LINGUISTIC COMPLEXITY AND SIMPLIFICATION IN TRANSLATION: COGNITIVELY-GROUNDED PHONOLOGICAL METRICS

Mariam ALMIHMADI
College of Social Sciences, Umm Al-Qura University, Makkah, Saudi Arabia

mmmihmadi@uqu.edu.sa

Manuscript received 22 January 2023
Manuscript accepted 7 December 2023
**Corresponding author*
https://doi.org/10.33736/ils.5356.2023

ABSTRACT

Translated text (TT) is characteristically simpler than non-translated (NTT) authentic text in terms of its lexicon, syntax, and style (Laviosa-Braithwaite, 2001). It is still not entirely clear what causes this phenomenon, and scholars continue to debate the issue. The traditional metrics that are implemented in the simplification literature are often criticised as unreliable and lacking cognitive grounding. This paper addresses this limitation in the literature and proposes a paradigm that uses complexity-based measures adopted from phonology and cognitive psychology. Calculations are run on a corpus of 100 translated and non-translated article abstracts from five academic disciplines. Statistical analyses reveal that TTs use shorter words from dense phonological neighbourhoods. The findings suggest that adopting a cognition-informed approach is essential in elucidating the process of simplification. The results are relevant to the issues of universality and multidimensionality of translation as a form of constrained communication.

Keywords: linguistic complexity; neighbourhood density; phonological complexity; phonotactic probability; simplification

Introduction

Decades-long research into the linguistic features that are postulated to distinguish translated (TTs) and non-translated texts (NTTs) has yielded empirical evidence that TTs

are lexically, syntactically, and stylistically simpler than NTTs (Hu, 2016; Laviosa-Braithwaite, 2001). Linguistic differences between TTs and NTTs exhibit such consistent patterns that they can be detected by machine classification algorithms, resulting in the successful identification of TTs at a high accuracy rate (Rabinovich & Wintner, 2015).

Several corpus-based studies analysing a variety of text genres from different languages have revealed that far more function words appear in TTs than in NTTs, rendering TTs lexically less dense. At the same time, highly frequent lexical items appear more often in TTs than in NTTs (see the next section for references). These features have been taken as diagnostics of simplification, which has been treated in corpus-based translation literature as a translation universal (Baker, 1993, 1996; Laviosa, 2002). As a putative universal, simplification is believed to be "inherent in the translation process itself" (Baker, 1993, p. 243). Simplification is presumed to apply subconsciously and is inaccessible to the translator (Baker, 1996; Olohan & Baker, 2000). However, there is no evidence, to date, supporting this process-inherent automaticity claim.

The standard research paradigm in simplification studies looks for evidence for simplification along lexical parameters that are likely to be under the translator's conscious control. These parameters include sentence length, use of varied lexicon, presence of frequent lexical items, and prevalence of function words. Translation, as a form of mediated writing, involves some form of deliberation while making conscious decisions regarding word choice, register, sentence size, etc. As Roehr-Brackin (2015) puts it, translators make a "conscious effort to analyse the input and control the output" (pp. 118–119). More recently, Wang (2020) writes, "the translation process is a psychological and introspective process, as well as a process of problem solving and decision-making" (p. 1413). The traditionally used metrics indicate that simplification happens, but they do not entail that it happens subconsciously. The current study is an attempt to fill this gap in the literature.

The study aims to explore the issue of simplification using sub-lexical phonological metrics that tap into the cognitive component of the translation process. Unconscious priming (e.g., Bazan et al., 2019) provides evidence that phonology is processed unconsciously. Phonological Complexity, Neighbourhood Density, and Phonotactic Probability are established metrics in phonological and cognitive psychology literature, with documented effects on word recall, processing, and learning. Despite their well-documented utility, these metrics have yet to be incorporated into simplification studies. The research questions that this study seeks to address are as follows:

1. Are translated texts phonologically less complex than non-translated texts?
2. Do translated texts use more words from denser phonological neighbourhoods?
3. Do translated texts use more words with higher phonotactic probabilities?

Literature Review

Simplification in Translation

Simplification is defined as "the idea that translators subconsciously simplify the language or message or both" (Baker, 1996, p. 176). Since its inception as a translation universal in the 1990s, simplification has been firmly associated with a simpler language in TTs. Inspired by the notions put forth in Baker's seminal study (1993), numerous studies have examined monolingual comparable corpora of TTs and NTTs looking for translation universals, including simplification (see Robin, 2017, for a review). Three types of simplification are investigated in the literature: lexical, syntactic, and stylistic (Laviosa-Braithwaite, 2001). However, for scope reasons, the current study focuses on lexical simplification. Table 1 summarises some of the most frequently used measures of lexical simplification.

Table 1

Summary of the Most Frequently Used Measures of Lexical Simplification

Parameter	Description	Expected Difference
Sentence Length	The average number of words in a sentence in a given text.	TTs < NTTs
Lexical Density	The ratio of lexical words to function words in a given text.	TTs < NTTs
Type-token Ratio	The proportion of the number of different words (types) to the total number of running words (tokens) in a given text.	TTs < NTTs

Note. < = less than.

