This is a PRE-FINAL version of the paper. For the PUBLISHED version, see:
Penha-Marion, L., Gilquin, G., & Lefer, M.-A. (2024). The effect of directionality on lexico-syntactic simplification in French><English student translation. In H. Kotze & B. Van Rooy (Eds.), *Constraints on language variation and change in complex multilingual contact settings* (pp. 153−190). John Benjamins Publishing Company. https://doi.org/10.1075/coll.60.06pen

CHAPTER 6

The effect of directionality on lexico-syntactic simplification in French><English student translation

Laura Penha-Marion, Gaëtanelle Gilquin & Marie-Aude Lefer
Université catholique de Louvain

This chapter reports on an exploratory case study designed to investigate lexico-syntactic simplification in French><English translations produced by students in two within-subjects language contact settings: translation from the foreign language (FL) into the first language (L1) (FL>L1 translation) and translation from the L1 into the FL (L1>FL translation). The aim of the study is to determine whether directionality affects student translation production and, if so, how. Lexico-syntactic simplification is operationalised as mean sentence length, root lemma-token ratio, lexical density, and core vocabulary coverage. The results indicate that translation directionality exerts an effect on the distribution of lexical items (lemmas, lexical words, and high-frequency words) in the translations (as compared to their corresponding source texts), with there being more lexical simplification in L1>FL translation than in FL>L1 translation. They reveal, in addition, that student translation production is also impacted by constraints both at the macro level (translation experience) and at the micro level (students' idiosyncrasies and individual source texts).

Keywords: lexico-syntactic simplification, simplicity, complexity, translation directionality, bilingual language production, constrained language, student translation

1. Introduction

The present study has been conducted within the constrained-language framework (Kotze, 2020, 2022; Kruger & Van Rooy, 2016; see also this volume, Chapter 1), which builds on Lanstyák and Heltai's (2012) argument that all language production is affected by constraints (e.g., physiological and psychological) but that, in certain communicative contexts (e.g., language mediation in contact situations), some constraints play a stronger than average role, giving rise to language varieties (e.g., translation, interpreting, code switching, second language [L2] writing) that exhibit similar linguistic features (e.g., hyperpurism, as manifested in the form of avoidance of foreign words). Kotze (2020, 2022) sets out five constraint dimensions along which said varieties can be compared: (1) Language activation (monolingual–bilingual), (2) modality and register (spoken–written–multimodal), (3) text production (independent/unmediated–dependent/mediated), (4) language proficiency (learner–proficient), and (5) task expertise (non-expert–expert).

Our study deals primarily with translation, most notably *student translation* (here, translation produced by novice translators, who are variably proficient in the source and target languages), and focuses on (language contact) directionality. In the constrained-language framework, this issue, which underlies bilingual communication[1] and closely relates to that of language proficiency, is embedded in the language activation dimension (Kotze, 2020, p. 345; 2022, p. 77). In our study, we explore a particular type of within-subjects language contact, namely the one happening in novice translators' minds and resulting in (written) translation production. Our main goal is to establish whether directionality affects student translation production and, if so, how. To put it somewhat differently, we seek to determine how the bilingual language activation constraint (specifically seen through the prism of translation directionality) manifests itself when the same two sequentially acquired languages, characterised by varying proficiency levels, come into contact (in novice translators' minds) during translation, particularly in the following settings: translation from the foreign language (FL) into the first language (L1) (FL>L1 translation, where FL is activated as the source language and L1 as the target language), and translation from L1 into FL (L1>FL translation, where L1 is activated as the source language and FL as the target language). Does bilingual language activation manifest itself equally in the two translation settings or is it stronger, and hence more linguistically visible, in L1>FL translation (where FL written production is involved)? With our study, we hope to contribute toward the understanding and identification of (within-subjects) language contact manifestations among novice translators.

To achieve our goal, we examine a student translation corpus composed of news articles in French (translated from English; FL>L1 translation) and English (translated from French; L1>FL translation). Our translation data have been collected following a repeated-measures design, which means that they have been produced by the same students in both translation settings. While this certainly offers advantages in terms of research design (e.g., lower levels of individual variability),

[1] By bilingual communication we understand any form of verbal communication established as the result of the contact between two languages (e.g., one L1 and one L2, two L1s, two L2s), regardless of the type of language contact (within-subjects or between-subjects) and the characterisation of the languages (e.g., in terms of modality, age of acquisition, and proficiency).

it also implies having textual data in two different languages. This, in turn, has one major methodological implication – the need to establish a basis of comparison for the translations, to ensure that any differences observed between them are the result of a directionality effect and not the reflection of cross-linguistic contrasts between original (i.e., non-translated) English and French newswriting.

In our study, L1><FL translation directionality is approached through the lens of one of the purported *varioversals* (i.e., linguistic features that typify similar-constraint-induced language varieties) (Kotze, 2022, p. 74), namely lexico-syntactic simplification. The construct of simplification and associated notions of complexity and simplicity have been investigated in a variety of linguistic fields, such as creole research (e.g., Aboh & Smith, 2009), plain language and readability (e.g., Crossley et al., 2011, 2019; Saggion et al., 2011), World Englishes (e.g., Szmrecsanyi & Kortmann, 2009), second language acquisition (SLA) (e.g., Leow, 1997; Takashima, 1992), learner corpus research (LCR) (e.g., Veiga, 2016), and corpus-based translation studies (CBTS) (e.g., Bernardini et al., 2016). The term *simplification* may take on different meanings depending on the field. In SLA, for example, simplification refers to a *strategy* typical of learners. In CBTS, it corresponds to a *feature* typical of translated texts.

In this chapter, the simplification construct is examined from two perspectives: the *bilingual comparable* perspective, based on the analysis of target texts in English and French, and the *bilingual bidirectional parallel* perspective, which relies on the analysis of source texts in English and French and corresponding target texts in French and English. As a result, we employ two terms to refer to the construct, so as to highlight this twofold methodological approach. We use *simplicity* in reference to the bilingual comparable perspective and *simplification* in reference to the bilingual bidirectional parallel perspective (see Ferraresi et al., 2018, p. 734). We define simplicity as the quality of a translated text whose form and/or meaning are simpler than that of other translated and/or original texts in the same or different language(s), and simplification as the result of a process whereby the form and/or meaning of source text linguistic structures are made simpler in the target text.

Drawing on Laviosa's (1998a, 1998b) simplification studies, we operationalise simplicity and simplification as mean sentence length, root lemma-token ratio, lexical density, and core vocabulary coverage. We check for two main types of pattern, taking into account cross-linguistic lexico-syntactic differences between original English and French newswriting:

1. simplicity patterns (emerging from bilingual comparable analyses), which reveal whether translations into FL English are lexico-syntactically simpler than translations into L1 French; and
2. simplification patterns (emerging from bilingual bidirectional parallel analyses), which indicate whether translation students simplify source texts when translating and, if so, whether there is more lexico-syntactic simplification when they translate into FL.

Our research questions are as follows: Does directionality influence simplicity and simplification in FL><L1 translation? If so, how? To address these questions, we take into account other constraints included in Kotze's (2020, 2022) model of constrained communication, given their relatedness and interdependence. In addition to bilingual language activation, we expect two other macro-level constraints (FL proficiency and translation experience) to impact simplicity and simplification to some extent. The students represented in our translation corpus are both *learners* of English (and thus arguably less proficient in FL English production than L1 French production) and *novices* at French><English translation, with less experience in French>English translation.[2] In particular, we expect lower levels of FL proficiency and translation experience in L1>FL translation to account for simpler and/or more simplified translations. In conducting our analyses, we also factor in two micro-level constraints (students' idiosyncrasies and individual source texts) as potential sources of variation.

The chapter is structured as follows. Section 2 gives the theoretical background to our study, addressing the issue of translation directionality in CBTS (Section 2.1) and outlining research on the simplification construct (Section 2.2), with regard to both translation in CBTS and L2/FL production in SLA/LCR. Section 3 describes the data and methodology used. Section 4 presents the main findings, and Section 5 concludes by discussing these and outlining avenues for future research.

2. Theoretical background

2.1. Translation directionality in CBTS

In CBTS, the issue of translation directionality, which concerns itself with product/linguistic and process/cognitive aspects associated with change in *translation direction* (i.e., source language > target language setting), has been addressed from two main standpoints: that of *language pair* directionality and that of *L1><L2/FL* directionality (Neumann, 2019). In the former, the focus is placed on language systems, and specifically on linguistic (a)symmetries between translations in opposite language pair combinations (e.g., English><German). In the latter, the focal point is the translator's L1 and L2/FL and linguistic (a)symmetries between translations into these languages.

While the issue of language pair directionality has been explored to a relatively large extent in CBTS (see, e.g., Evert & Neumann, 2017) and corpus-based contrastive linguistics (see, e.g., Dupont & Zufferey, 2017), L1><L2/FL directionality is still considerably under-researched. There are three main reasons for this. First, L1>L2/FL translation remains a taboo within large segments

[2] Albeit conceptually possible (and necessary within the constrained-language framework), this proficiency/ experience distinction is not so easily made in practice. We are, however, able to draw it in our study thanks to the configuration of our student translation corpus, which allows for the collection of rich, standardised metadata about the students' linguistic background, including their self-assessed proficiency in the source and target languages and FL><L1 translation experience (see Section 3.1 for more details).

of the translation community (Ferreira & Schwieter, 2017, p. 96).[3] Second, the corpora traditionally used in CBTS are mostly made up of translations into the L1 only (Lefer, 2020, p. 261). Third, descriptive studies tackling L1><L2/FL directionality tend to be undertaken from the perspective of cognitive translation studies (e.g., Buchweitz & Alves, 2006; Chang, 2011; Pavlović & Jensen, 2009). This means that they rely on research methods that are typically used in translation process research (e.g., keystroke logging, screen recording, and eye tracking) to examine L1><L2/FL translation processes. Studies investigating L1><L2/FL translation products – either on their own (e.g., Duběda, 2018; Pokorn et al., 2019) or together with their respective translation processes (e.g., Whyatt, 2018, 2019) – do exist, but few adopt corpus methods.

A notable exception is Rodríguez-Inés (2013). In her study, translations from English/French/German into Spanish and vice versa, produced by Spanish-speaking professional translators and FL teachers, are compared in terms of a series of shallow text features: use of parentheses, loanwords, calques, n-grams; lexical diversity (e.g., type-token ratio); sentence length. No major differences are found between the two groups as regards L1><L2/FL directionality, even though the results point to translation (in)variability patterns being affected by constraints in the following dimensions:

1. Language activation (cross-linguistic influence): Translations from and into French have the highest level of similarity, which is attributed to Spanish and French being typologically related.
2. Task expertise: Translations from French into Spanish produced by FL teachers are the most similar, which is thought to reflect their lack of translation experience – FL teachers tend to favour more literal translation solutions, a behaviour commonly observed in novice/student translators.
3. Language proficiency: Translations into FL show the highest degree of variability, regardless of the language pairs involved, which is considered to be a result of asymmetries in L2/FL proficiency.

Some of the measures adopted in Rodríguez-Inés (2013), particularly type-token ratio and mean sentence length, are traditionally used as proxies for simplification in CBTS (see, e.g., Laviosa, 1998a, 1998b) and linguistic complexity in SLA (see, e.g., Bulté & Housen, 2012).

