fourni de paramètres de simplification et avons recours à des statistiques multivariées. Nos analyses portent sur des données en français tirées du corpus *Europarl* (français traduit de l'anglais et français non traduit) et montrent que le français traduit est plus simple, d'un point de vue lexical et syntaxique, que le français non traduit. Une tendance à la convergence lexicale est également mise au jour en français traduit, ce qui semble indiquer que les traducteurs lissent les différences lexicales entre les orateurs de la langue source.

ABSTRACT

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 texts. The simplification hypothesis states that translated texts are simpler than non-translated texts. The convergence hypothesis, also discussed by Laviosa (1998a; 1998b), but less so in follow-up studies, is that translated texts are more homogeneous than original texts, that is they display less variance. To date, simplification has mostly been operationalised in CBTS as type-token ratio, lexical density, core vocabulary coverage, list head coverage and average sentence length. Relying on these parameters, previous research has produced mixed results, with simplification varying across translation modalities, language pairs and registers. The present article sets out to revisit the simplification and convergence hypotheses through the lens of NLP-informed readability research. In particular, we rely on a larger set of simplification indicators and make use of multivariate statistical techniques. We present a simplification study of *Europarl* corpus data in French translated from English and in non-translated French. The results show that translated French is simpler than original French, lexically and syntactically. We also find evidence of convergence that shows that translators smooth out cross-speaker lexical heterogeneity in translated parliamentary proceedings.

RESUMEN

Desde que Laviosa publicara su trabajo pionero sobre la simplificación léxico-sintáctica en traducción (1998a; 1998b), esta ha ocupado un lugar destacado en los Estudios de Traducción. De acuerdo con esta hipótesis, los textos traducidos son más simples que los no traducidos. La hipótesis de la convergencia, también elaborada por Laviosa, pero con menor seguimiento en el campo en investigaciones posteriores, postula que los textos traducidos son más homogéneos que los textos originales. Hasta la fecha, la simplificación se ha abordado en los estudios de traducción basados en corpus como la relación tipo-caso, la densidad léxica, la cobertura del vocabulario básico, la cobertura del *list head*, y la longitud media de la oración. Teniendo en cuenta estos parámetros, investigaciones previas han ofrecido resultados diversos, en los que se observa que la simplificación varía en función de las modalidades de traducción, pares de lenguas o registros. El objetivo de este artículo es revisitar las hipótesis de simplificación y convergencia a la luz de la investigación sobre legibilidad informada por el procesamiento de lenguaje natural. Para ello, nos basamos en un conjunto de indicadores de simplificación más amplio e hicimos uso de técnicas de estadística multivariante. Analizamos la simplificación en datos de francés traducido del inglés y en francés original, extraídos de *Europarl*. Los resultados muestran que el francés traducido es más simple que el no traducido, tanto a nivel léxico como sintáctico. También observamos casos de convergencia, de acuerdo con la cual los traductores minimizarían la heterogeneidad léxica entre interlocutores en la traducción de las actas parlamentarias.

MOTS-CLÉS/KEYWORDS/PALABRAS CLAVE

traductologie de corpus, simplification, simplicité, lisibilité, Europarl
corpus-based translation studies, simplification, simplicity, readability, Europarl
estudios de traducción basados en corpus, simplificación, simplicidad, legibilidad, Europarl

1. Introduction

Corpus-based translation studies (CBTS) has developed and branched out considerably since its emergence in the early 1990s. The scope of CBTS has grown in multiple ways and there have been key methodological advances. The monolingual comparable approach, first advocated by Baker (1993; 1995; 1996), where translated text is compared with non-translated original text in the same language, is still widespread, but is increasingly used in combination with the parallel approach, where translations are studied together with their sources. This reflects the renewed interest for source-language influence in CBTS, alongside a range of other factors that shape translational products, such as register, translation expertise (novice vs. expert translation) and translation method (use of computer-aided translation tools or machine translation). Recognition of the multifactorial nature of translation has gradually led to the increased use of advanced multivariate statistics in CBTS. Also noteworthy is the fact that corpus translation scholars are now examining forms of interlinguistic mediation other than written translation, such as consecutive and simultaneous interpreting, audiovisual translation and sign language interpreting, which all pose specific challenges related to corpus collection, as well as data extraction and analysis. Sophisticated annotation methods and more recent corpus techniques, such as parsing and n-gram extraction, have also entered the field.

In line with Baker's (1993; 1995; 1996) programmatic research agenda for CBTS, the study of features of translation still holds centre stage in the field today. Features

of translation are defined by Baker (1993: 243) as "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems." Typical examples of such phenomena include explicitation, simplification and normalisation. While empirical evidence of simplification has been found in different translated languages, including non-European ones, and translation modalities (for example 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 (1998a; 1998b) linguistic operationalisations 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 article, we propose an innovative approach to the study of simplification in translation that aims to move beyond the linguistic operationalisations of simplification used in Laviosa (1998a; 1998b). 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) (for surveys of the field, see Benjamin 2012; Collins-Thompson 2014; François 2015; Vajjala 2021). This paradigm offers robust and 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 can be overlooked by shallow parameters, and make use of advanced statistical methods to aggregate these parameters. Here, we use the translated and original proceedings of the European Parliament in French as a test case for the application of such an NLP-informed readability approach to simplification in CBTS.

The article is structured as follows. Sections 2 and 3 are devoted to simplification and its linguistic operationalisations in CBTS and readability research, respectively. Section 4 presents the *Europarl* corpus data analysed, the simplification parameters investigated and the statistical tests and methods used in the comparison of translated and original parliamentary proceedings. Section 5 provides an overview of the corpus results and discusses them. Concluding remarks are offered in Section 6.

2. Simplification research in corpus-based translation studies: operationalising a complex construct

Ever since the publication of Laviosa's studies in the late 1990s (Laviosa 1998a; 1998b), the phenomenon of lexico-syntactic simplification has been widely investigated in CBTS. 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, that is in situations of monolingual text production (see Bernardini, Ferraresi, *et al.* 2016: 64-65). The simplification hypothesis, which can be traced back to Baker (1993) and the pre-corpus-linguistic research she mentions, states that translated language, being a form of interlinguistic mediation and constrained communication (see Kotze 2019), is informationally, lexically, syntactically and discursively simpler than original language production. As acknowledged by Ferraresi, Bernardini, *et al.* (2018: 734), the

term *simplicity* would be more appropriate to refer to this phenomenon. The term *simplification* would be better suited to parallel approaches, where target texts are compared with their sources to determine whether given source items have been simplified in translation. However, in this article, we use the term *simplification*, given its wide currency in monolingual comparable studies.

In Laviosa (1998a; 1998b), the construct of lexico-syntactic simplification is operationalised as lexical density (the proportion of lexical words out of the total number of running words; see also Baker 1995: 237-238), core vocabulary coverage (the proportion of high-frequency words, where high-frequency words correspond to the top 100 or 200 most frequent tokens in a reference corpus), list head coverage (a corpus-internal measure similar to core vocabulary coverage, except that the calculation is based on the top 100 most frequent words in the corpus being examined) and mean sentence length (the average number of tokens per sentence). Regarding lexical simplification, Baker (1995: 236) also mentions another operationalisation, namely lexical variety (type-token ratio). Laviosa (1998a) finds that English translated narrative prose is lexically simpler than original narrative prose: it displays a lower lexical density as well as higher core vocabulary and list head coverages. The study of English news corpora reveals similar trends, with the additional finding that sentences are shorter in translated news than in original news (Laviosa 1998b).