A cross-examination of the relevant literature reveals that sentence length is genre- and language-dependent. For example, it is found to be shorter in translated than non-translated newspaper articles, whereas the opposite pattern is reported for fiction (Laviosa, 1998). Similarly, Liu and Afzaal (2021) reported a genre-based effect, where translated prose and academic writing are less complex, whereas translated fiction is more complex than NTTs. Williams (2005), however, reported mixed results for English and French. Sentences in translated French texts from English were 12% shorter than sentences in non-translated French texts. This difference was statistically significant. In contrast, sentence-length differences were going in the opposite direction for English-translated and non-translated texts by 15% but did not reach statistical significance.

Several studies have examined TT corpora from a variety of languages and reported lower lexical density in TTs than in NTTs. These studies include Laviosa (1998) on translated English from multiple languages, Hu (2016), and Xiao and Dai (2014) on translated Chinese from English, and Williams (2005) on translated English from French. However, the Williams (2005) study also reported a statistically non-significant difference in lexical density between translated and non-translated French texts going in the opposite direction. Similarly, Ferraresi et al. (2019) found no significant differences in lexical density for translated English from French. But their translated English texts from Italian were even lexically denser.

The above studies also examined type-token ratios and reported inconsistent findings. While Laviosa (1998) and Hu (2016) found differences in the expected direction: lower type-token ratios in TTs, Ferraresi et al. (2019) found no significant differences in English texts translated from French or Italian. Likewise, Xiao and Dai (2014) reported no significant differences between Chinese TTs and NTTs. Williams (2005), however, reported the same language-dependent pattern as seen for lexical density above. The type-token ratio in English-translated texts is significantly lower than in English NTTs. The difference, which goes in the opposite direction, between French TTs and NTTs does not reach statistical significance.

The mixed findings enumerated above have been variously attributed to genre, language, and/or modality effects. Some commentators suggest that the incompatible requirements of the genres are responsible for the mixed findings, as more explicitness is desired in some text types than in others (e.g., Kruger & van Rooy, 2012). Other researchers contend that the discrepancies are language-dependent (e.g., Williams, 2005). Simplification is observed in TTs from certain source languages, but not others. Yet a third group of researchers appeal to the modality of the activity, suggesting that simplification is greater in interpreting than in translation (e.g., Sandrelli & Bendazzoli, 2005).

Linguistic Complexity Metrics

Phonological Complexity

In the literature on cognitive linguistics, phonological complexity is often defined quantitatively in terms of the number of sounds or syllables in a word (Mueller et al., 2003). Experiments on verbal working memory show that shorter words are recalled more accurately than longer words (e.g., Baddeley et al., 1975; Longoni et al., 1993). This is known as the word-length effect or short-word advantage (Tehan & Tolan, 2007).

Numerous studies have also demonstrated that words that are less phonologically complex are acquired earlier (Braginsky et al., 2019; Gendler-Shalev et al., 2021). Commenting on short-word advantage in early vocabulary production, Gendler-Shalev et al., (2021) suggest that it “represents a subconscious selection of less complex words” (p. 790). Similarly, Braginsky et al., (2019), who explored consistency and variability in “acquisition trajectories of around 400 words in each of 10 languages” (p. 53)

using data from more than 32,000 children, reported a large effect of phonological complexity on production. Shorter words are more likely to be produced by more children. In reading comprehension, longer words are also said to slow down processing by leading to longer pauses and eye fixations (Rojo López, 2015). Importantly, Edwards et al. (2011) contend that the development of the lexicon goes hand in hand with the development of phonology, and hence, they should not be studied independently of each other.

The present study is the first to bring phonology into the realm of lexical simplification in translation. In this paper, Phonological Complexity represents ontological complexity (Forker, 2021; Rescher, 1998). The more sounds a text has, the more complex it is. So, if simplification obtains, TTs will have words composed of fewer sounds, on average, than in NTTs.

Neighbourhood Density

Neighbourhood Density (ND) is an instantiation of "phonological regularity" in terms of "sound similarity" (Freedman & Barlow, 2012, p. 371). ND generally refers to the cluster of words that are maximally similar to a given word (w), differing by one phone from w (Luce & Pisoni, 1998). Words in such a cluster are called neighbours. For example, "cat", "pat", "rat", and "sat" are neighbours. The larger the size of the cluster, the denser the neighbourhood. Numerous studies have documented the behavioural effects of ND on various aspects of language processing. Participants' performance is traditionally measured in terms of the speed and accuracy of their responses. The exception here is retrieval effort. In pupillometric studies, retrieval effort is indexed by pupil response (e.g., Goldinger & Papesh, 2012; Laeng et al., 2012).

Results from language processing research reveal a high-ND disadvantage in recognition among monolinguals (e.g., Luce & Pisoni, 1998; Marian & Blumenfeld, 2006). Items from dense neighbourhoods are recognised more slowly than items from sparse neighbourhoods. This is expected given that recognition involves deciding among competing alternatives. With bilinguals, high-ND in the non-target language (the language not being tested) is reported to have an inhibitory effect on the recognition of the items presented in the target language (e.g., Chen & Sie, 2019; Dirix et al., 2017). In contrast, high-ND items are recalled more rapidly and accurately than low-ND items (e.g., Allen & Hulme, 2006; Roodenrys et al., 2002; Storkel et al., 2006). ND is also found to have a facilitatory effect on production by monolinguals (e.g., Jones, 2018; Marian & Blumenfeld, 2006; Vitevitch, 2002). This high-ND advantage in production is also reported for bilinguals (e.g., Bradlow & Pisoni, 1999; Stamer & Vitevitch, 2012).