[3] Many higher-education and professional institutions, for example in French-speaking Belgium, recommend that translators work from the L2/FL into the L1. This prescriptive position reflects a rather reductionist view of bilingualism, which equates bilingualism with *asymmetric sequential bilingualism* and thus only takes into consideration bilingual users whose: (1) L1 has been acquired in a monolingual setting; (2) L1 acquisition precedes and shapes L2/FL acquisition; (3) proficiency is higher in the L1 than in the L2/FL. The recommendation is based on the assumption that writing proficiency and translation quality are closely intertwined, such that the higher the proficiency, the higher the quality.

2.2. The construct of simplification in bilingual language production

The idea that language may be simplified for different purposes (e.g., in graded readers, to make texts more understandable for learners of a certain level) and reasons (e.g., as a result of the translating process or lower levels of L2/FL proficiency) is not new. As mentioned above, the construct of simplification has been investigated in a variety of linguistic fields. Since our study is situated at the intersection of CBTS and SLA/LCR, whose main varieties of interest (i.e., translation in the case of CBTS and L2/FL production in that of SLA/LCR) are influenced by the bilingual language activation constraint, in what follows we briefly outline research dealing with simplification in bilingual language production.

2.2.1. Simplification/simplicity and translation: The CBTS perspective

In CBTS, the term *simplification* (often preferred to *simplicity*) is used to describe a feature of translated texts characterised by low lexico-syntactic complexity, especially in comparison with non-translated texts in the same language. Simplification is one of Baker's (1993, 1996) proposed *translation 'universals'* (i.e., recurrent features of translated texts that are purportedly independent of particular source texts and languages). It has been examined mainly on the basis of monolingual comparable corpora, which are composed of original and translated texts in one language, with the aim of determining which texts are lexico-syntactically simpler. The underlying hypothesis is that translation, being both mediated and bilingualism-influenced language production, and thus more constraint-affected, displays higher levels of lexico-syntactic simplification. This hypothesis has been tested mostly as in Laviosa (1998a, 1998b), by means of the following measures:

- Mean sentence length: The average number of words per sentence.
- Type-token ratio: The ratio of *types* (i.e., different word forms) to *tokens* (i.e., instantiations of such forms).
- Lexical density: The proportion of lexical words to running words.
- Core vocabulary coverage: The proportion of high-frequency words to running words, where the frequency of a word is defined in relation to an external point of reference, a reference corpus.
- List head coverage: The proportion of high-frequency words to running words, where the frequency of a word is established internally, within a specific corpus.[4]

Lower mean sentence length, type-token ratio, and lexical density on the one hand, and higher core vocabulary coverage and list head coverage on the other, are indicative of lexico-syntactic simplification.

[4] Some of those measures have also been implemented in translation process research (through readability indices) to assess one of the key components of translation difficulty, namely source text difficulty (Hvelplund, 2011; Jensen, 2009). For more information on the construct of translation difficulty, see Sun (2015, 2019).

While evidence confirming the hypothesis has been produced, it does not point to simplification being a universal feature of translation (see, e.g., Xiao & Dai, 2014). Rather, it suggests that the lexico-syntactic profile of translated texts is contingent upon factors such as genre and language pair. For instance, mean sentence length has been found to be significantly lower in translated English newswriting, but significantly higher in translated English fiction (Laviosa 1998a, 1998b). It has also proved to be significantly lower in fiction translated from English into Polish (Grabowski, 2013) and Chinese (Wen, 2014), but significantly higher in non-fiction translated from English into French (Williams, 2005).

Recently, the simplification hypothesis has been expanded to include an additional constraint dimension – task expertise. Redelinghuys and Kruger (2015), for example, compare student and professional translations with original texts in an attempt to identify the simplest texts (lexico-syntactically speaking) among these. Their results indicate that, apart from the communicative setting and source texts, translation experience also plays an important role in shaping translated language: student translations are found to be the simplest, followed by professional translations and original texts. Their results are not, however, in line with those of Corpas Pastor et al. (2008), in whose study no simplification is observed in student translations.

2.2.2. Simplification/complexity and L2/FL production: The SLA/LCR perspective

In SLA, the term *simplification* (also preferred to *simplicity*) is used to describe a strategy typical of learners, who tend to simplify the target language, especially in the early stages of language acquisition (see, e.g., Ellis, 2008, pp. 80–82; Meisel, 1980). Simplification is regularly evoked in LCR to account for certain features observed in learner language, including non-standard features (e.g., lack of verb inflection, in Rosen, 2016, p. 310) and simple lexical or syntactic choices (e.g., predicative adjectives and vague nouns, in Hinkel, 2003).

More frequently, however, it is the concept of *complexity* that is discussed in the LCR literature. Complexity is one of the dimensions of Skehan's (1998, 2009) well-known framework of complexity, accuracy and fluency (CAF), which is used to describe language learners' performance and proficiency (see Housen et al., 2012). Marchand and Akutsu (2015), for instance, examine CAF in a learner corpus of computer-mediated communication in L2 English in an attempt to assign proficiency levels to the different texts included in the corpus. The features they look at include mean sentence length and use of sophisticated vocabulary. Measures of complexity, including some of those found in CBTS (see Section 2.2.1), have likewise been adopted with natural language processing applications in mind, for example as a way of trying to automatically identify learners' mother tongues (e.g., Crossley & McNamara, 2012).

Finally, complexity has also been investigated in a developmental perspective, considering whether learner language becomes more complex over time (i.e., complexity as a process rather than a quality). Thus, Vyatkina (2013) studies specific syntactic complexity features such as the use of coordinate, nominal, and non-finite verb structures in a longitudinal learner corpus of L2 German. She demonstrates that learners' developmental paths are characterised by an increase in the frequency and range of these syntactic complexity features.

3. Data and methodology

3.1. Data description

Our study relies on three sets of corpus data. The first set, extracted from the Multilingual Student Translation (MUST) corpus (Granger & Lefer, 2020), contains 42 translations from English into French (15 164 tokens) and 42 translations from French into English (11 838 tokens), produced by (the same) 41 first-year students enrolled in the MA degree program in translation at the Université catholique de Louvain.[5] The students are all native speakers of French and learners of English (average of 9.7 years of study), which has the status of a FL in French-speaking Belgium, their immediate social context. While the majority also hold a BA in translation and interpreting, six do not (of these, five hold a BA in modern languages and one in political sciences). This means that overall most students have had at least five semesters of English>French (FL>L1) translation and two semesters of French>English (L1>FL) translation. Six, however, have had only one semester of FL>L1 translation and no formal training in L1>FL translation (an elective L1>FL translation course is offered later in the programme). The constellation of constraints surrounding the FL>L1 and L1>FL translation settings therefore differs with regard to:

- Language activation: French and English are activated in both settings, but not in the same way. In FL>L1 translation, English is activated as the source language and French as the target language. In L1>FL translation, French is activated as the source language and English as the target language.
- Language proficiency: Proficiency is higher in L1 French – as a productive skill in FL>L1 translation and a receptive skill in L1>FL translation (asymmetric bilingualism).
- Task expertise: The students are budding translators and have had more training in English>French translation (FL>L1 translation is the main focus of their MA programme).

Constraints are similar in terms of modality and register (i.e., written translations of news articles) and text production (i.e., dependent on a source text). Our constraint matrix is set out in Table 6.1.

Table 6.1. Constraint matrix

Constraint dimension	**FL>L1 translation**	**L1>FL translation**
Language activation	Bilingual: ● FL English as source language ● L1 French as target language	Bilingual: ● L1 French as source language ● FL English as target language
Modality and register	Newswriting	Newswriting
Text production	Dependent/mediated	Dependent/mediated
Language proficiency	Asymmetric bilingualism:	Asymmetric bilingualism: ● Native in French (receptive)

[5] One student translated four texts because she had to retake the course in which we collected the translation data.

	● Upper-intermediate to advanced in FL English (receptive) ● Native in French (productive)	● Upper-intermediate to advanced in FL English (productive)
Task expertise	Novice, with two experience levels: ● 5 semesters of training (BA in translation and interpreting) ● 1 semester of training (no BA in translation and interpreting)	Novice, with two experience levels: ● 2 semesters of training (BA in translation and interpreting) ● No formal training (no BA in translation and interpreting)

The translations analysed in this study were produced on the basis of three source texts in English (A, B, C) and three source texts in French (D, E, F) (see Appendix 1 and 2). Table 6.2 shows the number of translations and respective tokens per source text. The source texts are informative news articles of around 300 words each, covering Belgium-related topics. They were obtained from three main sources: *The Guardian* (A, C), *NBC News* (B), and the Belgian newspaper *Le Soir* (D, E, F).

Table 6.2. Number of translations and tokens per English and French source text

	English source text			**French source text**		
	A	B	C	D	E	F
Number of translations	4	24	14	4	24	14
Number of tokens	1 172	8 444	5 548	1 026	6 800	4 012

We supplemented the MUST data with two purpose-built reference corpora, composed of original newswriting in English and French, so as to uncover potential cross-linguistic lexico-syntactic differences between the two languages. We did so with a view to disentangling L1/FL directionality effects from these. Both reference corpora are made up of full texts of approximately 220–420 words, written within the 2018–2020 timeframe. The English reference corpus consists of 322 articles (107 152 tokens), collected from the American press (*NBC News*, *The New York Times*, *USA Today*) and the British press (*BBC News*, *The Guardian*, *The Independent*, *The Telegraph*). The French reference corpus likewise contains 322 articles (104 941 tokens), gathered from the Belgian press (*RTBF*, *La Libre*, *Le Soir*) and the French press (*RTL*, *Le Figaro*, *Le Monde*, *Le Parisien*). We selected the aforementioned news sources in view of their content quality, political position (right wing or left wing), source type (broadcasters or newspapers), language variety (two varieties per language), and high circulation numbers (in the case of newspapers).

To check how representative the six source texts were of newswriting in English and French, we compared them with the corresponding reference corpora, relying on the four simplicity/simplification measures mentioned above. Figure 6.1 shows violin plots for each measure. These plots depict the position of the English and French source texts in relation to the distribution of the English and French reference data. As can be seen from the plots, all source text values fall between the 2.5 and 97.5 percentiles (bottom and top dotted lines), which is where most of the reference data points lie. This indicates that source texts A–C and D–F are generally speaking good representatives of original English and French newswriting.