Laviosa (1998a; 1998b) also addresses the phenomenon of *convergence*, that is the "clustering of a corpus of translations around the average value of a linguistic feature" (Laviosa 1998a: 1). Likewise, Baker (1996) mentions the tendency of translated texts to be more similar to each other, more homogeneous, than comparable original texts, a phenomenon that she calls *levelling-out*. In Laviosa (1998b), for example, it is found that lexical density scores display less variance in translated texts than in original texts. In other words, the lexical density profile of translated news is more homogeneous than non-translated news. No such convergence is observed for translated fiction in Laviosa (1998a), which the author attributes to the limited number of texts included in the corpus.

Laviosa's simplification parameters have been used in several corpus-based translation and interpreting studies, such as Grabowski (2013), Kajzer-Wietrzny (2015), Bernardini, Ferraresi, *et al.* (2016) and Ferraresi, Bernardini, *et al.* (2018). The reference corpora used to extract lists of high-frequency words for the computation of core vocabulary coverage scores vary across studies. Kajzer-Wietrzny (2015) relies on Laviosa's word list, based on the *Collins Cobuild Bank of English*[1]. Bernardini, Ferraresi, *et al.* (2016), on the other hand, rely on the web-derived *WaCky* corpus family (Baroni, Bernardini, *et al.* 2009). However, in these studies, core vocabulary coverage is typically computed for the top 100 or 200 most frequent words. Additional simplification parameters have also been included in CBTS simplification research alongside Laviosa's indicators, such as mean word length, in characters (for example Kruger and Van Rooy 2012), readability indices based on word and sentence length, as in Williams (2005) and Redelinghuys and Kruger (2015) (for example the Flesch Reading Ease score) and hapax legomena (for example Lv and Liang 2019). NLP-oriented CBTS, such as Corpas Pastor, Mitkov, *et al.* (2008), Ilisei, Inkpen, *et al.* (2010) and Volansky, Ordan, *et al.* (2015), also examine additional simplification parameters, such as syllable ratio (average number of syllables per word), mean word rank (established on the basis of the 6 000 most frequent words of the language under scrutiny),

ambiguity (average of senses per word, based on *Wordnet* synsets; Miller 1995), simple vs. complex sentences (operationalised as the number of finite verbs per sentence) and sentence depth (parse tree depth). These other linguistic indicators are not standard practice in mainstream CBTS simplification research.

The simplification studies mentioned above have produced mixed results: while some suggest that translated texts are simpler than original texts in the same language, others point to the opposite trend, complexification, in translated language, depending on the mediation modalities (written translation vs. interpreting, consecutive vs. simultaneous interpreting), language pairs, translation directions, registers and degrees of translation expertise investigated (see for example Ferraresi, Bernardini, *et al.* 2018: 718-719 for a detailed overview). Little evidence has been found for the phenomenon of convergence at the levels of lexis or syntax, but it has admittedly received less attention in simplification studies based on Laviosa (1998a; 1998b) (notable exceptions include Williams 2005; Corpas Pastor, Mitkov, *et al.* 2008 and Grabowski 2013).

All in all, the picture that emerges from previous research is understandably quite complex, but it is still very much based on the core parameters taken from Laviosa (1998a; 1998b), considered separately. We agree with Bernardini, Ferraresi, *et al.* (2016: 65) that "while these parameters are clearly an approximation that cannot hope to do justice to the complexity of the notion of simplicity […], they do provide a methodological point of reference." However, we wish to argue that the time is ripe in CBTS to assess the usefulness of more robust models to better capture simplification patterns in translation. To do so, we draw on insights from readability research, where the construct of simplification has been operationalised and analysed in sophisticated ways.

3. The view from readability research

Readability research emerged in the United States in the 1920s as a way of supporting the reading practice of a large part of the population (Zakaluk and Samuels 1996). The field focuses on the development of statistical models, called readability formulas, which aim to predict the reading difficulty of a text for a specific population, relying solely on the linguistic characteristics of the said text. Reader variables (age, education level, ethnicity, etc.) and the context of reading (type of reading, goal of reading, time limit, etc.) are generally considered homogeneous within the population of interest.

The issue of text readability was first investigated in the field of education, giving rise to several readability formulas (Flesch 1948; Dale and Chall 1948; Gunning 1952), which have been widely used in a range of contexts since then. These formulas are based on two or three shallow text characteristics, such as the number of syllables per 100 words or average sentence length. Later, the exploration of readability shifted to psycholinguistics, as various scholars stressed the importance of higher-level text characteristics, such as conceptual density (Kintsch and Vipond 1979) or macrostructural aspects of the text (Meyer 1982). However, these studies did not make previous approaches obsolete, not only because they were much harder to automatise, but also because they did not achieve higher performance.

Since the early 2000s, readability has been increasingly investigated within the framework that François (2015) refers to as "AI readability." It combines NLP tech-

niques (to design and automatise more linguistically motivated features) with machine learning, which allows the use of more sophisticated models able to include far more variables than the classic formulas and to cope better with the linguistic information encoded in those variables. Representative examples of this paradigm are Collins-Thompson and Callan (2005), who showed that the grade levels of texts could be automatically predicted from word distributions, or Schwarm and Ostendorf (2005), who used a syntactic parser to extract several parser-based features. In the same vein, Pitler and Nenkova (2008) investigated various semantic and discursive features, such as lexical chains and discourse relations, while Vajjala and Meurers (2012) obtained excellent performance for English by combining NLP-enabled features with features coming from the field of second language acquisition, such as lexical variation and lexical density.

Deep learning, which has revolutionised the field of NLP, is now also used in readability research. For instance, Cha, Gwon, *et al.* (2017) rely on word embeddings, which are a dense representation of the semantic space of a language in which words and texts can be projected and compared. Additional information has quickly been added alongside embedding, such as age of acquisition, word frequencies and word length (Jiang, Gu, *et al.* 2018; Le, Nguyen, *et al.* 2018). Deep neural networks have also been applied to readability, for instance by Le, Nguyen, *et al.* (2018) and Azpiazu and Pera (2019).

Rather surprisingly, texts translated by humans (as opposed to machine-translated texts) have rarely been examined through the readability lens. A notable exception is Ciobanu, Dinu, *et al.* (2015), which sets out to assess whether source languages influence the readability of translated texts. The authors do so on the basis of English *Europarl* data (parliamentary proceedings in original English and English translated from 20 official languages of the European Union; see Section 4.1). Relying on the Flesch-Kincaid formula (based on the average number of syllables per word and words per sentence), as well as shallow, lexical and morpho-lexical features (such as type-token ratio and lexical density), they find that "readability features do not have enough discriminative power to obtain high performance on distinguishing original texts from translations" (Ciobanu, Dinu, *et al.* 2015: 102). However, the Flesch-Kincaid measure reveals interesting language-family clusters for Germanic, Slavic and Romance languages, which suggests a potential impact of the source language on the readability of translations. In the present article, we wish to cross-fertilise methodological insights from corpus-based translation studies and NLP-informed readability research to investigate simplification and convergence in translated language.