Items from denser neighbourhoods are also learned faster and more accurately by monolinguals and bilinguals than items from sparse neighbourhoods (e.g., de Groot et al., 2002; Jones, 2018; Stamer & Vitevitch, 2012, for monolinguals, and Nair et al., 2017, for bilinguals). Interestingly, the high-ND disadvantage in recognition is also replicated in the pupillometric study by Schmidtke (2014) for both monolinguals and bilinguals.

Schmidtke (2014) reported that words from dense neighbourhoods were retrieved with greater effort.

This paper is the first to utilise Neighbourhood Density to quantify simplification. Neighbourhood Density concerns the functional aspect of linguistic complexity, which involves the notion of cost associated with processing and production. Since high-ND items are easier to recall and produce, TTs are expected to have more words from dense neighbourhoods.

Phonotactic Probability

Phonotactic Probability (PP) is generally defined as the likelihood of the occurrence of a sound or cluster of sounds in a given sequence for a given text (Vitevitch et al., 1999). For example, in English, as a word-initial sequence, [pr] is phonotactically probable; the sequence [pn] is improbable in word-initial positions. Several studies have reported a facilitatory effect of high PP on item recognition and recall in terms of speed and accuracy for both monolinguals and bilinguals. For example, with monolinguals, Luce and Large (2001), Vitevitch et al. (2004), and Vitevitch et al. (1999) found that high-probability sound combinations are recognised faster and more accurately than sound sequences with low PP. For bilinguals, Lee (2011) reported that high-probability sequences elicited faster and more accurate responses. Similarly, the speed and accuracy of non-word recall by monolinguals are reported to be greater for high-PP items (e.g., Thorn & Frankish, 2005). Experimental work with bilinguals reveals a similar effect. For example, Messer et al., (2015) reported a recall advantage for high-PP. High-PP is also reported to have a facilitatory effect on non-word production by monolinguals and bilinguals (e.g., Edwards et al., 2004; Freedman & Barlow, 2012).

In terms of non-word learning, converging evidence points to a low-PP advantage that characterises monolingual and bilingual learning. Low Phonotactic-Probability non-words trigger learning because they sound different from other known words. Hoover et al. (2010), Storkel et al. (2006) and Storkel (2009) reported that their monolingual participants learned a larger number of low-PP non-words. This facilitatory effect on word learning is also reported in experimental studies with bilinguals (e.g., Chen & Sie, 2019; Nair et al., 2017).

This paper is the first to implement Phonotactic Probability in the study of simplification. Phonotactic Probability is a measure of functional complexity, involving processing and production cost. Since high-PP items are easier to recall and produce, more words with frequent sound combinations in TTs would mean simplification.

Theoretical Framework

This study is guided by the concepts and principles originally developed in corpus linguistics and adopted later in corpus-based descriptive translation studies. The corpus-based framework is empirical: it describes and analyses actual language use, rather than

idealised language data (Laviosa, 1998). This framework aims to identify linguistic features that distinguish one form of text from another and to discover linguistic regularities in terms of patterns of repetitions, similarity, and co-occurrence, which underlie the phonology-based metrics proposed in the current study. For more on the corpus-based translation paradigm, see Baker (1993).

The paper analyses monolingual comparable corpora of TTs and NTTs, testing for simplification using phonological metrics that tap into the functional notions of cost and effort which impact human cognition. The principle of least effort (Zipf, 1949) is presumed to be behind the observation that, whenever there is a choice, human activities display a preference for behaviours and paths that require minimal effort. Simplification in translation could be envisaged to follow this cognitive predisposition. Also relevant is the insight from the theory of cognitive economy (Rescher, 1989) regarding the tendency of cognitive processes to minimise cost. Translation as a process that involves recasting a message from one linguistic system into another comes at a cognitive cost in terms of language and message processing, recall, and production. As explained in the previous section, the phonology-based metrics in this study are relevant to these aspects of human cognition.

Methodology

The Corpus

As presented in Table 2, the corpus is made up of abstracts of 100 research articles published between 2018 and 2022 in 10 academic journals. This genre is selected based on the presumption that abstract writing is very constraining and likely to foster a high-fidelity reproduction of the original content into another language. The abstracts are all written in English, with 50 translated from Arabic. Translated abstracts are all from articles written in Arabic, with the abstract being the only section that appears in both Arabic and English. The abstracts come from five academic disciplines selected from the 10 main classes of the Dewey Decimal Classification (Dewey, 1876). Twenty abstracts (10 TTs and 10 NTTs) were extracted from each of these disciplines.

Since the three sub-lexical metrics are calculated over phonological units (sounds), not orthographic characters (letters), the corpus abstracts are converted into phonetically transcribed texts using the International Phonetic Alphabet (The International Phonetic Association, 1999). For phonetic conversion, the web-based edition of the to-Phonetics Converter was utilised (Mu-Sonic Ltd, 2013). Phonetic forms were imported from the open Carnegie Mellon University Pronouncing Dictionary as implemented in to-Phonetics Converter. In the corpus, there are 20523 words spelled with 113414 letters corresponding to 104883 sounds, with 8531 extra letters.