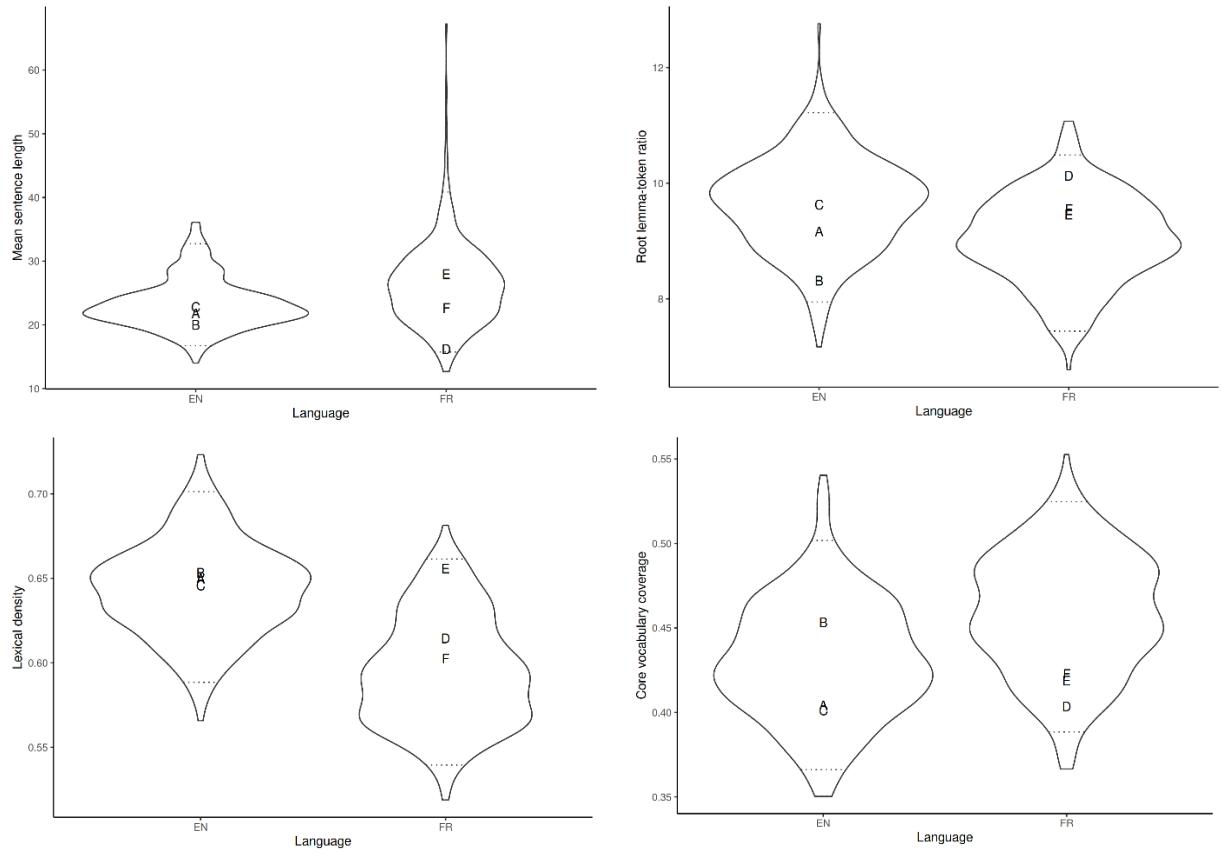

Figure 6.1. Violin plots for the four simplicity/simplification measures, depicting the position of source texts A–C and D–F in relation to original English and French newswriting

3.2. Linguistic analyses

Drawing on Laviosa (1998a, 1998b), we operationalised simplicity and simplification as mean sentence length, lexical density, core vocabulary coverage, and lemma-token ratio – a variation of type-token ratio. The reason we decided to adopt lemma-token instead of type-token ratio, which is what Laviosa (1998b) and other simplification studies in CBTS use, is that the notion of *lemma* (i.e., a non-inflected word form) is more comparable across languages, especially those differing in terms of inflectional richness, such as English and French.

We took lemma-token ratio and lexical density as measures of lexical richness, related to vocabulary range or breadth (i.e., how varied the words used in a given text are); core vocabulary coverage as a measure of lexical sophistication, associated with vocabulary depth; and mean sentence length as a syntactic measure, indicative of grammatical or structural range (see Bulté & Housen, 2012). We extracted the values needed to calculate each of these measures automatically (via Python scripts) from Sketch Engine (Kilgarriff et al., 2014) and captured them in a spreadsheet, where we performed all the appropriate mathematical operations. We calculated these four measures for the three corpora and source texts on the basis of text files. All the text files,

encoded separately in Sketch Engine, were automatically tokenised, lemmatised, and part-of-speech (POS) tagged with the English TreeTagger (see Schmid, 1994) and the French FreeLing (see Padró & Stanilovsky, 2012).

Mean sentence length is a length-based metric in which length is established in terms of word number. It is calculated by dividing the total number of tokens by the total number of sentences. To obtain this measure, we relied on sentence boundaries defined by the two aforementioned POS-taggers.

Seeing that type- and lemma-token ratios are highly dependent on text length (see, e.g., Jarvis, 2013), and that our corpus data differ in this respect, we made use of a square root transformation, known as the Guiraud Index or simply Guiraud – root lemma-token ratio – to compensate for text length variation. Guiraud has proved to be a happy medium between the classic type-token ratio and other stronger transformations (see Van Hout & Vermeer, 2007, p. 136). Root lemma-token ratio is calculated by dividing the number of lemmas by the square root of the number of tokens.

Lexical density can be computed in two different ways: by dividing the number of lexical words by the number of grammatical words or by dividing the number of lexical words by the overall number of tokens (see, e.g., Bulté & Housen, 2012). Following Laviosa (1998b), we used the latter computation, identifying lexical words through POS tags. We relied on Sketch Engine's automatic POS tagging to manually classify tags as either lexical (adjectives, adverbs, nouns, and verbs) or grammatical (articles, conjunctions, interjections, prepositions, and pronouns).

Core vocabulary coverage is a frequency-based measure that indicates the proportion "of text covered by the *n* most frequent words of [a] given language" (Bernardini et al., 2016, p. 65). Core vocabulary coverage is obtained by dividing the number of high-frequency words by the total number of tokens. In our study, we defined the frequency of English and French words in relation to the 100 most frequent words in the enTenTen15 (English reference corpus) and the frTenTen12 (French reference corpus), which are both available in Sketch Engine. These English and French TenTen corpora are in the order of ten billion words (see Jakubíček et al., 2013).

3.3. Statistical analyses

In light of our twofold bilingual methodological approach (see Section 1), we started by establishing how original English and French newswriting compare as regards the four measures. A Welch's *t*-test showed that all mean value differences are statistically significant and that the two display distinct complexity traits. While original French newswriting appears to be more complex than English at the syntactic level, with French sentences being significantly longer (t = -7.084, $p < 0.001$; $\bar{x}_{FR}$ = 26.31 vs $\bar{x}_{EN}$ = 23.3), original English newswriting seems more complex than French at the lexical level, in terms of both richness and sophistication. It has a significantly higher root lemma-token ratio (t = 9.658, $p < 0.001$; $\bar{x}_{EN}$ = 9.632 vs $\bar{x}_{FR}$ = 9.034), significantly higher lexical density (t = 20.539, $p < 0.001$; $\bar{x}_{EN}$ = 0.644 vs $\bar{x}_{FR}$ = 0.595), and significantly lower core vocabulary coverage (t = -9.838, $p < 0.001$; $\bar{x}_{EN}$ = 0.431 vs $\bar{x}_{FR}$ = 0.458).

We then proceeded to compute ratios for each *observation* (i.e., value pertaining to a text file) in the MUST data set. We did so in an effort to neutralise the cross-linguistic simplicity differences found between original English and French, establishing a *tertium comparationis* or basis of comparison for the student translations into English and French. We computed three ratios.

The first, the *translation ratio* (TR), corresponds to the quotient of a measurement obtained for a translation in a given language and the average of all the 322 measurements obtained for the reference corpus in the same language. It indicates a numerical relationship between the two values that describes the magnitude of one in relation to the other (i.e., how many times one is bigger or smaller than the other). To put it differently, TR gives us an idea of how certain lexico-syntactic traits of a given translation (in, say, French) compare to those of original (French) newswriting. By way of illustration, let us consider the translation of Source Text A (*ST_A*) and that of Source Text D (*ST_D*) produced by Student 5 (*stu5*). In the former (*stu5_L1trans_A*), mean sentence length is 23.667; in the latter (*stu5_FLtrans_D*), it is 17.538. Thus, the mean sentence length TRs of *stu5_L1trans_A* and *stu5_FLtrans_D* equate respectively to:

$$TR_{stu05\_L1trans\_A} = \frac{\text{mean sentence length of } stu05\_L1trans\_A}{\text{average of mean sentence length values in French}} = \frac{23.667}{26.311} = 0.899$$

$$TR_{stu05\_FLtrans\_D} = \frac{\text{mean sentence length of } stu05\_FLtrans\_D}{\text{average of mean sentence length values in English}} = \frac{17.538}{23.304} = 0.753$$

What $TR_{stu5\_FLtrans\_D}$, for example, is telling us is that sentences in *stu5_FLtrans_D* are, on average, approximately 25% shorter than sentences in the English reference corpus (ca. 75% of the numerator is contained in the denominator). By calculating TRs on the basis of reference values, that is, drawing translation-to-reference corpus *comparisons* (and not simply contrasting raw measurements such as 23.667 and 17.538), we are determining a 'common denominator' for French and English student translations so as to explore these irrespective of the language they are in.

The second ratio, the *source ratio* (SR), corresponds to the quotient of a measurement obtained for a source text in a given language and the average of all the 322 measurements obtained for the reference corpus in the same language. It gives an indication of how lexico-syntactically complex a source text (e.g., in English) is in relation to a reference corpus (in English). For instance, the mean sentence length SR of *ST_D* equals:

$$SR_{ST\_D} = \frac{\text{mean sentence length of } ST\_D}{\text{average of mean sentence length values in French}} = \frac{16.176}{26.311} = 0.615$$

This means that, on average, sentences in *ST_D* are 38.5% shorter than sentences in original French newswriting (61.5% of 16.176 is contained in 26.311). In our statistical analyses, SR is not used as is, but rather as a term of the third ratio.

The third ratio, the *target–source ratio* (TSR), corresponds to the quotient of a TR in one language (e.g., French) and the corresponding SR in the other language (English). It is intended to serve as a neutralised measure of simplification describing the relationship between a target text and its corresponding source text (i.e., whether lexico-syntactic complexity increases or decreases from the source text to the target text). For example, using $TR_{stu5\_FLtrans\_D}$ and SR_{ST_D} as terms of a mean sentence length TSR for *stu5_FLtrans_D* yields the following:

$$TSR_{stu5\_FLtrans\_D} = \frac{TR_{stu5\_FLtrans\_D}}{SR_{ST\_D}} = \frac{0.753}{0.615} = 1.224$$

$TSR_{stu5\_FLtrans\_D}$ shows that, with regard to mean sentence length, 122.4% of *stu5_FLtrans_D* is included in *ST_D*, which means that there is a 22.4% increase in sentence-length complexity from the latter to the former. There is therefore complexification (not simplification).

Regarding statistical testing, we first carried out monofactorial analyses, examining the bilingual language activation constraint on its own, so as to get an initial overview of simplicity and simplification patterns emerging from the data. Then we performed multifactorial analyses in order to address our research questions. In view of the limited set of data points at hand, the analyses should be viewed as exploratory. The monofactorial analyses consisted in running nonparametric Wilcoxon signed-rank tests (the translation data did not turn out to be normally distributed) with translation direction (FL>L1 vs L1>FL) as the independent variable in both simplicity and simplification comparisons, but the TR as dependent variable in the former and TSR as dependent variable in the latter. For the multifactorial analyses, we relied on linear mixed-effects regression models (lmer). We used the same independent/explanatory variables as fixed effects in both simplicity and simplification models: translation direction, FL English proficiency, and translation experience. The same holds true for random effect terms – source texts (i.e., their alphabetical ID) and students (i.e., their numerical ID). As to dependent/response variables, we used the log-transformed TR in simplicity models and log-transformed TSR in simplification models. We ran all the statistical analyses separately for each measure in RStudio 4.0.3 (R Core Team, 2020).