4. Data and methodology

This section presents the corpus data used (Section 4.1), the simplification parameters investigated (Section 4.2) and the statistical tests and methods we adopted to analyse simplification and convergence trends in our data (Section 4.3).

4.1. The Europarl data used: French translated from English and original French

As outlined above, the simplification and convergence hypotheses we want to test in this study are that proceedings in French, translated from English, are simpler and display less variance than comparable proceedings in original French. To do so, we rely on two French subcorpora of *Europarl-direct* (Cartoni and Meyer 2012). *Europarl-direct* is a directional version of *Europarl* (Koehn 2005), a multilingual parallel corpus made up of the verbatim reports (proceedings) of the plenary sessions of the European Parliament (EP). The speeches delivered at the EP, whether impromptu or read out, go through an editing process before their inclusion in the proceedings that are published on the EP website (for example deletion of disfluencies; see Ferraresi, Bernardini, *et al.* 2018: 723). Up to the first half of 2011, the proceedings were translated into the official languages of the European Union. This practice has since been discontinued. In *Europarl-direct*, the source languages of the speeches are clearly identified, which makes it possible to determine whether a given text is an original or a translation. Here we make use of a subcorpus of speeches originally delivered in French (Original French; OF) and a comparable subcorpus of speeches in French translated from English (Translated French; TF) (see Table 1). Both subcorpora have been POS-tagged with the *TreeTagger* (Schmid 1995). The speeches delivered in French and in English at the EP are mostly given by native speakers of the language (80% of native speakers of French, representing 84.6% of the speeches included in the OF subcorpus used; 65% of native speakers of English in the TF subcorpus used, representing 91% of the speeches). A wide range of topics are covered in the EP plenary sessions: agriculture, economics and finance, environment, health, justice, politics, procedure and formalities, science and technology, society and culture, and transport (see Kajzer-Wietrzny and Ferraresi 2019). The proportion of read-out versus impromptu speeches and the number of professional translators involved in the translation of the proceedings are unknown.

Table 1
***Europarl-direct* subcorpora used in the study**

	French translated from English (TF)	Original French (OF)
Total number of running words	1 552 093	634 138
Total number of speeches	5 257	1 880
Total number of speakers	237	192

We decided to rely on the directional *Europarl* corpus for several reasons: its availability to the research community, its truly multilingual make-up, its metadata and its homogeneity in terms of register (proceedings of parliamentary debates). In particular, its wide availability and multilingualism will make it easy to enlarge the empirical foundation of the approach presented here by replicating it on other *Europarl* datasets (other languages and language pairs). In addition, the corpus metadata[2] include speakers' names (mostly MEPs and commissioners), which makes it possible to test Laviosa's (1998a; 1998b) simplification and convergence hypotheses at the level of individual speakers. Specifically, the convergence hypothesis we can test on the basis of by-speaker analyses is that translated proceedings display less

variance in simplification traits across speakers than original, non-translated proceedings. In other words, we aim to determine whether the translation process tends to flatten out cross-speaker heterogeneity. We also believe that the register of parliamentary debates is a good starting point for the proposed approach to simplification, as political discourse, as a whole, has been examined from numerous angles in corpus linguistics (see Ädel 2010) and in CBTS (Kajzer-Wietrzny, Ferraresi, *et al.* forthcoming), but less so in readability research (a notable exception is the study by Ciobanu, Dinu, *et al.* 2015 mentioned above). The reason for choosing French as a test case is twofold: (1) CBTS simplification research has mainly focused on English, to the detriment of other languages, and (2) a readability package including 400 NLP-informed features was available for French (François 2011). We decided to compare original French with French translated from English as a starting point, as this is the language pair with which the authors of the present contribution are most familiar. Restricting the translated corpus to a single source language allows us to control for interference. In fact, at present, very little is known about cross-linguistic contrasts related to lexico-syntactic or discursive simplicity (for example, for a given register, is Language A more syntactically elaborate, more lexically dense or varied than Language B?). The contrastive aspect of simplification is outside the scope of the present study, but it is clearly a facet of the CBTS simplification research agenda that will need to be addressed in future studies.

4.2. Overview of simplification parameters analysed

The simplification analyses presented in this article are based on François's (2011) readability work, which includes both classic and NLP-enabled simplification parameters (see also François and Fairon 2012; François and Miltsakaki 2012). For the present analyses, we have selected the 19 most relevant lexical, syntactic and discursive parameters from François's (2011) set. They are listed in Table 2.

The choice of the parameters listed in Table 2 has been guided by previous research in both CBTS and NLP-informed readability studies. We have been careful to select parameters that have proved useful in previous readability research and that can be meaningful in translation research, alongside more traditional simplification indicators *à la* Laviosa. The selected parameters are mainly lexical, with some syntactic and discursive parameters. For lexical variety, we have decided to rely on type-token ratios based on lemmas (normalised or not), rather than standard type-token ratios based on inflected forms, to cancel out the effect of inflectional richness and thereby ensure comparability with future studies (for example English has a poorer inflectional system than French). For lexical density, we have included two different ratios (total number of lexical words out of the total of grammatical words and out of the total of words). Only the second (LEX/ALL) is commonly used in CBTS. For the two density measures, the following word categories have been considered lexical: nouns, adjectives, adverbs, and verbs[3]. Conceptual density complements the two lexical density measures. It is the ratio of the total number of logical propositions to the total number of words per speech, as defined in Kintsch, Kozminsky, *et al.*'s (1975) specification model. To compute this parameter, we used Lee, Gambette, *et al.*'s (2010) software, which automatically estimates the number of propositions in French texts on the basis of 35 different rules. Word complexity is measured in three different

Table 2

Lexical, syntactic, and discursive simplification parameters examined in the study

Parameter Type	Simplification Parameter	Description
Lexical	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 normalised 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, Kozminsky, *et al.* (1975) and computed with *Densidées* (Lee, Gambette, *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 (see Jakubíček, Kilgarriff, *et al.* 2013). We calculated the variables on the basis of four different list sizes: top 200, 1 000, 2 000, and 5 000 most frequent lemmas in *frTenTen*.
	GMLF	Geometric mean of lemma frequencies
	75LF	75th percentile of the probability distribution of lemmas per speech
	90LF	90th percentile of the probability distribution of lemmas per speech
Syntactic	MSL	Mean sentence length: average number of words per sentence
	%LongSent	Percentage of sentences that are longer than 30 words (see Daoust, Laroche, *et al.* 1996)
Discursive	PRO/NAM	Ratio of pronouns to proper names
	PRO/NOM+NAM	Ratio of pronouns to nouns (common nouns and proper names)