Table 2
Corpus Sources

Class	Division	Journal	Vol. (issue)	Year	Publisher
200 Theology	290 Non-Christian Religion	Journal of Islamic Studies	29 (2, 3) 30 (1, 2, 3) 31 (2, 3) 32 (1)	2018 2019 2020 2021	Oxford University Press
		Journal of Islamic Studies	32 (2, 3)	2020	King Saud University
300 Sociology	370 Education	Journal of Education	200 (3) 201(1)	2020 2021	Sage
		Journal of Educational Sciences	32 (3) 33 (1)	2020 2021	King Saud University
600 Useful Arts	630 Agriculture	Journal of Agriculture and Food Research	7	2022	Elsevier, ScienceDirect
		Scientific Journal of King Faisal University: Basic and Applied Sciences	21 (1, 2)	2020	King Faisal University
700 Fine Arts	720 Architecture	The Journal of Architecture	26 (7, 8)	2021	Taylor & Francis Group
		Journal of Architecture and Planning	33 (3, 4) 34 (2)	2021 2022	King Saud University
900 History	910 Geography	The Geographical Journal	188 (2)	2022	Wiley
		Arab Geographical Journal	53 (78, 79)	2022	Société De Geographie De Egypte

Data Processing

Prior to phonetic conversion, punctuation marks and numbers were removed from the corpus. This was done using a Python script. Next, the corpus was IPA-transcribed. The number of words, characters, and sounds in each text were extracted via a Python script. Phonological-Complexity values were calculated using Formula 1.

Formula 1: $$PC = \frac{\sum sound\ segmenst}{\sum words}$$

Neighbourhood-Density and Phonotactic-Probability values for each of the 100 texts were obtained from Phonological Corpus Tools (PCT), version 1.5.1 (Hall et al., 2022). The query for Neighbourhood Density used Levenshtein Edit Distance as the string similarity algorithm and was run for all words in each text. Pronunciation variants were set to "canonical forms only". Phonotactic Probability calculation was based on the algorithm by Vitevitch and Luce (1998). The query was set to calculate Phonotactic Probability for all words in each text. Again, Pronunciation variants were set to "canonical forms only". Biphone probabilities were computed using token counts. For more information on the method, see Hall et al. (2022) and Vitevitch and Luce (1998).

Statistical Analysis

In this study, Phonological Complexity, Neighbourhood Density, and Phonotactic Probability make the Dependent Variables (DVs) of the study. TextType (TT or NTT) is the Independent Variable (IV). However, given that the 100 texts are drawn from 10 journals that are themselves drawn from five academic disciplines, it is possible to find greater similarity among observations from the same journal and/or discipline, which can lead to clustering effects and non-independence in the data. The dataset has a nested hierarchical structure. The present study comprises a corpus of 100 abstracts, which are distributed among 10 scholarly journals that are classified within five distinct disciplinary fields.

In linguistics and behavioural sciences, the go-to statistical tool to capture the nestedness of datasets is the mixed-effects model (Baayen et al., 2008; Barr et al., 2013). According to Titz (2020), Mixed-effects modelling "is rapidly becoming the gold standard of statistical analysis in the behavioural sciences" (p. 1). However, although translation corpora are intrinsically hierarchical, this type of statistical analysis is still under-reported in translational studies but see De Sutter and Lefer (2020) for an important contribution in this area.

The current study fits mixed-effects models to test for simplification along three DVs. For each of these DVs, a null model with Journals nested within Disciplines is fitted first. The purpose of that is to assess whether there is a need for multilevel modelling. Multilevel modelling can detect and handle any clustering effects among the observations. It is important to ensure that the lack of independence among the observations, if present, is statistically handled. According to Garson (2020), ignoring "heteroskedastic error variance […] will lead to inaccurate standard errors and significance tests" (p. 57). The present study involves an evaluation of the contribution made by the random effects of Journals that are nested within Disciplines, with Disciplines serving as the upper-level units. In mixed models, interclass correlation coefficients (ICCs) are used for this purpose (see Garson, 2013). The ICC gives an indication of the "effect size" of the upper-level

grouping variables, which in turn, determines whether a given DV is independent of the grouping variable. The closer ICC value is to zero, the little effect the grouping has.

Model fitting and statistical analyses were performed in R (R Core Team, 2022) using the lmer4 package (Bates et al., 2015) and the lmerTest package (Kuznetsova et al., 2017). Satterthwaite's (1946) method was used to compute degrees of freedom for significance testing and for the calculation of *p*-values. To select the best-fitting model for each DV, the Akaike Information Criterion (AIC) was used. A lower AIC signifies better fit and less error. The alpha level was set to .05.

Results

The statistical analysis in this study answers these research questions: (1) Are translated texts phonologically less complex than non-translated texts?, (2) Do translated texts use more words from denser phonological neighbourhoods?, and (3) Do translated texts use more words with higher phonotactic probabilities? Results reveal that TTs have smaller phonological-complexity and Phonotactic-Probability mean values, but a higher Phonological Neighbourhood-Density mean value. Table 3 gives the summary statistics of TTs and NTTs pooled across the disciplines.