As regards students' metadata variables, that is, translation experience and FL English proficiency, we treated the former as a categorical variable taking on the following two levels: 1 (limited experience) for students with one semester of training in FL>L1 translation but no formal training in L1>FL translation and 2 (extensive experience) for students having had five semesters of FL>L1 translation and two semesters of L1>FL translation. We treated the latter as two categorical variables, namely self-reported English proficiency (upper-intermediate vs advanced) and stays in English-speaking countries (yes vs no), and one numerical variable – years studying English.

4. Results

4.1. Lexico-syntactic simplicity in translations into L1 French and FL English

As described in Section 3.3, we first performed monofactorial analysis to check for lexico-syntactic simplicity patterns in the MUST data (see Figure 6.2 for a graphical summary). We ran one Wilcoxon signed-rank test per measure, calculated as a TR. The results reveal statistically significant simplicity (target text vs target text) differences with regard to mean sentence length ($V = 182$, $p < 0.001$), lexical density ($V = 611$, $p < 0.05$), and core vocabulary coverage ($V = 261$, $p < 0.05$), but not root lemma-token ratio ($V = 520$, $p = 0.399$).

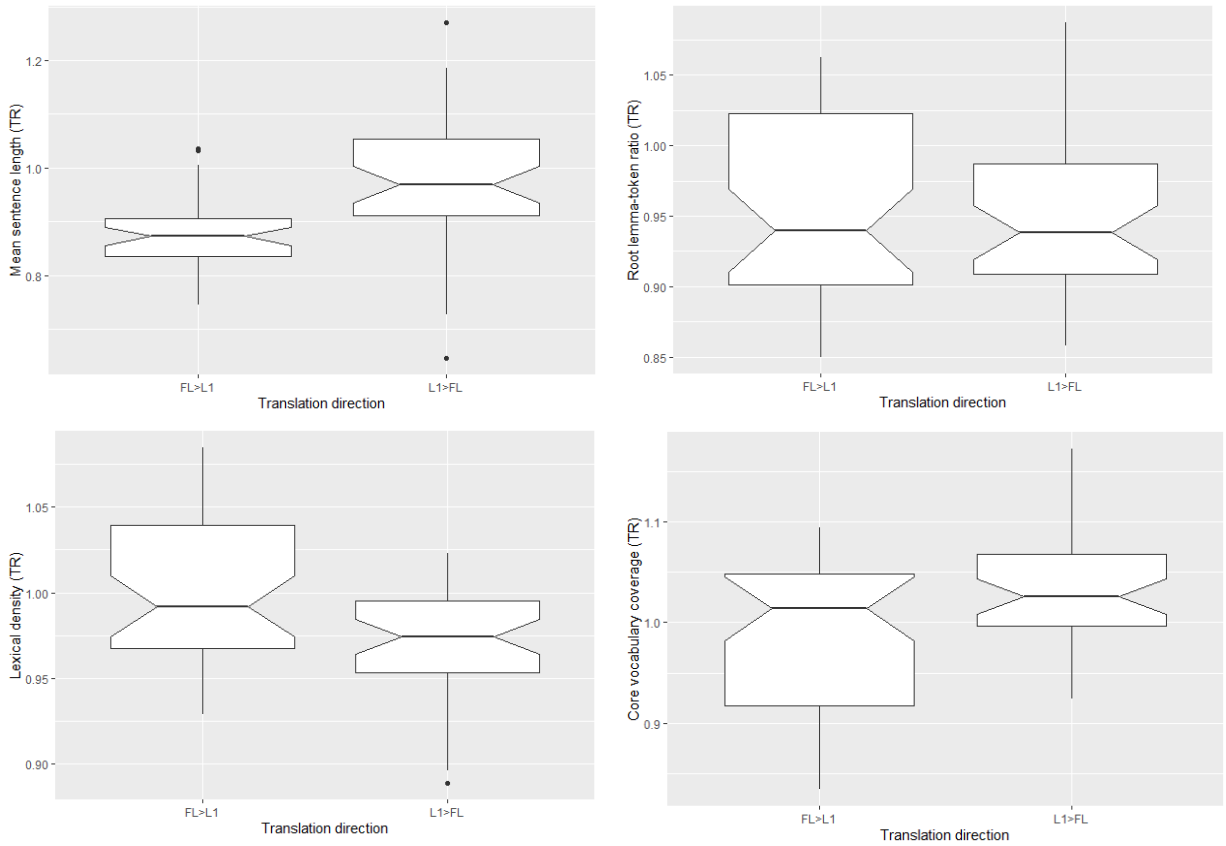

Figure 6.2. Boxplots summarising simplicity values (calculated as TRs) per translation direction

Table 6.3. Median values (based on TR quotients) of simplicity variables

Simplicity variable	$\tilde{x}_{TR}$	
	L1 French	**FL English**
Mean sentence length	0.873	0.969
Root lemma-token ratio	0.94	0.938
Lexical density	0.992	0.974
Core vocabulary coverage	1.014	1.025

Table 6.3 shows the median values (based on TR quotients) of the simplicity variables. Considered individually, the medians indicate how many times a translation measurement (e.g., in English) is typically bigger or smaller than the average reference corpus measurement (in English). If we compare magnitudes pairwise (i.e., bilingually), we are then able to determine the simpler or more complex translation set (magnitudes serve as indicators of *relative* lexico-syntactic complexity). Looking at Table 6.3, particularly at the median values associated with statistically significant differences, we see that:

- While sentences in translations into L1 French are typically 12.7% shorter than sentences in original French newswriting, sentences in translations into FL English are typically 3.1% shorter than sentences in original English newswriting.
- While translations into L1 French are typically 0.8% less lexically dense than news articles in original French, translations into FL English are typically 2.6% less lexically dense than news articles in original English.
- While translations into L1 French typically contain 1.4% more high-frequency words than news articles in original French, translations into FL English typically contain 2.5% more high-frequency words than news articles in original English.

We can conclude that the magnitude of translated English measurements (expressed as a percentage) is comparatively smaller in the case of mean sentence length and comparatively larger in that of lexical density and core vocabulary coverage. Compared to translations into L1 French, translations into FL English thus seem to be more complex at the syntactic level and simpler at the lexical level.

To assess the effect of bilingual language activation on each simplicity measure, factoring in other macro- and micro-level constraints, we ran one linear mixed-effect regression per measure, computed as a TR. Full results are provided in Appendix 3. The simplicity models for lexical density and core vocabulary coverage did not converge when both student and source text ID were taken as random effects as well as with only student ID as random effect. For this reason, we had to remove student ID from the models.

Of all the explanatory variables used as fixed effects, only translation experience turned out to be statistically significant in the lexical density model ($p < 0.001$, $R^2 = 0.73$), which suggests that translation experience is the sole macro-level constraint to explain variation in lexical density. The plot presented in Figure 6.3 shows how the two levels of translation experience, namely 1 (limited experience) and 2 (extensive experience), affect lexical density. It indicates that students with extensive translation experience produce translations with higher lexical density values. A micro-level constraint, individual source texts, also seems to account for some of the lexical variability. As can be seen in Appendix 3, their confidence intervals do not include zero.

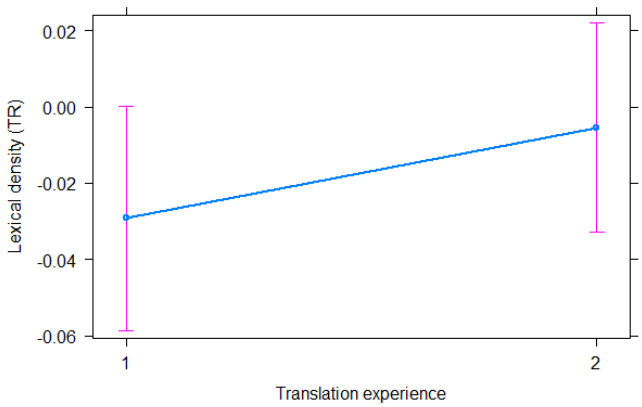

Figure 6.3. Plot showing the effect of translation experience on lexical density (measure computed as a TR; simplicity regression model)

In the core vocabulary coverage model, none of the explanatory variables reached statistical significance, although translation experience approaches significance (p = 0.0502, R^2 = 0.69). The effect plot in Figure 6.4 shows that students with limited translation experience produce higher core vocabulary coverage levels (i.e., lexically simpler texts). Part of the lexical variability also comes from individual source texts (see Appendix 3).

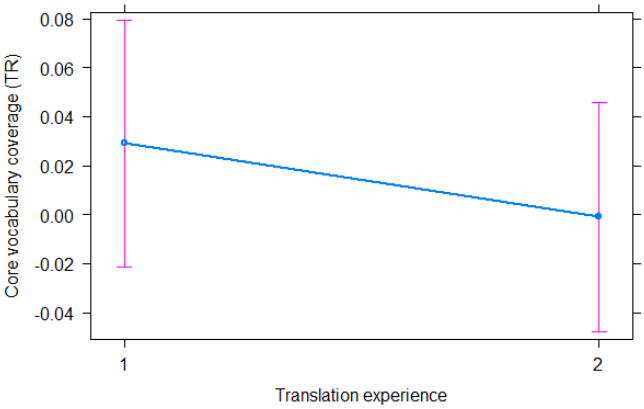

Figure 6.4. Plot showing the (non-significant) effect of translation experience on core vocabulary coverage (measure computed as a TR; simplicity regression model)

In the models for mean sentence length and root lemma-token ratio, none of the explanatory variables achieved statistical significance. However, random effect terms did (see Appendix 3). While both source text and student ID are statistically meaningful in the case of mean sentence length (R^2 = 0.78), only source text ID is in that of root lemma-token ratio (R^2 = 0.79). This suggests that sentence-length and lemma-related variation is explained by micro- rather than macro-level constraints. Variation in mean sentence length seems to be due partly to individual source text constraints and partly (though to a lesser degree) to students' idiosyncrasies. Lexical diversity variation (in terms of lemma range) appears to be due to individual source text constraints.

To sum up, the results of the regression analyses point to the simplicity profile of student translations being influenced by a few (rather than all) and by micro-level (rather than macro-level) constraints. While source text ID is associated with statistically meaningful trends in all regression models, translation experience – though not translation direction – is (or practically is) in two: those pertaining to lexical density and to core vocabulary coverage, with extensive translation experience corresponding to higher lexical complexity. The results reveal the influence of translation experience on the (lexical) simplicity traits of student translations.

4.2. Lexico-syntactic simplification in target texts

To check for lexico-syntactic simplification patterns in the MUST data (see Figure 6.5 for a graphical summary), we ran Wilcoxon signed-rank tests. We ran one test per measure, calculated as a TSR. The results reveal statistically significant simplification (target text vs source text) differences with regard to mean sentence length ($V = 236$, $p < 0.05$), root lemma-token ratio ($V = 9.3$, $p < 0.001$), lexical density ($V = 903$, $p < 0.001$), and core vocabulary coverage ($V = 520$, $p < 0.001$).