ways: the mean length of words (in letters), the proportion of long words (here, words of 10 letters or more) and the average number of syllables per 100 words, a widely used feature which is part of the famous Flesch (1948) formula. The four measures of core vocabulary coverage correspond to the percentage of the words in speeches that are covered by the top 200, 1 000, 2 000, and 5 000 most frequent lemmas extracted from the web-crawled *frTenTen* reference corpus (Jakubíček, Kilgarriff, *et al.* 2013). In CBTS, core vocabulary coverage typically takes into consideration the top 100 or 200 most frequent words. Here, we have decided also to experiment with longer lists, as readability research has shown that the discriminative power of a list-based variable varies with the size of the list (Dale and Chall 1948; Harris and Jacobson 1974). Relying on short lists—as commonly done in CBTS—makes it difficult to discriminate between more complex texts. In view of the corpus used in this study, made up of speeches delivered by highly educated speakers, such a limitation seems particularly likely to generate inconclusive results. We have also decided to rely on lemma-based lists, rather than word-form-based lists, to ensure cross-linguistic comparability with

follow-up research. The proposed approach also relies on three additional lexical simplification parameters related to frequency: the geometric mean of lemma frequencies and the 75th and 90th percentiles of the probability distribution of lemmas per speech. These measures have been shown to capture variations in text lexical complexity better than variables based on means (for example in François 2011). We have not included list head coverage in our analyses as the repetitiveness of the (mostly) grammatical words it measures is roughly captured by the LEX/GRAM ratio. Syntactic complexity is here assessed with two measures: the average length of sentences, which is a very robust and efficient proxy for sentence complexity, and the proportion of long sentences (here, sentences of 30 words or more). Additionally, two discursive simplification parameters, which are both pronoun-noun ratios, are examined. Pronouns require the production of inferences by the reader in order to link each referring expression (pronoun) to its antecedent (Wilkens and Todirascu 2020). The reason for selecting these two parameters in CBTS research is that they operationalise a phenomenon observed by Vanderauwera (1985: 97-98, cited by Baker 1993: 244) in the English translations of Dutch novels, namely the fact that potentially ambiguous pronouns are often translated with forms that allow for precise identification. The use of pronouns versus more precise forms of identification is mentioned by Baker (1993) in her discussion of simplification, but it has received very little attention in CBTS. An exception is Volansky, Ordan, *et al.* (2015: 106), where the ratio of personal pronouns to proper names magnified by an order of 3, referred to as *explicit naming*, is used as an operationalisation of explicitation in translation. The 19 parameters listed in Table 2 have been computed for the 7 137 speeches and 404 speakers included in the two subcorpora analysed (25 speakers are included in both corpora, having given speeches in both French and English at the EP).

4.3. Statistical tests and methods used

For the statistical analysis of the simplification parameters, we applied a twofold approach. First, drawing on Laviosa's (1998a; 1998b) methodology, we analysed all parameters separately in order to detect simplification or convergence effects at the parameter level. To start with, we applied Jarque Bera normality tests to all parameters, as the large size of our dataset did not allow us to use the Shapiro-Wilk test (1965). The Jarque Bera normality test (Jarque and Bera 1987) works by comparing the skewness and kurtosis coefficients of the empirical distribution with those of the normal distribution. It is a good option for large datasets. As all Jarque Bera tests rejected the normality assumption for our parameters with a p-value < 0.001, we used 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 characterise the size of the effect of translation on the simplification of speeches, we also computed point-biserial correlation coefficients between the two conditions (OF and TF) and each simplification parameter. When assessing the presence of an effect in a large dataset such as ours, an effect-size metric should always be preferred over a more conventional t-test (or similar test). Finally, again following Laviosa (1998a; 1998b) in this matter, we performed the Levene test (1960) on each parameter—at speaker level—to examine the convergence hypothesis. The Levene test was selected as it is known to be more robust to deviations from normality in the data.

Second, we investigated our two hypotheses (simplification and convergence) using multivariate analyses, which, to the best of our knowledge, has not been done before in CBTS simplification research (see De Sutter and Lefer 2020 on the need to use multivariate statistics in CBTS). This approach allows us to consider the effect of translation on all simplification parameters at once, offering a more encompassing way of testing the two hypotheses.

For the simplification hypothesis, we applied a principal component analysis (PCA) transformation of the 19 parameters. In readability studies, some parameters tend to encode similar information, which makes data interpretation more complex. PCA can reveal hidden structure in the data and thereby help identify the most meaningful simplification trends. In our study, the PCA was carried out with the *psych* package in R (Revelle 2019), with a *promax* rotation. We decided to keep the three most explanatory components, as they explained 61% of the variance, which seems enough for the purposes of our analysis. As the components of the PCA are orthogonal (Schlens 2014), univariate tests can be applied to each component separately. In this study, Wilcoxon rank sum tests were run for each component, together with a point-biserial correlation to characterise the effect size.

To test the convergence hypothesis considering all parameters at once, we designed the following methodology. First, each speaker was represented by a vector of 19 dimensions. These dimensions correspond to the 19 parameters described in Section 4.2. They make it possible to locate every speaker in a vector space, according to the lexical, syntactic, and discursive characteristics of his/her speeches. In such a space, two speakers sharing very similar characteristics are neighbours, whereas speakers with speeches with very different characteristics are located very far apart. We hypothesise that if translated language is more homogeneous than original language, the language used by the various speakers represented in the translated French subcorpus should be more similar. From a mathematical point of view, it means that the average distance in our vector space between all speakers in the TF condition should be smaller than the average distance between all speakers in the OF condition. To compute the distance between a speaker *i* and all other speakers from the same condition, we first calculated, for each speaker *j* (where $1 < j < N$; $i \neq j$), the Euclidean distance between the vector of this speaker *i* and the vector of speaker *j*, thus obtaining *N-1* values. Then, we simply took the mean of these *N-1* values and assigned it to speaker *i*, as it represents the average distance from all other speakers in the same condition (TF or OF). As the result of this computation for all speakers, we obtained a new variable, *mean_dist*, to which we applied a standard Wilcoxon rank sum test to determine whether the mean distance between speakers in the OF condition was indeed higher than in the TF condition. In addition, effect size was estimated with a point-biserial correlation coefficient.

5. Results and discussion

This section, which presents the results of our corpus analyses and discusses them, is divided into two parts. Section 5.1 is devoted to simplification. Section 5.2 deals with convergence.

5.1. Simplification

In this section, we first present the results of the parameter-based simplification analyses, which we performed at the levels of both speeches and speakers, before moving on to the aggregate results obtained from the PCA.

The mean values for the 19 selected parameters, the corresponding results of the Wilcoxon rank sum test (W value), and the point-biserial correlation coefficient (cor) are given in Tables 3 and 4. As indicated above, two units of analysis have been used: speeches (Table 3) and individual speakers (Table 4).