Table 3

Mean and Standard Deviation (SD) Values of TTs and NTTs along the metrics of the Study

	Non-Translated	Translated
Phonological Complexity	5.41 (0.398)	4.95 (0.352)
Phonological Neighbourhood Density	0.270 (0.0811)	0.308 (0.104)
Phonotactic Probability	0.0316 (0.006)	0.0297 (0.004)

Comparison of Phonological Complexity of Translated Texts and Non-Translated Texts

As can be seen from Table 4, TTs are consistently less phonologically complex than NTTs. On average, words from TTs are composed of fewer sounds than words from NTTs. Education and Agriculture display the highest values in the corpus. This means that TTs and NTTs from these disciplines use, on average, longer words than the rest of the disciplines in the study. To test the statistical significance of these differences, three models were fitted. The empty model with the random effect of Journals nested in Disciplines ran into the Singular Fit problem. This happens when the variances/covariances of one of the random effects are already accounted for by the other random term. This is usually a sign of overfitting. To solve this problem, Barr et al., (2013) recommend simplifying models by dropping random effects with little contribution to model fit. In the current study, the variance of the upper-level variable (DisciplineID) was 0.00, with ICC= 0.00. It was therefore dropped and the model was re-fitted using the nested term, which has an ICC score of 0.49. Now to find out whether Text Type (TTs vs

NTTs) can explain any amount of the variation in the data, a mixed model with this fixed-effect variable was fitted.

Table 4
Mean and (SD) Values of Phonological Complexity of TTs and NTTs

	Non-Translated	Translated
Education	5.78 (0.404)	5.10 (0.390)
Agriculture	5.49 (0.334)	5.17 (0.321)
Religion	4.99 (0.182)	4.65 (0.281)
Geography	5.43 (0.324)	4.97 (0.262)
Architecture	5.33 (0.291)	4.86 (0.280)

As can be seen from Table 5, which gives model fit statistics for both empty and best-fitted models, adding Text Type resulted in a significant improvement of the model fit ($\chi^2(1, N=100) = 7.13$, $p= .007$). This improvement is also seen in a lower AIC (=75.69) than for the empty model (AIC= 80.82). The fixed effect of Text Type is also statistically significant: $t(10)= –3.23$, $p= .009$, *CI* (2.5% – 97.5%)= -0.73 – -0.18. Note that this difference could be due to other factors not investigated here.

Table 5
Model Fit Statistics

	Empty Model PC ~ (1 | JournalID:DisciplineID)	Best-fitted Model PC ~ TextType+(1 | JournalID:DisciplineID)
AIC	80.82	75.69
LogLik	-37.41	-33.84
Deviance	74.82	67.69
N. Parameters	3	4
Model Comparison		
χ2	7.13	
d.f.	1	
p.	0.007	

Neighbourhood Density and Translated Text

Results of Neighbourhood Density show that TTs consistently have words from denser neighbourhoods. This result holds for both pooled data as presented in Table 3 and most discipline-grouped data as provided in Table 6. An empty model with the random effect of Journals nested in Disciplines was fitted. Neither random effect contributed significantly to the model. Their variances and ICC values were close to zero. For this

reason, multilevel modelling was dismissed and a simpler statistical test was used instead. A two-sample t-test found the Neighbourhood-Density difference between TTs and NTTs to be statistically significant at α=.05: *t*(98)= -2.014, *p*=.046, *CI* (2.5%–97.5%)= -0.074– -.0005.

Table 6

Mean and (SD) Values of Phonological Neighbourhood Density of TTs and NTTs

	Non-Translated	Translated
Education	0.239 (0.0842)	0.261 (0.0873)
Agriculture	0.268 (0.0784)	0.279 (0.102)
Religion	0.281 (0.104)	0.393 (0.104)
Geography	0.301 (0.0697)	0.289 (0.0823)
Architecture	0.263 (0.0684)	0.317 (0.109)

Phonotactic Probability and Translated Texts

Results of Phonotactic Probability reveal that TTs have words with lower phonotactic probability as calculated over the pooled data as shown in Table 3 and the disaggregated data for Education, Geography, and Architecture as presented in Table 7. The remaining disciplines (Agriculture and Religion) have the difference in the opposite direction. In the empty model with the random effect of Journals nested in Disciplines, both random effects failed to reach significance. Their variances and ICC values were close to zero. For this reason, multilevel modelling was dismissed, and a simpler statistical test was used instead. A Welch two-sample t-test found no significant Phonotactic-Probability difference between TTs and NTTs at α=.05: *t*(88.95)= 1.7822, *p*= .078, *CI* (2.5% – 97.5%)= -0.0002– 0.004.

Table 7

Mean and (SD) values of Phonotactic Probability of TTs and NTTs

	Non-Translated	Translated
Education	0.0392 (0.0059)	0.0321 (0.0035)
Agriculture	0.0302 (0.0027)	0.0309 (0.0046)
Religion	0.0290 (0.0034)	0.0298 (0.0051)
Geography	0.0301 (0.0077)	0.0282 (0.0050)
Architecture	0.0296 (0.0035)	0.0276 (0.0029)

Discussion

This study has explored the issue of simplification in translation within a paradigm that allows us to examine the automaticity claim, which seems to be tacitly assumed in various

operationalisations of simplification, but for which no conclusive evidence has been reported. Taken together, the results of the study reveal that TTs are indeed simpler than NTTs. On average, TTs have shorter words that come from denser phonological neighbourhoods. TTs are statistically distinguishable from NTTs in terms of their phonological complexity and neighbourhood density. Failure of the phonotactic probability difference between TTs and NTTs to reach statistical significance is unexpected, given the numerous reports in the literature which have documented the high Phonotactic-Probability advantage in recall, recognition, and production.