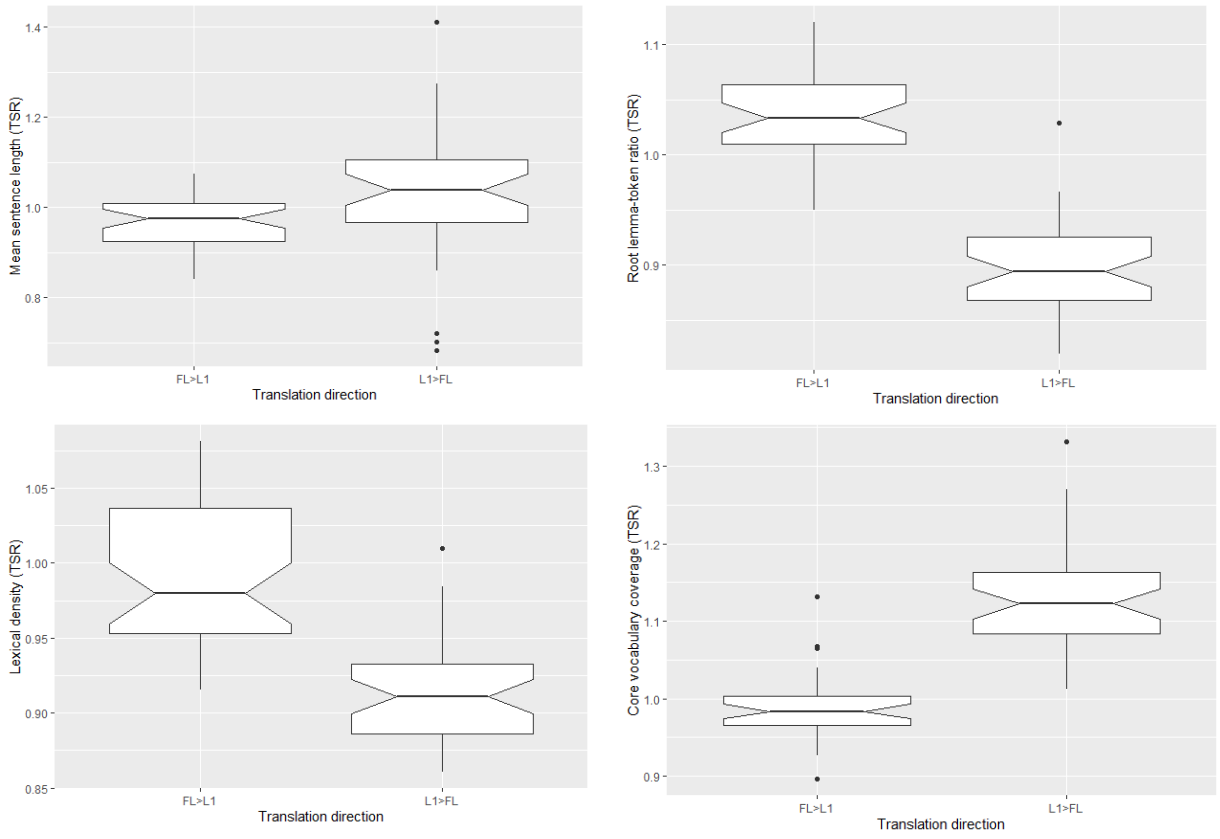

Figure 6.5. Boxplots summarising simplification values (calculated as TSRs) per translation direction

Table 6.4. Median values (based on TSR quotients) of simplification variables

Simplification variable	$\tilde{x}_{TR}$	
	FL>L1 translation	**L1>FL translation**
Mean sentence length	0.974	1.039
Root lemma-token ratio	1.034	0.894
Lexical density	0.98	0.911
Core vocabulary coverage	0.983	1.122

Table 6.4 shows the median values (based on TSR quotients) of the simplification variables. Considered individually, the medians indicate how many times a (neutralised) translation measurement (e.g., in English) is typically bigger or smaller than its corresponding (neutralised) source text measurement (in French). If we compare magnitudes pairwise (i.e., per translation direction), we are then able to determine whether there is simplification or complexification (from the source text to the target text) in one translation direction in relation to the other (magnitudes serve as indicators of decrease or increase in *relative* lexico-syntactic complexity). Looking at Table 6.4, we see that, typically, in FL>L1 translation, 97.4% of the mean length of sentences, 103.4% of lemmas, 98% of lexical words, and 98.3% of high-frequency words used in target texts are contained in their corresponding source texts. We also see that, typically, in L1>FL translation, those magnitudes correspond respectively to 103.9%, 89.4%, 91.1%, and 112.2%. We can conclude that, in L1>FL translation, sentences are lengthened, the ranges of lemmas and lexical words are narrowed, and the reliance on high-frequency words increases. As such, the results suggest that students both simplify and complexify source texts when translating. In L1>FL translation, students seem to simplify source texts at the lexical level, in terms of both richness and sophistication, but complexify them at the syntactic level.

To assess the effect of bilingual language activation on each simplification measure, factoring in other macro- and micro-level constraints, we ran one linear mixed-effect regression per measure, calculated as a TSR. Full results are given in Appendix 4. The simplification models for lexical density and core vocabulary coverage did not converge with both student and source text ID as random effects. We therefore opted to remove source text ID from the models.

In the lexical density model, the sole explanatory variable to reach statistical significance is translation direction ($p < 0.001$, $R^2 = 0.77$), which indicates that change in the direction of bilingual language activation explains lexical simplification (in terms of lexical density reduction in L1>FL translation). The plot depicted in Figure 6.6 shows the effect of translation directionality on lexical density: students simplify source texts to a greater degree in L1>FL translation. Part of this simplification trend can likewise be attributed to the micro-level constraint of students' idiosyncrasies (see Appendix 4).

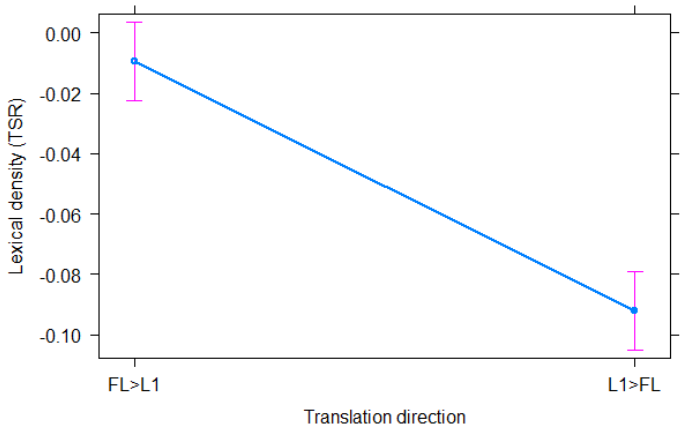

Figure 6.6. Plot showing the effect of translation directionality on lexical density (measure computed as a TSR; simplification regression model)

In the core vocabulary model, translation direction and translation experience were both statistically significant ($p < 0.001$ in both cases, $R^2 = 0.7$). The plots in Figure 6.7 show the effect of each explanatory variable on core vocabulary coverage. As we can see from the plots, there is more lexical simplification (as measured by core vocabulary coverage) in translations into FL and translations produced by less experienced students. Student ID (as a random effect term), on the other hand, was not statistically significant (see Appendix 4).

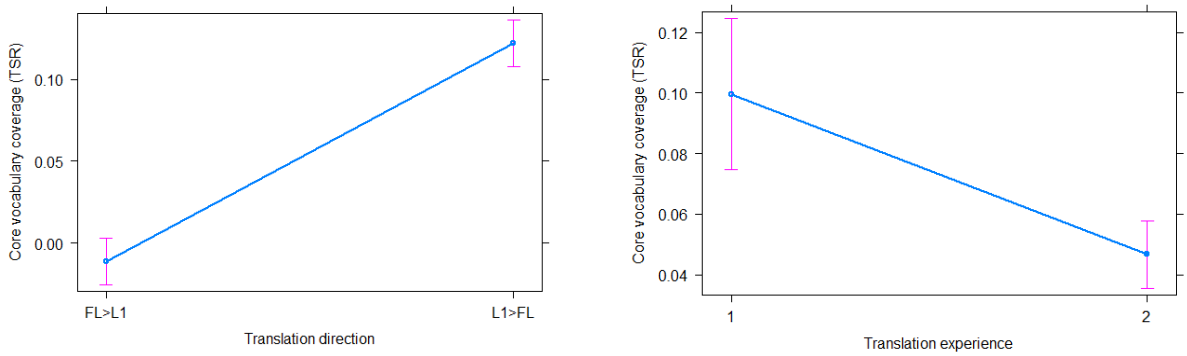

Figure 6.7. Plots showing the effect of translation directionality and translation experience on core vocabulary coverage (measure computed as a TSR; simplification regression model)

Somewhat similar trends emerge from the root lemma-token ratio regression analysis: There is a statistically significant effect for translation direction ($p < 0.05$, $R^2 = 0.81$) and source text ID, but none for student ID (see Appendix 4). The effect of translation directionality on root lemma-token ratio is shown in Figure 6.8. As we can observe, students simplify source texts lexically (in terms of lemma-token ratio reduction) when they translate into FL.

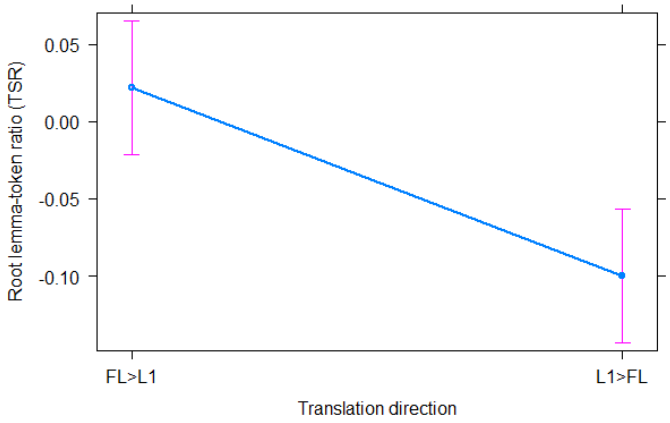

Figure 6.8. Plot showing the effect of translation directionality on root lemma-token ratio (measure computed as a TSR; simplification regression model)

Finally, in the model for mean sentence length, no explanatory variable attained statistical significance (translation experience approaches significance: $p = 0.065$, $R^2 = 0.74$), but student and source text ID did (see Appendix 4). Syntactic simplification variation, measured here with sentence length, thus seems to be explained mainly by micro-level constraints.

To sum up, the regression results indicate that translation direction and translation experience account for much of the simplification variation in the translation data, with there being more lexical simplification in translations produced in L1>FL translation and by inexperienced students. Translation direction did not reach statistical significance in the model for mean sentence length. However, it was statistically significant in three other models – on its own in two (lexical density and root lemma-token ratio) and together with translation experience in one (core vocabulary coverage). Student ID was significant in two models (lexical density and mean sentence length) and source text ID in one (root lemma-token ratio), which suggests that micro-level constraints are also an important factor affecting (student) translation production.

5. Discussion and conclusion

In this chapter, we reported on a case study designed to explore translations produced by French-speaking students in FL>L1 (English>French) and L1>FL (French>English) translation (or within-subjects language contact) settings. Our goal was to establish whether directionality affects student translation production and, if so, how. In our study, situated at the crossroads of CBTS and SLA/LCR, we approached the issue of translation directionality through the lens of the widely investigated simplification construct, operationalising it as mean sentence length, root lemma-token ratio, lexical density, and core vocabulary coverage.