TABLE 3
Simplification in TF and OF (by-speech analyses)

Parameters	TF mean	OF mean	W value	p-value	cor	interpretation
TTR-L	0.53	0.51	4426500	>0.001	0.07	complexification
NormTTR-L	0.69	0.69	5027500	0.26	N/A	no difference
LEX/GRAM	1.24	1.30	5888000	>0.001	-0.13	simplification
LEX/ALL	0.55	0.56	5888000	>0.001	-0.14	simplification
ConcDens	0.48	0.48	5070800	0.09	N/A	no difference
MWL	4.81	4.79	4879000	0.42	N/A	no difference
PW10	5.42	5.58	5191700	>0.01	-0.03	simplification
Syllper100	162.56	162.74	5085100	0.06	N/A	no difference
CVC200	0.64	0.63	4378500	>0.001	0.09	simplification
CVC1000	0.82	0.80	3882600	>0.001	0.16	simplification
CVC2000	0.89	0.88	3802300	>0.001	0.18	simplification
CVC5000	0.96	0.94	3514200	>0.001	0.26	simplification
GMLF	-748.7	-762.5	4006000	>0.001	0.15	simplification
75LF	61 pmw	56 pmw	4083600	>0.001	0.02	simplification
90LF	99 pmw	74 pmw	3459900	>0.001	0.02	simplification
MSL	24.87	27.04	5808000	>0.001	-0.12	simplification
%LongSent	29.5	35.3	5783700	>0.001	-0.12	simplification
PRO/NAM	33.40	27.56	4542900	>0.001	0.04	complexification
PRO/NOM+NAM	0.52	0.50	4524200	>0.001	0.01	complexification

The results of the Wilcoxon tests largely confirm the simplification hypothesis for French parliamentary proceedings translated from English. The by-speech and by-speaker analyses show that the proceedings in French translated from English are lexically and syntactically simpler than comparable French originals. Lexically, for instance, translations are less dense (LEX/GRAM, LEX/ALL), contain fewer words of 10+ letters (PW10) and rely more extensively on high-frequency words (all core vocabulary coverage variables). These trends are in sharp contrast with previous research by Ferraresi, Bernardini, *et al.* (2018), based on a similar but much smaller dataset derived from EPTIC (*European Parliament Translation and Interpreting Corpus*[4]), where it was found that there was no difference in lexical density and core vocabulary coverage between English translated from French and original English. As regards syntax, our results show that sentences are shorter in translations (MSL) and that there are fewer sentences of 30+ words in translations (%LongSent). This is in line with Bernardini, Ferraresi, *et al.* (2016), who found that in EP proceedings in Italian translated from English, sentence length is lower than in original Italian. In

Table 4
Simplification in TF and OF (by-speaker analyses)

Parameters	TF mean	OF mean	W value	p-value	cor	interpretation
TTR-L	0.53	0.51	18896	>0.01	0.09	complexification
NormTTR-L	0.69	0.69	22888	0.92	N/A	no difference
LEX/GRAM	1.26	1.33	27882	>0.001	-0.23	simplification
LEX/ALL	0.55	0.57	28208	>0.001	-0.23	simplification
ConcDens	0.48	0.48	24552	0.16	N/A	no difference
MWL	4.83	4.84	23998	0.33	N/A	no difference
PW10	5.43	5.82	25682	>0.05	-0.11	simplification
Syllper100	163.33	164.02	24655	0.14	N/A	no difference
CVC200	0.64	0.62	16454	>0.001	0.22	simplification
CVC1000	0.82	0.80	14637	>0.001	0.28	simplification
CVC2000	0.89	0.87	13818	>0.001	0.31	simplification
CVC5000	0.95	0.94	13714	>0.001	0.36	simplification
GMLF	-754.11	-773.30	15294	>0.001	0.29	simplification
75LF	58 pmw	48 pmw	16242	>0.001	0.10	simplification
90LF	87 pmw	66 pmw	11735	>0.001	0.08	simplification
MSL	25.08	27.11	29171	>0.001	-0.16	simplification
%LongSent	29.29	35.55	29474	>0.001	-0.21	simplification
PRO/NAM	30.05	21.54	17224	>0.001	0.14	complexification
PRO/NOM+NAM	0.483	0.476	18906	>0.01	0.01	complexification

their text classification experiment based on translated and original English *Europarl* data, Volansky, Ordan, *et al.* (2015), by contrast, report that sentences in English translated from French are longer than in original English. This tends to indicate that gains and losses in syntactic complexity are translation-direction dependent rather than translation-inherent. There are also a few parameters for which no differences between TF and OF were found. These are standardised lemma-token ratio, conceptual density, mean word length and the number of syllables per 100 words. A parameter that points in the direction of complexification is the non-standardised lemma-token ratio, which is higher in French translated from English than in original French. This effect is, however, very likely to be due to the fact that speeches are shorter in translated French than in original French (mean values: 295 words per speech in TF vs. 337 words per speech in OF), which biases the TTR-L. Confirmation of this interpretation can be seen in the non-significant difference found for the normalised TTR-L. Interestingly, the two pronoun-noun ratios we have used (PRO/NAM; PRO/NOM+NAM) point to discursive complexification, contrary to our initial hypothesis. With the benefit of hindsight, this trend may be linked to a well-known cross-linguistic contrast between French and English: French is said to be more nominal than English, which is more verbal. This difference is actually reflected in our dataset in the ratio of nouns to verbs, which is lower in TF than in OF (mean in TF: 1.46; mean in OF: 1.58; W= 27505; p-value > .001). Further research will be needed to fully account for this pattern.

To wrap up the bivariate analysis of simplification, two additional key findings are worth highlighting. First, the two sets of analyses show that the by-speaker approach generates more robust results, as indicated by the higher correlation coef-

ficients. This is an important insight to take into consideration in future research based on *Europarl* data. Indeed, to the best of our knowledge, most *Europarl*-based studies take speeches (texts), rather than speakers, as their typical unit of analysis. Second, the results related to core vocabulary coverage indicate that the most powerful parameter is CVC5000, that is the score based on the top 5 000 most frequent words extracted from a reference corpus (here *frTenTen*). Large frequency reference lists, such as the ones used here, should be experimented with in future studies to determine whether they should indeed become the norm in CBTS simplification research.

As regards the results of the multivariate approach, we applied a PCA to the 19 parameters in order to summarise them as three principal components. In view of the results of the bivariate analyses presented above, the PCA was applied at the level of the speakers only. The PCA technique has the advantage of creating axes combining the information contained in different parameters, thus allowing us to test the simplification hypothesis for several parameters at once. However, interpretation of the axes is not necessarily straightforward. The classic approach is to calculate the degree to which each parameter loads on the component. To achieve more meaningful results, it is also common to apply a rotation to the components, in order to force variables to load maximally on only one factor (Field 2014: 679). When variables are known to be correlated with each other, as is the case in readability studies (François 2011), it is recommended to use a *promax* rotation. The loadings of the 19 parameters on the three components are reported in Table 5. Following Field (2014: 682), we considered loadings higher than 0.40 as meaningful and "the higher in magnitude a loading is (in either the positive or negative direction), the more important the variable is for the component" (Dumont 2018: 285). The most important loadings make it possible to interpret the components of the PCA.