Since the current study is the first to investigate these phonology-based metrics in the context of simplification, it would be premature to dismiss any potential effect of Phonotactic Probability. To explain this lack of statistical significance, we need more studies tackling Phonotactic Probability in translation corpora. The present study adds to the repertoire of measurement that could be further evaluated within the framework of the revised research agenda for corpus-based translation studies that De Sutter and Lefer (2020) have proposed. It is hoped that the adoption of their revised agenda as a framework for future translation studies will lead to advances in theory, methodology, and analytics that will make it possible to expand the testing grounds for simplification beyond monolingual comparable corpora, as currently is the norm, to include bilingual parallel corpora, with the proviso that any confounding factors that potentially arise as a result of comparing disparate linguistic systems are appropriately controlled for.

The new agenda defines translation as "an inherently multidimensional linguistic activity and product, which is simultaneously constrained by sociocultural, technological and cognitive factors" (De Sutter & Lefer, 2020, p. 1). The cognitively grounded phonological metrics proposed here are in line with De Sutter and Lefer's (2020) call for "exploring new, more sophisticated linguistic indicators" (p. 19). More research is needed to test the validity of these metrics in translation corpora from other languages and other genres.

Relatedly, the need for a better understanding of the structure of translation data cannot be overemphasised. Even in a limited study such as the current one, which implements a basic multilevel model, it is clear that multilevel modelling is useful in identifying and handling a lack of data independence when it is present. In the phonological-complexity data, the model was successful in partitioning the variation that was due to journals nested in disciplines on the one hand, and the variation that was attributable to TextType, on the other. In contrast, when the upper-level grouping effects were absent, as was the case with Neighbourhood Density and Phonotactic Probability data, the mixed model, again, successfully indicated that.

As to the automaticity question, simplification unambiguously manifested itself along the sub-lexical (except for Phonotactic Probability) measures of the study. Unlike many lexical measures of simplification, sub-lexical phonological measures presumably fall beyond the conscious control of translators. Nonetheless, simplification happens along these phonological parameters. This result unveils what must be a genuinely

cognitive component of simplification. Therefore, it seems reasonable to conclude that a cognition-informed account may hold the key to demystifying simplification.

According to De Sutter and Lefer (2020, p. 2), among the "[f]undamental questions" that "remain largely unanswered" is which "cognitive mechanisms shape translation". For scope and space limitations, the current study can only speculate on this unresolved issue. Two hypotheses in the literature seem particularly relevant. These are (1) Rescher's (1989) cognitive economy complemented by the principle of least effort (Zipf, 1949), and (2) Halverson's (2015) cognitive salience. Together, they have good potential to generate useful insights that can account for the simplification effect uncovered here. Perhaps what needs to be investigated in future research that aims to explore the cognitive underpinnings of simplification is the observation that highly frequent, thus salient, words tend to be short and require minimal mental effort to recall and produce.

Another question that remains is whether this effect is translation-specific, i.e., "a product of constraints which are inherent in the translation process itself", in Baker's words (1993, p. 243) or "a universal strategy inherent in the process of language mediation, as practised by language learners, non-professional translators and professional translators alike", as Blum-Kulka (1986, p. 21) puts it. The resemblance that translation bears to other forms of mediated or constrained discourse has been amply highlighted in the literature over the past decades (Bisiada, 2017). Translation has also been likened to non-native language and bilingual communication (Rabinovich et al., 2016).

It would be interesting to see if the paradigm used here would still yield similar conclusions when applied to other forms of mediated or bilingual communication. Clearly, this topic merits further research, as it would (1) improve our understanding of the dynamics of simplification in other forms of text, translated or non-tanslated, and (2) inform our judgement on where to place translation among the various forms of mediated discourse.

Conclusion

The study has taken an interdisciplinary approach to the issue of simplification in translation. Although corpus-based inquiry into simplification dates back to the 1990s, a question mark still hangs on its ontological status. This is partly due to the mixed findings in the literature. The lexical metrics that are traditionally used to quantify simplification are often criticised as lacking robustness and cognitive grounding. The current study has tested for simplification using complexity-based measures adopted from phonology and cognitive psychology. The study supports the simplification hypothesis. The findings reported here imply that simplification is an effect that bears the hallmark of the workings of human cognition. So far, the current study has highlighted a cognitive component of simplification and added a multi-disciplinary twist to the narrative. However, more research is needed to fully understand how and why simplification in translation happens.

References

Allen, R., & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language*, *55*(1), 64-88. https://doi.org/10.1016/j.jml.2006.02.002

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Baddeley, A., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 575-589. https://doi.org/10.1016/S0022-5371(75)80045-4

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. 233-252). 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 Sager* (pp. 175-186). John Benjamins.

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-51. https://doi.org/10.18637/jss.v067.i01

Bazan, A., Kushwaha, R., Winer, E., Snodgrass, J., Brakel, L., & Shevrin, H. (2019). Phonological ambiguity detection outside of consciousness and its defensive avoidance. *Frontiers in Human Neuroscience*, *13(1)*, 1-14. https://frontiersin.org/articles/10.3389/fnhum.2019.00077

Bisiada, M. (2017). *Universals of editing and translation*. Language Science Press. http://hdl.handle.net/10230/33522

Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-Kulka (Eds.), *Interlingual and intercultural communication discourse and cognition in translation and second language acquisition studies* (pp. 17-35). Gunter Narr Verlag Tübingen.