To collect our translation data, we chose a repeated-measures design. We had the same students translate news articles in both translation settings and produce textual data in two languages. In so doing, we were faced with a key methodological hurdle (how to account for obvious cross-linguistic differences between non-translated English and French newswriting),

which we attempted to overcome through the statistical use of ratios, calculated on the basis of original texts collected in two (respectively English and French) purpose-built reference corpora serving as our control groups. Adopting this research design had another methodological implication for our study, conducted as it was within the constrained-language framework: of the constellation of potential constraints affecting our translation data, only those pertaining to two dimensions (text production, and modality and register) could be held constant. Those related to the language activation, language proficiency and task expertise dimensions, which varied according to the translation setting, had to be included in the statistical analyses. Text production deserves a special mention. Although our translation data share the dependent/mediated text production constraint, we decided to factor in the various source texts used in their production, suspecting these of influencing translated language at a lower level of granularity.

To get an overview of our translation data, we ran Wilcoxon signed-rank tests with translation direction as the independent variable. The results showed statistically significant differences with regard to three measures (all but root lemma-token ratio) in our simplicity analyses and all four measures in our simplification analyses. We identified two general patterns, regardless of how we looked at the data (i.e., comparing target texts among themselves or to source texts): lower lexical complexity and higher syntactic complexity in L1>FL translation. Interestingly, such patterns are opposite to those emerging from our reference data, with original English newswriting displaying higher lexical complexity and original French newswriting displaying higher syntactic complexity. What this finding seems to reveal are trends in the languages, possibly a functional trade-off of lexico-syntactic resources between original and translated newswriting, or a structural transfer from source texts and/or the source language. This merits further investigation. Replicating our analyses on data representing another register or language pair might prove interesting.

To tackle our research questions, we relied on mixed-effects modelling, examining bilingual language activation (our main interest) alongside other constraints. We treated the variables representing macro-level constraints (translation direction, FL proficiency, and translation experience) as fixed effects and those representing micro-level constraints (student and source text ID) as random effects. The results of our regression analyses suggest that the manifestations of (within-subjects) language contact in student translation are visible and measurable. The results also suggest that bilingual language activation may not be the sole or the major macro-level constraint to affect student translation production, although it seems to be the most prominent one as far as simplification is concerned. Both translation directionality and translation experience account for some of the lexical variation (with students in L1>FL translation and with limited experience producing simpler and/or more simplified target texts), but not to the same extent and in the same manner. While translation experience alone turned out to be a good predictor of lexical variation in simplicity models (it reached or almost reached statistical significance with lexical density and core vocabulary coverage as response variables), translation direction successfully predicted lexical variation in simplification models – both on its own (with lexical density and root lemma-token ratio as response variables) and together with translation experience (with core

vocabulary coverage as a response variable). However, the two variables did not explain sentence-length-related variation. They were not statistically significant when mean sentence length was taken as a response variable. As for the proficiency-related explanatory variables, none of these reached statistical significance either (in any of the regression models), which highlights the need for more refined measurements of FL proficiency.

Any explanation as to why bilingual language activation behaved the way it did in our study remains tentative at this stage. Even though it is possible that bilingual language activation is simply not strong enough to manifest itself clearly in the presence of other macro-level constraints, it might also be the case that the complexity of its behaviour was not fully captured by our regression analyses. Given that the constraint is very much entangled with language proficiency and that we performed those analyses in an exploratory fashion, considering each explanatory variable individually (i.e., not allowing for possible interactions between them in view of our small data set), that seems plausible. Future studies will need to address this caveat. Doing so would help us discern how bilingual language activation interacts with other constraints, how strong it really is, and whether or not there are interaction effects. It would also allow us to empirically assess the intertwining nature of constraints, further informing the constrained-language framework.

The regression results additionally suggest that micro-level constraints should not be overlooked. Random effects showed statistically meaningful trends in basically all regression models. Although this might be due to the specificities of our dataset, it might arguably also be because micro-level constraints are in fact core to (student) translation, perhaps even more so than macro-level constraints. This warrants further investigation. Treating student and source text ID as fixed effects in new regression analyses (while allowing for interactions between these and macro-level constraints) might help clarify the role the underlying constraints play in shaping student translated language. The fact that source text ID, in particular, was statistically significant also merits consideration, as it shows that constraints may play out even at the lowest levels of granularity. It likewise raises the question of whether any of the constraint dimensions can indeed be held completely constant.

To sum up, given the exploratory nature of our statistical analyses, we cannot at this point make broad generalisations about the absolute or relative effect of directionality on student translation products. More research involving diverse language pairs and more robust research designs are needed in order to better understand it. Ours could be improved in at least five ways. First, collecting more data overall (i.e., having more students translate more than two source texts) would increase the power of our statistical tests. Second, collecting data from English-speaking translation students would allow the adoption of a monolingual comparable approach. Third, asking students who contribute translations to the MUST corpus to take FL proficiency tests, such as the Oxford Quick Placement Test and LexTALE (Lemhöfer & Broersma, 2012), would help us to get a clearer grasp of their FL proficiency levels. Fourth, carrying out more fine-grained linguistic analyses could help provide a better insight into simplicity and simplification as typical features of constraint-induced language varieties. Examples include the examination of additional simplicity/simplification measures tapping into phraseological complexity (see, e.g., Paquot, 2019;

Rubin et al., 2021) and the investigation of simplification in tandem with other purported varioversals (e.g., explicitation, which has been said to be related to simplification; see Pym, 2008; Xiao & Dai, 2014). Fifth, refining our regression analyses, notably through the use of interaction terms, would allow us to speak more specifically to the intertwining nature and interaction of constraints.

Funding

This study was supported by the Fonds Spéciaux de Recherche (Université catholique de Louvain) and the Belgian National Fund for Scientific Research (F.R.S.-FNRS).

Acknowledgments

We would like to thank Denis Marion for his help with the automatisation process of the linguistic analyses and the editors and reviewers for their insightful feedback. We also gratefully acknowledge the statistical support provided by the Statistical Methodology and Computing Service (SMCS) at the Université catholique de Louvain. Finally, many thanks are due to the MA students at the Louvain School of Translation and Interpreting who contributed translations to the MUST project.

References

Aboh, E. O., & Smith, N. (Eds.). (2009). *Complex processes in new languages*. John Benjamins.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager* (pp. 175–186). John Benjamins.

Bernardini, S., Ferraresi, A., & Miličević, M. (2016). From EPIC to EPTIC: Exploring simplification in interpreting and translation from an intermodal perspective. *Target, 28*(1), 61–86. https://doi.org/10.1075/target.28.1.03ber

Buchweitz, A., & Alves, F. (2006). Cognitive adaptation in translation: An interface between language direction, time, and recursiveness in target text production. *Letras de hoje* [Letters today], *41*(2), 241–272. http://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/601

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins.

Chang, V. C.-Y. (2011). Translation directionality and the revised hierarchical model: An eye-tracking study. In S. O'Brien (Ed.), *Cognitive explorations of translation* (pp. 154–174). Continuum.

Corpas Pastor, G., Mitkov, R., Afzal, N., & Pekar, V. (2008). Translation universals: Do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the Eighth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)* (pp. 75–81). http://www.mt-archive.info/05/AMTA-2008-TOC.htm

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language, 23*(1), 84–101.

Crossley, S. A., & McNamara, D. S. (2012). Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. In S. Jarvis, & S. A. Crossley (Eds.), *Approaching language transfer through text classification: Explorations in the detection-based approach* (pp. 106–126). Multilingual Matters.

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading, 42*(3–4), 541–561. https://doi.org/10.1111/1467-9817.12283

Duběda, T. (2018). La traduction vers une langue étrangère : Compétences, attitudes, contexte social [Translating into a foreign language: Competences, attitudes, social context]. *Meta, 63*(2), 492–509. https://doi.org/10.7202/1055149ar

Dupont, M., & Zufferey, S. (2017). Methodological issues in the use of directional parallel corpora: A case study of English and French concessive connectives. *International Journal of Corpus Linguistics, 22*(2), 270–297. https://doi.org/10.1075/ijcl.22.2.05dup

Ellis, R. (2008). *The study of second language acquisition.* 2nd ed. Oxford University Press.

Evert, S., & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. In G. de Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New theoretical and methodological traditions* (pp. 47–80). De Gruyter Mouton.

Ferraresi, A., Bernardini, S., Petrović, M., & Lefer, M.-A. (2018). Simplified or not simplified? The different guises of mediated English at the European Parliament. *Meta, 63*(3), 716–737. https://doi.org/10.7202/1060170ar

Ferreira, A., & Schwieter, J. W. (2017). Directionality in translation. In J. W. Schwieter, & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 90–105). John Wiley.

Grabowski, Ł. (2013). Interfacing corpus linguistics and computational stylistics: Translation universals in translational literary Polish. *International Journal of Corpus Linguistics, 18*(2), 254–280. https://doi.org/10.1075/ijcl.18.2.04gra

Granger, S., & Lefer, M.-A. (2020). The Multilingual Student Translation Corpus: A resource for translation teaching and research. *Language Resources and Evaluation, 54,* 1183–1199. https://doi.org/10.1007/s10579-020-09485-6

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *Tesol Quarterly*, *37*(2), 275–301. https://doi.org/10.2307/3588505

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. John Benjamins.

Hvelplund, K. T. (2011). *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. Unpublished PhD thesis. Copenhagen Business School. http://hdl.handle.net/10419/208778

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen Corpus family. Paper presented at the Seventh International Corpus Linguistics Conference (CL2013), Lancaster, England, 23–26 July.

Jarvis, S. (2013). Defining and measuring lexical diversity. In S. Jarvis, & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13–44). John Benjamins.

Jensen, K. T. H. (2009). Indicators of text complexity. In S. Göpferich, A. L. Jakobsen, & I. M. Mees (Eds.), *Behind the mind: Methods, models and results in translation process research* (pp. 61–80). Samfundslitteratur.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kotze, H. (2020). Converging *what* and *how* to find out *why*: An outlook on empirical translation studies. In L. Vandevoorde, J. Daems, & B. Defrancq (Eds.), *New empirical perspectives on translation and interpreting* (pp. 333–370). Routledge.

Kotze, H. (2022). Translation as constrained communication: Principles, concepts and methods. In S. Granger, & M.-A. Lefer (Eds.), *Extending the scope of corpus-based translation studies* (pp. 67–97). Bloomsbury.

Kruger, H., & Van Rooy, B. (2016). Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide*, *37*(1), 26–57. https://doi.org/10.1075/eww.37.1.02kru

Lanstyák, I., & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures*, *13*(1), 99–121. https://doi.org/10.1556/Acr.13.2012.1.6

Laviosa, S. (1998a). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, *43*(4), 557–570. https://doi.org/10.7202/003425ar

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* (pp. 101–112). Routledge.