In our case, a rather clear picture emerges. The first component mostly encodes frequency information about the lexical dimension of the speeches. The most relevant parameters are those identifying the proportion of frequent words in the speeches, based on the *frTenTen* lists, as well as the geometric mean of the frequency of all words (GMLF). Interestingly, lexical density is also included in this first dimension. When the value of this component increases, it means that the text becomes simpler, as it corresponds to a speech having more words from the *TenTen* lists, fewer lexical words compared with grammatical words and higher geometric means of word frequencies. Unsurprisingly, the mean of the TF condition is higher on this axis (0.29) than the mean of the OF condition (-0,36) and this difference is significant according to the Wilcoxon sum rank test ($W = 14766$; p-value < .001). The size of the effect is moderate ($r = 0.32$) and corresponds roughly to the magnitude of the parameters based on the *TenTen* list ($0.22 \leq r \leq 0.36$ in Table 4), which are those that load the most on this first component of the PCA. The second dimension is a mixed one, including lexical (lexical diversity, lexical density, and frequency of the 75th- and 90th-percentile words) and syntactic information (sentence length). When the value on this dimension increases, it means that the text becomes simpler, as it corresponds to shorter sentences, lower conceptual density, more frequent 75th- and 90th-percentile words. However, this component also loads the two TTR-L variables and the lexical density variables with positive coefficients, which should normally correspond to more complex speeches. This can be partially explained as follows: first, as mentioned before, the TTR-L is biased in our data and goes in the complexification direction, whereas

the loadings for lexical density are smaller and could have a corrective effect on other variables. If we stick to the interpretation that an increase on the second component means getting simpler, the mean of OF (-0.12) is indeed significantly more complex than the mean of the TF condition (0.1) as shown by the Wilcoxon test (W = 16634; p-value < .001). However, the effect size is small (r = 0.11), which is in line with the effect size of the parameters that load on this component ($0.09 \leq r \leq 0.16$ in Table 4). Finally, the third component clearly corresponds to word length, whether measured in letters or syllables. Its interpretation is also straightforward, as increasing on this axis means becoming more complex, namely having longer words and fewer words within the 200 most frequent words in the *frTenTen* list. Once again, the average of TF (-0.08) reveals a significant simplification effect (W = 26078; p-value = 0.009) compared to OF (mean = 0.11), but this effect is small (r = -0.10), as the two parameters that load the most on this component, namely MWL and Syllper100, are not significant in Table 4.

Summing up, the bivariate and multivariate simplification analyses performed on the French *Europarl* data suggest that texts translated from English are lexically and syntactically simpler than non-translated texts. Nevertheless, the effects we observed are of limited magnitude, even when multiple parameters are considered together in the PCA. The most highly correlated dimension, the first, explains about 10% of the variance in the complexity that distinguishes originals from translated texts.

Importantly, the monolingual comparable corpus approach taken here does not make it possible to determine whether these differences can be attributed to source-language influence (here, English) or whether they are translation-inherent. The complexity profile of the English source texts will need to be examined to solve this question. It is also important to add that the two discursive parameters included in the study point in the opposite direction, namely complexification. Follow-up studies will need to take a closer look at these parameters, as they may also be related to source-language influence, French being more nominal than English.

Table 5

Loadings of the 19 parameters over the 3 components of the PCA

Parameters	1st Component	2nd Component	3rd Component
TTR-L	-0.08	**0.79**	0.09
NormTTR-L	-0.22	**0.76**	0.13
LEX/GRAM	**-0.8**	**0.5**	0.03
LEX/ALL	**-0.83**	**0.43**	0.02
ConcDens	-0.3	**-0.47**	-0.03
MWL	0.07	0.06	**0.98**
PW10	-0.04	-0.27	**0.7**
Syllper100	0.07	0.04	**0.98**
CVC200	**0.56**	-0.19	**-0.54**
CVC1000	**0.86**	0.14	0.03
CVC2000	**0.91**	0.2	0.22
CVC5000	**0.89**	0.16	0.2
GMLF	**0.77**	0.03	-0.27
75LF	0.31	**0.61**	-0.11
90LF	0.22	**0.61**	0.14
MSL	0.13	**-0.55**	0.24

%LongSent	0.13	**-0.51**	0.18
PRO/NAM	0.15	0.04	-0.05
PRO/NOM+NAM	0.17	**0.48**	-0.07

5.2. Convergence

This section is devoted to convergence and aims to determine whether translated texts tend to be more similar to each other than original texts, using the simplification parameters investigated above as the starting point of the convergence analyses. We first examine separately the results of the variance tests applied on each parameter before discussing aggregate results.

Table 6 gives the variance scores of the simplification parameters for each of the two conditions (TF and OF), together with the F value of the Levene's test for homogeneity of variance. In view of the fact that the by-speaker analyses are more robust than the by-speech analyses (see Section 5.1), we only report results for the former. As can be seen from the table, the convergence hypothesis is largely confirmed, especially in the case of lexical parameters such as lexical density and core vocabulary coverage, as the translated texts, grouped by speaker, are often found to be more homogeneous than the originals. This tends to indicate that cross-speaker (lexical) heterogeneity is smoothed out by translators.

Table 6

Convergence in TF and OF (by-speaker analyses)

Parameters	TF variance	OF variance	F value	p-value	interpretation
TTR-L	0.0099	0.0087	0.95	NS	no difference
NormTTR-L	0.0021	0.0023	0.23	NS	no difference
LEX/GRAM	0.0115	0.0337	23.47	>0.001	convergence
LEX/ALL	0.0004	0.0008	22.09	>0.001	convergence
ConcDens	0.0006	0.0009	1.74	NS	no difference
MWL	0.0367	0.0552	5.95	>0.01	convergence
PW10	2.50	3.31	4.22	>0.01	convergence
Syllper100	40.41	45.76	4.59	>0.01	convergence
CVC200	0.0008	0.0016	22.67	>0.001	convergence
CVC1000	0.0007	0.0014	20.68	>0.001	convergence
CVC2000	0.0004	0.0012	36.26	>0.001	convergence
CVC5000	0.0003	0.0009	63.08	>0.001	convergence
GMLF	624.79	1379.70	36.04	>0.001	convergence
75LF	2.714349e-09	1.677740e-09	4e-04	NS	no difference
90LF	1.449450e-10	2.450632e-10	0.01	NS	no difference
MSL	37.69	39.37	3.27	>0.05	convergence
%LongSent	188.72	231.60	1.90	NS	no difference
PRO/NAM	1049.13	697.83	3.21	>0.05	divergence
PRO/NOM+NAM	0.0241	0.1952	2.23	NS	no difference

The degrees of freedom of the F value are (1 427); NS = not significant

We also examined the convergence hypothesis using an innovative design that compares the two subcorpora in terms of the mean of the average Euclidean distances

between each speaker and all other speakers in the same condition. The mean distance of all speakers in translated French is 49.7 whereas it is 59.9 in original French. These results confirm that speakers originally speaking in French are more heterogeneous in terms of the 19 parameters investigated than the speakers represented in the translated French data, which is in line with the bivariate analyses presented above. The difference between the two conditions is significant according to the Wilcoxon rank sum test ($W = 34923$; p-value $< .001$) and the size of the effect reaches $r = -0.22$, which means that text status (translated vs. original) explains about 5% of the distance variance between speakers.

To sum up the results obtained for convergence, we see that the two statistical approaches we have adopted here lend support to the convergence hypothesis in translated French parliamentary proceedings, in that translated speeches display less variance than non-translated, original speeches. At this stage, however, source-language influence cannot be ruled out. Follow-up studies will need to check whether the convergence trends observed here in French translated from English are not due to the fact that speakers who deliver speeches in English at the EP tend to produce more homogeneous texts than speakers who give speeches in French.