Bradlow, A., & Pisoni, D. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074-2085. https://doi.org/10.1121/1.427952

Braginsky, M., Yurovsky, D., Marchman, V., & Frank, M. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 1-16. https://doi.org/10.1162/opmi_a_00026

Chen, T.-Y., & Sie, Y.-S. (2019). *A reassessment of the effects of neighborhood density and phonotactic probability on L2 English word learning*. *Proceedings of the 11th*

International Conference on Mental Lexicon, 1, 104. https://doi.org/10. 7939/r3-sb0h-jj73

de Groot, A., Borgwaldt, S., Bos, M., & van den Eijnden, E. (2002). Lexical decision and word naming in bilinguals: Language effects and task effects. *Journal of Memory and Language*, *47*(1), 91-124. https://doi.org/10.1006/jmla.2001.2840

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

Dewey, M. (1876). *A classification and subject index for cataloguing and arranging the books and pamphlets of a library.* Lockwood & Brainard Company.

Dirix, N., Cop, U., Drieghe, D., & Duyck, W. (2017). Cross-lingual neighborhood effects in generalized lexical decision and natural reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(6), 887-915. https://doi.org/10.1037/xlm0000352

Edwards, J., Beckman, M. & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in non-word repetition. *Journal of Speech, Language, and Hearing Research*, *47*(2), 421-436. https://doi.org/10.1044/1092-4388(2004/034)

Edwards, J., Munson, B., & Beckman, M. (2011). Lexicon-phonology relationships and dynamics of early language development. *Journal of Child Language*, *38*(1), 35-40. https://doi.org/10.1017/S0305000910000450

Ferraresi, A., Bernardini, S., Petrović, M., & Lefer, M. (2019). Simplified or not simplified? The different guises of mediated English at the European parliament. *Meta*, *63*(3), 717-738. https://doi.org/10.7202/1060170ar

Forker, D. (2021). Complexity and its relation to variation. *Frontiers in Communication*, *6*(1), 1-11. https://doi.org/632468

Freedman, S., & Barlow, J. (2012). Using whole-word production measures to determine the influence of phonotactic probability and neighborhood density on bilingual speech production. *International Journal of Bilingualism*, *16*(4), 369-387. https://doi.org/10.1177/1367006911425815

Garson, G. (2013). Fundamentals of hierarchical linear and multilevel modeling. In G. Garson (Ed.), *Hierarchical linear modeling: Guide and applications* (pp. 3-26). SAGE. https://doi.org/10.4135/9781483384450.n1

Garson, G. (2020). *Multilevel modeling: Applications in STATA, IBM SPSS, SAS, R, & HLM*. SAGE. https://us.sagepub.com/en-us/nam/multilevel-modeling/book260705

Gendler-Shalev, H., Ben-David, A., & Novogrodsky, R. (2021). The effect of phonological complexity on the order in which words are acquired in early childhood. *First Language*, *41*(6), 779-793. https://doi.org/10.1177/01427237211042997

Goldinger, S., & Papesh, M. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science*, *21*, 90-95. https://doi.org/10.1177/0963721412436811

Hall, K., Allen, B., Coates, E., Fry, M., Huang, S., Johnson, K., Lo, R., Mackie, S., Nam, S., & McAuliffe, M. (2022). *Phonological CorpusTools Version 1.5.1*. https://corpustools.readthedocs.io/en/latest/about.html

Halverson, S. (2015). Cognitive translation studies and the merging of empirical paradigms: The case of "literal translation". *Translation Spaces*, *4(2)*, 310-340. https://doi.org/10.1075/ts.4.2.07hal

Hoover, J., Storkel, H., & Hogan, T. (2010). A cross-sectional comparison of the effects of phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Memory and Language*, *63*(1), 100-116. https://doi.org/10.1016/j.jml.2010.02.003

Hu, K. (2016). *Introducing corpus-based translation studies*. Springer. https://doi.org/10.1007/978-3-662-48218-6

Jones, S. (2018). Adult word learning as a function of neighborhood density. *Languages*, *3*(1), 1-13. https://doi.org/10.3390/languages3010005

Kruger, H., & van Rooy, B. (2012). Register and the features of translated language. *Across Languages and Cultures*, *13*(1), 33-65. https://doi.org/10.1556/Acr.13.2012.1.3

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1-26. https://doi.org/10.18637/jss.v082.i13

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science, 7*(1), 18-27. https://doi.org/10.1177/1745691611427305

Laviosa, S. (1998). The corpus-based approach: A new paradigm in translation studies. *Meta: Journal Des Traducteurs*, *43*, 474-479. https://doi.org/10.7202/003424ar

Laviosa, S. (2002). 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-Braithwaite, S. (2001). Universals of translation. In M. Baker (Ed.), *Routledge encyclopedia of translation studies* (pp. 288-291). Routledge.