Lefer, M.-A. (2020). Parallel corpora. In M. Paquot, & S. Th. Gries (Eds.), *Practical handbook of corpus linguistics* (pp.) 257–282. Springer. https://doi.org/10.1007/978-3-030-46216-1

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*, 325–343. https://doi.org/10.3758/s13428-011-0146-0

Leow, R. P. (1997). Simplification and second language acquisition. *World Englishes, 16*(2), 291–296. https://doi.org/10.1111/1467-971X.00063

Marchand, T., & Akutsu, S. (2015). First steps in assigning proficiency to texts in a learner corpus of computer-mediated communication. In M. Callies, & S. Götz (Eds.), *Learner corpora in language testing and assessment* (pp. 85–112). John Benjamins.

Meisel, J. (1980). Linguistic simplification. In S. Felix (Ed.), *Second language development: Trends and issues* (pp. 13–40). Gunter Narr.

Neumann, S. (2019). Corpus insight on the complexity of directionality in translation. Paper presented at the Bilingualism and Directionality in Translation Conference, Brussels, Belgium, 12 December.

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards wider multilinguality. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)* (pp. 2473–2479). European Language Resources Association (ELRA).

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Pavlović, N., & Jensen, K. T. H. (2009). Eye tracking translation directionality. In A. Pym, & A. Perekrestenko (Eds.), *Translation Research Projects 2* (pp. 93–109). Intercultural Studies Group.

Pokorn, N. K., Blake, J., Reindl, D., & Peterlin, A. P. (2019). The influence of directionality on the quality of translation output in educational settings. *The Interpreter and Translator Trainer, 14*(1), 58–78. https://doi.org/10.1080/1750399X.2019.1594563

Pym, A. (2008). On Toury's laws of how translators translate. In A. Pym, M. Shlesinger, & D. Simeoni (Eds.), *Beyond descriptive translation studies: Investigations in homage to Gideon Toury* (pp. 311–328). John Benjamins.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.r-project.org/

Redelinghuys, K., & Kruger, H. (2015). Using the features of translated language to investigate translation expertise: A corpus-based study. *International Journal of Corpus Linguistics, 20*(3), 293–325. https://doi.org/10.1075/ijcl.20.3.02red

Rodríguez-Inés, P. (2013). ¿Cómo traducen traductores y profesores de idiomas? Estudio de corpus [How do translators and language teachers translate? A corpus study]. *Meta, 58*(1), 165–190. https://doi.org/10.7202/1023815ar

Rosen, A. (2016). The fate of linguistic innovations: Jersey English and French learner English compared. *International Journal of Learner Corpus Research, 2*(2), 302–322. https://doi.org/10.1075/bct.98.08ros

Rubin, R., Housen, A., & Paquot, M. (2021). Phraseological complexity as an index of L2 Dutch writing proficiency: A partial replication study. In S. Granger (Ed.), *Perspectives on the L2 phrasicon: The view from learner corpora* (pp. 101–125). Multilingual Matters.

Saggion, H., Gómez-Martínez, E., Etayo, E., Anula, A., & Bourg, L. (2011). Text simplification in simplext: Making texts more accessible. *Procesamiento del Lenguaje Natural* [Natural language processing], *47*, 341–342. http://www.redalyc.org/articulo.oa?id=515751747047

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. Paper presented at the Fifteenth International Conference on Computational Linguistics (COLING 1994), Kyoto, Japan, 5–9 August.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*, 510–532. https://doi.org/10.1093/applin/amp047

Sun, S. (2015). Measuring translation difficulty: Theoretical and methodological considerations. *Across Languages and Cultures, 16*(1), 29–54. https://doi.org/10.1556/084.2015.16.1.2

Sun, S. (2019). Measuring difficulty in translation and post-editing: A review. In D. Li, V. L. C. Lei, & Y. He (Eds.), *Researching cognitive processes of translation* (pp. 139–168). Springer.

Takashima, H. (1992). Transfer, overgeneralization and simplification in second language acquisition: A case study in Japan. *International Review of Applied Linguistics in Language Teaching (IRAL), 30*(2), 97–120.

Van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93–115). Cambridge University Press.

Veiga, N. V. (2016). Discourse markers in CEDEL2 and SPLLOC corpora of learner Spanish: Analysis of some lexical-pragmatic failures. In M. Alonso-Ramos (Ed.), *Spanish learner corpus research: Current trends and future perspectives* (pp. 267–297). John Benjamins.

Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *The Modern Language Journal, 97*(S1), 11–30. https://doi.org/10.1111/j.1540-4781.2012.01421.x

Wen, T.-H. (2014). Simplification in translated Chinese texts: A corpus-based study on mean sentence length. *NCUE Journal of Humanities, 9*, 253–271.

Whyatt, B. (2018). Old habits die hard: Towards understanding L2 translation. *Między oryginałem a przekładem* [Between originals and translations], *24*(41), 89–112. https://doi.org/10.12797/MOaP.24.2018.41.05

Whyatt, B. (2019). In search of directionality effects in the translation process and in the end product. *Translation, Cognition & Behavior, 2*(1), 79–100. https://doi.org/10.1075/tcb.00020.why

Williams, D. A. (2005). *Recurrent features of translation in Canada: A corpus-based study*. Unpublished PhD thesis. University of Ottawa.

Xiao, R., & Dai, G. (2014). Lexical and grammatical properties of Translational Chinese: Translation universal hypotheses reevaluated from the Chinese perspective. *Corpus Linguistics and Linguistic Theory, 10*(1), 11–55. https://doi.org/10.1515/cllt-2013-0016

Appendix 1: English source texts used in the study

A. Ken Loach responds angrily to Belgian PM in antisemitism row

British film director asks if Charles Michel had examined the evidence before accusing him of anti-Jewish comments

Ken Loach has responded with fury at a hastily convened press conference to a suggestion by the Belgian prime minister that a leading university was wrong to honour him following complaints that it had overlooked allegations of antisemitism.

An hour before the 81-year-old British director received his honorary degree from the free university of Brussels (ULB) on Thursday, he told reporters he could not allow Charles Michel's comments to go unanswered.

Loach said he was shocked to have to make a statement, adding that he understood that Michel, Belgium's PM since 2011, had read law at the ULB.

"Is the law so badly taught here? Or did he not pass his exam?" Loach said. "A good lawyer must examine the evidence before coming to a conclusion. Mr Michel, look at the evidence."

On the previous evening, during a speech at Brussels Grand Synagogue to mark the 70th anniversary of Israel's foundation, Michel had surprised his audience by commenting on the row over Loach's planned honour, which local Jewish organisations have been lobbying to have withdrawn in recent weeks.

Loach has faced heavy criticism in recent weeks over claims he has given cover to antisemitic views. He has had to defend himself from claims that he failed to condemn Holocaust denial during an interview with the BBC. [*The Guardian*, by Daniel Boffey, 26 April 2018, 240 tokens]

B. Belgian lawmakers want an end to German pensions for Nazi collaborators

In the aftermath of World War II, around 80,000 Belgians were convicted of collaborating with the SS and of committing war crimes, according to lawmakers.

Belgian lawmakers want Germany to stop paying war pensions to Nazi collaborators, saying the payments are a contradiction of the European Union's founding principles. The Belgian parliament's Foreign Relations Committee passed a resolution Tuesday calling on its government to ask Germany to stop providing the tax-free money. It also stressed the "injustice" of such payments because victims of Nazism do not receive similar allowances.

In the aftermath of World War II, around 80,000 Belgians were convicted of collaborating with Adolf Hitler's SS and of committing war crimes while the country was occupied by the Nazis, according to lawmakers. However, some of these people benefited from a series of Hitler's decrees, including the right to German nationality and a war pension in exchange for their actions.

The committee's resolution said that the pensions were "for collaboration with one of the most murderous regimes in history." The leader of the DéFI party lodged the original version of the resolution in 2016. A spokeswoman for the party said Wednesday that up to 27 people are still

believed to be receiving the payments. Recipients' names are only known by the German ambassador and had not been shared with the Belgian government, according to lawmakers. It is not clear how much the convicted collaborators receive each month. Some people living in Britain were also receiving the payments, the committee said.

Belgium's parliament is due to vote on the resolution next month. British officials did not immediately respond to requests for comment. The German Federal Ministry of Labor and Social Affairs said in a statement that no benefits are paid to former SS members.

[*NBC News*, by Saphora Smith, 20 February 2019, 300 tokens]

C. Belgium hands powers to caretaker PM to fight Covid-19 after 15-month stalemate

Warring parties agree to let Sophie Wilmès's minority government deal with impact

After 15 months of stalemate over forming a government, Belgium's warring political parties have agreed to hand special powers to fight coronavirus to a minority administration led by the current caretaker prime minister, Sophie Wilmès.

Wilmès saw King Philippe on Monday, who officially nominated her to form a government – the first time a woman has been given the task in Belgium's 190-year history. The federal parliament is expected to approve the deal in a confidence vote on Thursday.

Announcing the breakthrough, Wilmès tweeted that her governmental team appreciated the great responsibility conferred on it. "The sense of duty drives us. The wish to work in the interest of all Belgians equally. This great union [Belgium] is equal to the challenges of the moment."

Wilmès, Belgium's first female prime minister, has been leading a caretaker administration since last October, after the previous government collapsed in December 2018.

After hours of talks over the weekend, Belgium's political parties agreed to put aside differences that have stymied the creation of a fully fledged government since federal elections last May.

The Wilmès government, composed of two liberal and one Christian Democratic party, will now be supported by seven other parties across the political spectrum, save the far right and far left.

These parties have agreed to award the government special powers to sidestep the normal legislative process in order to deal with the impact of coronavirus on Belgium's health system and economy. The special powers will last three months and can be renewed once only for up to a further three months.

The global pandemic, which has led to 1,058 cases and five deaths in Belgium as of Monday, forced a rethink. Last week Wilmès announced the closure of schools, kindergartens, bars and restaurants, and a ban on weekend shopping, excluding supermarkets and pharmacies.

[*The Guardian*, by Jennifer Rankin, 16 March 2020, 319 tokens]

Appendix 2: French source texts used in the study

D. Le Canada est-il l'Eldorado rêvé pour les Belges ?

● Quelques centaines de Belges s'expatrient chaque année au Canada.
● Ils vantent la qualité de vie et les opportunités professionnelles.
● Mais pointent un gros écueil : l'équivalence des diplômes.

« L'immigration, c'est un saut dans le vide et certains ont un meilleur parachute que d'autres. » La formule est de Caroline Guimond, ministre conseiller à la Migration, à l'ambassade canadienne de Paris. Voilà pourquoi certains se plaisent au Canada, s'y installent à long terme et ne rentrent au pays que pour les vacances, tandis que d'autres renoncent au bout de quelques années. Mais, assure la ministre conseiller, « la Belgique est l'un des pays préférés dont le Canada recherche les migrants ». Autrefois, c'était surtout des fermiers, des bûcherons, des travailleurs de l'industrie, des hommes d'affaires qui partaient. Aujourd'hui, outre les aventuriers, ce sont aussi des travailleurs de l'Horeca, des personnes qui démarrent ou reprennent une entreprise.