6. Conclusion

In this article, we set out to draw on insights from readability studies to revisit Laviosa (1998a; 1998b)-inspired simplification and convergence research in CBTS. In particular, we wanted to explore additional simplification parameters (19 in total) and provide aggregate overviews of the simplification and convergence traits of translated texts. To test the proposed approach, we analysed *Europarl* corpus data in French translated from English and in original French. The simplification hypothesis, according to which translated texts are simpler than original texts, is largely confirmed for the data examined. Translated texts are found to be simpler, both lexically and syntactically, than original texts. We also found convincing evidence of convergence, especially at the level of lexis. This suggests that translators are apt to smooth out cross-speaker heterogeneity in translation. We hope to have shown that the use of a larger set of simplification parameters drawn from readability studies, combined with advanced multivariate statistics, can benefit CBTS research. Our corpus study has also brought to the fore the relevance of *Europarl* analyses at the speaker level (rather than the speech level) and the crucial importance of relying on large reference frequency lists in the study of core vocabulary coverage. It is our hope that the proposed approach will be used by corpus translation scholars working on other language pairs, registers and translation modalities, so as to enhance our understanding of the simplification- and convergence-related features that typify translation and other forms of interlinguistic mediation.

There are several ways in which the present study can be complemented in future research. It is undeniable that the use of the 19 parameters analysed in this study is a first step forward for simplification research in CBTS. However, we believe that additional simplification indicators will need to be explored in follow-up studies. Examples of such parameters include phraseological complexity measures, degrees of polysemy, and syntactic complexity measures based on parsing. Other discursive simplification measures will need to be included too. This is an area where more work

needs to be done, as discourse has clearly been under-researched compared to lexis and syntax in readability research to date, especially for languages other than English. In addition to extending the set of parameters, we would like to apply the same methodology to English *Europarl* data. Analysis of the simplification and convergence profiles of the English source texts would help to further understand and explain the patterns observed in the present study (for example pronoun-noun ratios, cross-speaker variance). We would also like to supplement the monolingual comparable analyses with bidirectional parallel analyses (English-to-French and French-to-English parallel data). This will make it possible to determine whether there is indeed an overall decrease in complexity from source texts to target texts. If simplification and convergence are observed in the two translation directions, this would constitute strong evidence in favour of the simplification and convergence hypotheses. If different trends are observed, this would point to the crucial role of source-language interference and cross-linguistic contrasts. A key challenge lying ahead, if multi-parameter models are to become standard practice in empirical translation studies, is to ensure the cross-linguistic comparability of the parameters used.

NOTES

1. https://collins.co.uk/pages/elt-cobuild-reference-the-collins-corpus
2. It should be noted that the *Europarl-direct* corpus made available by Cartoni and Meyer (2012) does not adopt a rigorous standard (such as XML). We therefore had to develop an in-house Python script in order to automatically identify each speech, alongside its metadata (that is speaker ID, language of the speech and name of the speaker). The result of this extraction process was manually checked and revealed very few errors. We found some empty speeches as well as <CHAPTER> speeches, without any exploitable content. In total, 630 speeches were dropped out of 16 751 (for the French and English subcorpora taken together).
3. The verb category includes auxiliaries, because the *TreeTagger* for French does not distinguish between lexical verbs and auxiliaries.
4. https://corpora.dipintra.it/eptic/

REFERENCES

Ädel, Annelie (2010): How to use corpus linguistics in the study of political discourse. *In*: Anne O'Keeffe and Michael McCarthy, eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 591-604.

Azpiazu, Ion Madrazo and Pera, Maria Soledad (2019): Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*. 7:421-436.

Baker, Mona (1993): Corpus linguistics and translation studies: Implications and applications. *In*: Mona Baker, Gill Francis and Elena Tognini-Bonelli, eds. *Text and technology: In honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins, 233-250.

Baker, Mona (1995): Corpora in translation studies: An overview and some suggestions for future research. *Target*. 7(2):223-243.

Baker, Mona (1996): Corpus-based translation studies: The challenges that lie ahead. *In*: Harold Somers, ed. *Terminology, LSP and Translation: Studies in Language Engineering. In Honour of Juan C. Sager*. Amsterdam: John Benjamins, 175-186.

Baroni, Marco, Bernardini, Silvia, Ferraresi, Adriano, *et al.* (2009): The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*. 43(3):209-226.

Benjamin, Rebekah George (2012): Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*. 24(1):63-88.

Bernardini, Silvia, Ferraresi, Adriano and Miličević, Maja (2016): From EPIC to EPTIC—Exploring simplification in interpreting and translation from an intermodal perspective. *Target*. 28:61-86.

Cartoni, Bruno and Meyer, Thomas (2012): Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. *In*: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, *et al*., eds. *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul: European Language Resources Association (ELRA), 2132-2137.

Cha, Miriam, Gwon, Youngjune and Kung, H. T. (2017): Language modeling by clustering with word embeddings for text readability assessment. *In*: Ee-Peng Lim and Marianne Winslett, eds. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. New York: Association for Computing Machinery, 2003-2006.

Ciobanu, Lina Maria, Dinu, Liviu P. and Pepelea, Flaviu Ioan (2015): Readability Assessment of Translated Texts. *In*: Ruslan Mitkov, Galia Angelova and Kalina Bontcheva, eds. *Proceedings of Recent Advances in Natural Language Processing*. Hissar: INCOMA, 97-103.

Collins-Thompson, Kevyn (2014): Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*. 165(2):97-135.

Collins-Thompson, Kevyn and Callan, Jamie (2005): Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*. 56(13):1448-1462.

Corpas Pastor, Gloria, Mitkov, Ruslan, Naveed, Afzal, *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*. Stroudsburg: Association for Machine Translation in the Americas, 75-81.

Dale, Edgar and Chall, Jane (1948): A formula for predicting readability: Instructions. *Educational Research Bulletin*. 27(2):37-54.

Daoust, François, Laroche, Léo and Ouellet, Lise (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.

De Sutter, Gert and Lefer, Marie-Aude (2020): On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives*. 28(1):1-23.

Dumont, Amandine (2018): *Fluency and disfluency: a corpus study of non-native and native speaker (dis)fluency profiles*. Doctoral dissertation, unpublished. Louvain-la-Neuve: Université catholique de Louvain.

Ferraresi, Adriano, Bernardini, Silvia, Miličević Petrović, Maja, *et al.* (2018): Simplified or not simplified? The different guises of mediated English at the European Parliament. *Meta*. 63(3):717-737.

Field, Andy (2014): *Discovering statistics using IBM SPSS statistics*. 4th ed. London: Sage.

Flesch, Rudolf (1948): A new readability yardstick. *Journal of Applied Psychology*. 32(3):221-233.

François, Thomas (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, Thomas (2015): When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*. 20(2):79-97.

François, Thomas and Fairon, Cédrick (2012): An "AI readability" formula for French as a foreign language. *In*: Jun'ichi Tsujii, James Henderson and Marius Paşca, eds. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics, 466-477.

François, Thomas and Miltsakaki, Eleni (2012): Do NLP and machine learning improve traditional readability formulas? *In*: Sandra Williams, Advaith Siddharthan and Ani Nenkova, eds. *NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability*

for target reader populations (PITR 2012). Stroudsburg: Association for Computational Linguistics, 49-57.

Grabowski, Łukasz (2013): Interfacing corpus linguistics and computational stylistics. Translation universals in translational literary Polish. *International Journal of Corpus Linguistics*. 18(2):254-280.