Car le Canada se veut pays d'immigration. Un cinquième de ses habitants est né hors du pays. Depuis 1990, plus de 6 millions d'immigrants sont arrivés au Canada. Qui a défini une stratégie migratoire précise et chiffrée, avec des cibles d'admission annuelles pour les résidents permanents : 0,8 % de la population. Le pays vise prioritairement trois catégories. Un, l'immigration économique (58 % en 2017) : il veut attirer les talents, les investisseurs, les travailleurs (très) qualifiés, une main-d'œuvre spécifique pour des secteurs en pénurie. Deux, l'immigration familiale (28 %) : il veut permettre aux familles (époux, enfants, parents, grands-parents) de se regrouper. Trois, l'immigration humanitaire (13 %) : il offre la protection à un certain nombre de réfugiés. [*Le Soir,* by Martine Dubuisson, 19 March 2018, 275 tokens]

E. Les Belges restent sur Facebook, malgré les scandales à répétition

Le Belge reste actif sur Facebook même si sa consommation a diminué et s'est diversifiée, alors que le réseau social a fêté son 15e anniversaire le 4 février dernier. Les scandales qui ont secoué l'entreprise ces dernières années et ses démêlés judiciaires n'ont pas poussé les Belges à se déconnecter, constate Xavier Degraux, consultant en médias sociaux.

Créé en 2004, Facebook s'est popularisé en Belgique à partir de 2008 et comptait environ 7,3 millions d'utilisateurs belges l'année dernière, selon les chiffres du réseau social transmis aux annonceurs. Entre six et sept millions de Belges sont actifs mensuellement.

Ces statistiques sont toutefois difficiles à analyser dans le détail. Le nombre d'utilisateurs regroupe les personnes qui ont fait au moins une action sur la plateforme dans le dernier mois et ne donne donc aucune indication sur la fréquence d'utilisation.

Selon Xavier Degraux, quelque sept millions d'utilisateurs belges de Facebook sont majeurs, tandis que la catégorie des 13 à 17 ans représente environ 300.000 profils. Le consultant constate d'ailleurs un désintérêt progressif des jeunes, globalement compensé par l'inscription de personnes

plus âgées. « Ces utilisateurs plus jeunes ne quittent toutefois pas le groupe Facebook car on remarque un succès grandissant pour Instagram, voire pour les messageries instantanées comme Messenger et WhatsApp », qui appartiennent aussi au géant américain.
Les scandales qui ont secoué l'entreprise ces dernières années n'ont pas non plus engendré de conséquences négatives sur le volume de membres en Belgique et à travers le monde. Facebook s'est notamment excusé après la transmission des données personnelles de 87 millions d'utilisateurs à la société Cambridge Analytica et le tribunal de première instance de Bruxelles a jugé l'année dernière que le réseau social ne respectait pas la législation belge sur la protection des données à caractère personnel. [*Le Soir*, 31 January 2019, 308 tokens]

F. Une app pour dépister le coronavirus depuis chez soi

L'application Well@Home est disponible gratuitement sur iOS et Android. Une version web et anonyme existe également.
Alors que le pic de l'épidémie se profile à peine, les services d'urgence sont déjà surchargés de malades. Pour tenter de venir en aide aux hôpitaux et corps soignant, l'entreprise belge BeWell Innovations a développé une app de screening des patients, Well@Home.
L'app repose avant tout sur un questionnaire qui permet de faire le tri entre les patients potentiellement atteints par le coronavirus. Les personnes qui pensent présenter des symptômes du virus n'ont qu'à télécharger l'application disponible sur iOS et Android et répondre aux différentes questions médicales pour avoir un diagnostic. En fonction des réponses, l'application classera les utilisateurs en plusieurs catégories ; sain, modéré ou à haut risque.
En fonction de la catégorie, l'application conseillera aux utilisateurs de rester confinés chez eux, de surveiller l'évolution des symptômes et de remplir régulièrement le questionnaire ou de prendre directement contact avec leur médecin traitant ou les services d'urgence. À noter que l'application ne remplace pas un examen médical réalisé par un professionnel. Si vous avez toujours un doute, vous pouvez vous rendre sur le site officiel www.info-coronavirus.be ou prendre contact avec votre médecin traitant.
« Avant toute chose, le questionnaire de tri et la consultance à distance permettent d'effectuer immédiatement une première sélection importante, ce qui offre aux médecins généralistes et au personnel hospitalier la possibilité de se concentrer sur les personnes effectivement infectées. À cela vient s'ajouter un autre atout : les personnes ne doivent pas quitter leur domicile et évitent ainsi de propager le virus. Enfin, grâce à l'historique de la plateforme en ligne, tous les patients à risque peuvent être suivis efficacement par le médecin ou le poste de garde », explique Alain Mampuya, CEO de BeWell Innovations.

[*Le Soir*, by Jennifer Mertens, 30 March 2020, 317 tokens]

Appendix 3: Linear mixed-effects simplicity models

Mean sentence length

Random effects

Random effect	**Confidence intervals**	
	2.5%	97.5%
sd_(Intercept)|student	0.02510293	0.08386450
sd_(Intercept)|source_text	0.04770541	0.20139511
sigma	0.05864973	0.09156073

Fixed effects

Fixed effect	**Sum Sq**	**Mean Sq**	**Num*DF***	**Den*DF***	***F*-value**	***Pr*(>*F*)**
translation_direction	0.0002910	0.0002910	1	3.694	0.0557	0.8259
self-reported_EN_proficiency	0.0004669	0.0004669	1	54.588	0.0894	0.7661
years_studying_EN	0.0000971	0.0000971	1	35.589	0.0186	0.8923
stays_EN_countries	0.0000067	0.0000067	1	36.566	0.0013	0.9715
translation_experience	0.0105025	0.0105025	1	37.129	2.0106	0.1645

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Root lemma-token ratio

Random effects

Random effect	**Confidence intervals**	
	2.5%	97.5%
sd_(Intercept)|student	0.00000000	0.01913431
sd_(Intercept)|source_text	0.03204330	0.10894676
sigma	0.02780577	0.03975804

Fixed effects

Fixed effect	**Sum Sq**	**Mean Sq**	**Num*DF***	**Den*DF***	***F*-value**	***Pr*(>*F*)**
translation_direction	0.00012811	0.00012811	1	3.949	0.1090	0.7580
self-reported_EN_proficiency	0.00002216	0.00002216	1	36.482	0.0189	0.8915
years_studying_EN	0.00009971	0.00009971	1	36.144	0.0848	0.7725
stays_EN_countries	0.00072690	0.00072690	1	36.754	0.6185	0.4366
translation_experience	0.00302199	0.00302199	1	36.937	2.5713	0.1173

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Lexical density

Random effects

Random effect	**Confidence intervals**	
	2.5%	97.5%
sd_(Intercept)|source_text	0.01531295	0.05301503
sigma	0.02099664	0.02873491

Fixed effects

Fixed effect	**Sum Sq**	**Mean Sq**	**Num*DF***	**Den*DF***	***F*-value**	***Pr*(>*F*)**
translation_direction	0.0018892	0.0018892	1	3.895	3.0166	0.159343
self-reported_EN_proficiency	0.0002262	0.0002262	1	74.391	0.3612	0.549691
years_studying_EN	0.0003214	0.0003214	1	74.709	0.5131	0.476021
stays_EN_countries	0.0002285	0.0002285	1	74.527	0.3649	0.547626
translation_experience	0.0046938	0.0046938	1	77.566	7.4948	0.007671 **

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Core vocabulary coverage

Random effects

Random effect	**Confidence intervals**	
	2.5%	97.5%
sd_(Intercept)|source_text	0.02606353	0.09048242
sigma	0.03638193	0.04978948

Fixed effects

Fixed effect	**Sum Sq**	**Mean Sq**	**Num*DF***	**Den*DF***	***F*-value**	***Pr*(>*F*)**
translation_direction	0.0036376	0.0036376	1	3.892	1.9345	0.23851
self-reported_EN_proficiency	0.0001149	0.0001149	1	74.403	0.0611	0.80545
years_studying_EN	0.0012534	0.0012534	1	74.729	0.6666	0.41684
stays_EN_countries	0.0000149	0.0000149	1	74.543	0.0079	0.92929
translation_experience	0.0074416	0.0074416	1	77.612	3.9576	0.05018 .

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Appendix 4: Linear mixed-effects simplification models

Mean sentence length

Random effects

Random effect	**Confidence intervals**	
	2.5%	97.5%
sd_(Intercept)|student	0.02184421	0.08121727
sd_(Intercept)|source_text	0.03378136	0.13992754
sigma	0.05886930	0.09241484

Fixed effects

Fixed effect	**Sum Sq**	**Mean Sq**	**Num*DF***	**Den*DF***	***F*-value**	***Pr*(>*F*)**
translation_direction	0.0157761	0.0157761	1	3.483	3.0062	0.16850
self-reported_EN_proficiency	0.0007470	0.0007470	1	54.374	0.1423	0.70743
years_studying_EN	0.0002137	0.0002137	1	35.881	0.0407	0.84122
stays_EN_countries	0.0000848	0.0000848	1	36.787	0.0162	0.89954
translation_experience	0.0189295	0.0189295	1	37.946	3.6071	0.06515 .

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Root lemma-token ratio

Random effects

Random effect	**Confidence intervals**	
	2.5%	97.5%
sd_(Intercept)|student	0.00000000	0.01890130
sd_(Intercept)|source_text	0.01519384	0.05954933
sigma	0.02785209	0.03975725

Fixed effects

Fixed effect	**Sum Sq**	**Mean Sq**	**Num*DF***	**Den*DF***	***F*-value**	***Pr*(>*F*)**
translation_direction	0.0186142	0.0186142	1	3.823	15.7970	0.01799 *
self-reported_EN_proficiency	0.0000200	0.0000200	1	36.450	0.0170	0.89695
years_studying_EN	0.0001231	0.0001231	1	36.499	0.1045	0.74834
stays_EN_countries	0.0006888	0.0006888	1	37.023	0.5845	0.44940
translation_experience	0.0037351	0.0037351	1	38.508	3.1698	0.08290 .

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Lexical density

Random effects

Random effect	**Confidence intervals**	
	2.5%	97.5%
sd_(Intercept)|student	0.01804954	0.03901723
sigma	0.02329567	0.03580840

Fixed effects

Fixed effect	**Sum Sq**	**Mean Sq**	**Num*DF***	**Den*DF***	***F*-value**	***Pr*(>*F*)**
translation_direction	0.141883	0.141883	1	41.654	170.4503	2.583e-16 ***
self-reported_EN_proficiency	0.001183	0.001183	1	62.312	1.4216	0.23766
years_studying_EN	0.002450	0.002450	1	36.396	2.9435	0.09473 .
stays_EN_countries	0.001161	0.001161	1	37.247	1.3947	0.24509
translation_experience	0.000353	0.000353	1	36.366	0.4244	0.51885

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Core vocabulary coverage

Random effects

Random effect	**Confidence intervals**	
	2.5%	97.5%
sd_(Intercept)|student	0.00000000	0.02387699
sigma	0.03848733	0.05211882

Fixed effects

Fixed effect	**Sum Sq**	**Mean Sq**	**Num*DF***	**Den*DF***	***F*-value**	***Pr*(>*F*)**
translation_direction	0.37435	0.37435	1	42.358	177.5806	< 2.2e-16 ***
self-reported_EN_proficiency	0.00020	0.00020	1	36.412	0.0949	0.7598467
years_studying_EN	0.00023	0.00023	1	37.753	0.1104	0.7415374
stays_EN_countries	0.00005	0.00005	1	38.110	0.0238	0.8781978
translation_experience	0.03108	0.03108	1	37.572	14.7422	0.0004586 ***

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1