Gunning, Robert (1952): *The Technique of Clear Writing*. New York: McGraw-Hill.

Harris, Albert and Jacobson, Milton (1974): Revised Harris-Jacobson readability formulas. *In*: *Proceedings of the 18th annual meeting of the College Reading Association, Bethesda, Maryland*. Oct.31 - Nov. 2. Bethesda: College Reading Association.

Ilisei, Iustina, Inkpen, Diana, Corpas Pastor, Gloria, *et al.* (2010): Identification of translationese: A machine learning approach. *In*: Alexander Gelbukh, ed. *Computational Linguistics and Intelligent Text Processing*. Heidelberg: Springer, 503-511.

Jarque, Carlos and Bera, Anil (1987): A Test for Normality of Observations and Regression Residuals. *International Statistical Review*. 55(2):163-172.

Jiang, Zhiwei, Gu, Qing, Yin, Yafeng, *et al.* (2018): Enriching word embeddings with domain knowledge for readability assessment. *In*: Emily M. Bender, Leon Derczynski and Pierre Isabelle, eds. *Proceedings of the 27th International Conference on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 366-378.

Jakubíček, Miloš, Kilgarriff, Adam, Kovář, Vojtěch, *et al.* (2013): The TenTen corpus family. *In*: Andrew Hardie and Robbie Love, eds. *7th International Corpus Linguistics Conference*. Lancaster: University Centre for Computer Corpus Research on Language (UCREL) 125-127.

Kajzer-Wietrzny, Marta (2015): Simplification in interpreting and translation. *Across Languages and Cultures*. 16(2):233-255.

Kajzer-Wietrzny, Marta, Ferraresi, Adriano, Bernardini, Silvia, *et al.*, eds. (forthcoming): *Empirical investigations into the forms of mediated discourse at the European Parliament*. Berlin: Language Science Press.

Kajzer-Wietrzny, Marta and Ferraresi, Adriano (2019): *Guidelines for EPTIC collaborators*. Bologna: DIT Lab, University of Bologna.

Kintsch, Walter and Vipond, Douglas (1979): Reading comprehension and readability in educational practice and psychological theory. *In*: Lars-Göran Nilsson, ed. *Perspectives on Memory Research*. Hillsdale: Lawrence Erlbaum, 329-365.

Kintsch, Walter, Kozminsky, Ely, Streby, William J., *et al.* (1975): Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*. 14(2):196-214.

Kruger, Haidee and Van Rooy, Bertus (2012): Register and the features of translated language. *Across Languages and Cultures*. 13(1):33-65.

Koehn, Philipp (2005): Europarl: A parallel corpus for statistical machine translation. *In*: Makoto Nagao, ed. *MT Summit X*. Tokyo: Asia-Pacific Association for Machine Translation, 79-86.

Kotze, Haidee (2019): Converging what and how to find out why: An outlook on empirical translation studies. *In*: Lore Vandevoorde, Joke Daems and Bart Defrancq, eds. *New Empirical Perspectives on Translation and Interpreting*. London: Routledge, 333-371.

Laviosa, Sara (1998a): Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*. 43(4):557-570.

Laviosa, Sara (1998b): The English Comparable Corpus: a Resource and a Methodology. *In*: Lynne Bowker, Michael Cronin, Dorothy Kenny, *et al.*, eds. *Unity in Diversity? Current Trends in Translation Studies*. Manchester: St. Jerome.

Le, Dieu-Thu, Nguyen, Cam-Tu and Wang, Xiaoliang (2018): Joint learning of frequency and word embeddings for multilingual readability assessment. *In*: Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, *et al.*, eds. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*. Stroudsburg: Association for Computational Linguistics, 103-107.

Lee, Hyeran, Gambette, Philippe, Maillé, Elsa, *et al.* (2010): Densidées: calcul automatique de la densité des idées dans un corpus oral. *In*: Alexandre Patry, Philippe Langlais and Aurélien Max, eds. *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Rencontres jeunes Chercheurs en Informatique pour le Traitement Automatique des Langues*. Montréal: ATALA, 11-20.

Levene, Howard (1960): Robust Tests for Equality of Variances. *In*: Ingram Olkin, ed. Contributions to *Probability and Statistics*. Palo Alto: Stanford University Press, 364-367.

Lv, Qianxi and Liang, Junying (2019): Is consecutive interpreting easier than simultaneous interpreting? – a corpus-based study of lexical simplification in interpretation. *Perspectives*. 27(1):91-106.

Meyer, Bonnie (1982): Reading research and the composition teacher: The importance of plans. *College Composition and Communication*. 33(1):37-49.

Miller, George A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Pitler, Emily and Nenkova, Ani (2008): Revisiting readability: A unified framework for predicting text quality. *In*: Mirella Lapata and Hwee Tou Ng, eds. *Proceedings of the 2008 conference on empirical methods in natural language processing*. Stroudsburg: Association for Computational Linguistics, 186-195.

Redelinghuys, Karien and Kruger, Haidee (2015): Using the features of translated language to investigate translation expertise: A corpus-based study. *International Journal of Corpus Linguistics*. 20(3):293-325.

Revelle, William (2019): P*sych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston: Northwestern University. R package version 1.9.12. <https://CRAN.R-project.org/package=psych>.

Shapiro, Samuel Sanford and Wilk, Martin (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591-611.

Shlens, Jonathon (2014): A tutorial on principal component analysis. a*rXiv preprint arXiv:1404.1100*.

Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German. *In Proceedings of the EACL-95 SIGDAT-Workshop: From Text to Tags*, 47-50.

Schwarm, Sarah and Ostendorf, Mari (2005): Reading level assessment using support vector machines and statistical language models. *In*: Kevin Knight, Hwee Tou Ng and Kemal Oflazer, eds. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. New Brunswick: Association for Computational Linguistics, 523-530.

Vajjala, Sowmya (2021): Trends, limitations and open challenges in automatic readability assessment research. *arXiv preprint* arXiv:2105.00973.

Vajjala, Sowmya and Meurers, Detmar (2012): On improving the accuracy of readability classification using insights from second language acquisition. *In*: Joel Tetreault, Jill Burstein and Claudia Leacock, eds. *Proceedings of the seventh workshop on building educational applications using NLP*. Stroudsburg: Association for Computational Linguistics, 163-173.

Vanderauwera, Ria (1985): *Dutch Novels Translated into English: The Transformation of a "Minority" Literature*. Amsterdam: Rodopi.

Volansky, Vered, Ordan, Noam and Wintner, Shuly (2015): On the features of translationese. *Digital Scholarship in the Humanities*. 30(1):98-118.

Wilkens, Rodrigo and Todirascu, Amalia (2020): Simplifying Coreference Chains for Dyslexic Children. *In*: Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, *et al.*, eds. *Proceedings of the 12th Language Resources and Evaluation Conference*. Paris: European Language Resources Association (ELRA), 1142-1151.

Williams, Donna (2005): *Recurrent Features of Translation in Canada*. Doctoral dissertation, unpublished. Ottawa: University of Ottawa.

Zakaluk, Beverley and Samuels, Jay (1996): Issues related to text comprehensibility: The future of readability. *Revue québécoise de linguistique*. 25(1):41-59.