Lee, S.-Y. (2011). *Parallel activation in bilingual phonological processing* [Doctoral thesis, University of Kansas]. http://hdl.handle.net/1808/8157

Liu, K., Afzaal, M. (2021). Syntactic complexity in translated and non-translated texts: A corpus-based study of simplification. *PLOS ONE 16*(6), e0253454. https://doi.org/10.1371/journal.pone.0253454

Longoni, A., Richardson, J., & Aiello, A. (1993). Articulatory rehearsal and phonological storage in working memory. *Memory & Cognition*, *21*(1), 11-22. https://doi.org/10.3758/BF03211160

Luce, P., & Large, N. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, *16*(5-6), 565-581. https://doi.org/10.1080/01690960143000137

Luce, P., & Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1-36.

Marian, V., & Blumenfeld, H. (2006). Phonological neighbourhood density guides lexical access in native and non-native language production. *Journal of Social and Ecological Boundaries*, *2*, 3-35.

Messer, M., Verhagen, J., Boom, J., Mayo, A., & Leseman, P. (2015). Growth of verbal short-term memory of non-words varying in phonotactic probability: A longitudinal study with monolingual and bilingual children. *Journal of Memory and Language*, *84*(1), 24-36. https://doi.org/10.1016/j.jml.2015.05.001

Mueller, S., Seymour, T., Kieras, D., & Meyer, D. (2003). Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1353-1380. https://doi.org/10.1037/0278-7393.29.6.1353

Mu-Sonic Ltd. (2013). *ToPhonetics*. https://tophonetics.com/

Nair, V., Biedermann, B., & Nickels, L. (2017). Understanding bilingual word learning: The role of phonotactic probability and phonological neighborhood density. *Journal of Speech, Language, and Hearing Research*, *60*(12), 3551-3560. https://doi.org/10.1044/2017_JSLHR-L-15-0376

Olohan, M., & Baker, M. (2000). Reporting that in translated English. Evidence for subconscious processes of explicitation? *Across Languages and Cultures*, *1*(2), 141-158. https://doi.org/10.1556/Acr.1.2000.2.1

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Rabinovich, E., Nisioi, S., Ordan, N., & Wintner, S. (2016). On the similarities between native, non-native and translated texts. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* 1870-1881. https://doi.org/10.18653/v1/P16-1176

Rabinovich, E., & Wintner, S. (2015). Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, *3*(3), 419-432. https://doi.org/10.1162/tacl_a_00148

Rescher, N. (1989). *Cognitive economy: The economic dimension of the theory of knowledge*. University of Pittsburgh Press.

Rescher, N. (1998). *Complexity: A philosophical overview*. Routledge. https://www.routledge.com/Complexity-A-Philosophical-Overview/Rescher /p/book/ 9781138508378

Robin, E. (2017). Translation universals revisited. *FORUM, 15*(1), 51-66. https://doi.org/10.1075/forum.15.1.03rob

Roehr-Brackin, K. (2015). Explicit knowledge about language in L2 learning: A usage-based perspective. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 117-138). John Benjamins. https://doi.org/10.1075/sibil.48.06roe

Rojo López, A. (2015). Translation meets cognitive science: The imprint of translation on cognitive processing. *Multilingua*, *34*(6), 721-746. https://doi.org/10.1515/multi-2014-0066

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology*. *Learning, Memory, and Cognition*, *28*(6), 1019-1034. https://doi.org/10.1037//0278-7393.28.6.1019

Sandrelli, A., & Bendazzoli, C. (2005). Lexical patterns in simultaneous interpreting: A preliminary investigation of EPIC (European Parliament Interpreting Corpus). *Proceedings from the Corpus Linguistics Conference Series 1*. University of Birmingham. https://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*(6), 110–114. https://doi.org/10.2307/3002019

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology*, *5*(1), 1-16.

Stamer, M., & Vitevitch, M. (2012). Phonological similarity influences word learning in adults learning Spanish as a foreign language. *Bilingualism: Language and Cognition*, *15*(3), 490-502. https://doi.org/10.1017/S1366728911000216

Storkel, H. (2009). Developmental differences in the effects of phonological, lexical, and semantic variables on word learning by infants. *Journal of Child Language*, *36*(2), 291-321. https://doi.org/10.1017/S030500090800891X

Storkel, H., Armbrüster, J., & Hogan, T. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175-1192. https://doi.org/10.1044/1092-4388(2006/085)

Tehan, G., & Tolan, G. (2007). Word length effects in long-term memory. *Journal of Memory and Language*, *56*(1), 35-48. https://doi.org/10.1016/j.jml.2006.08.015

The International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Thorn, A., & Frankish, C. (2005). Long-term knowledge effects on serial recall of non-words are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 729-735. https://doi.org/10.1037/0278-7393.31.4.729

Titz, J. (2020). mimosa: A modern graphical user interface for 2-level mixed models. *Journal of Open Source Software*, *5*(49), 1-2. https://doi.org/10.21105/joss.02116

Vitevitch, M. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 735-747. https://doi.org/10.1037/0278-7393.28.4.735

Vitevitch, M., Armbrüster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 514-529. https://doi.org/10.1037/0278-7393.30.2.514

Vitevitch, M., & Luce, P. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–329. https://doi.org/10.1111/1467-9280.00064

Vitevitch, M., Luce, P., Pisoni, D., & Auer, E. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, *68*(1–2), 306-311. https://doi.org/10.1006/brln.1999.2116

Wang, F. (2020). Translation process is a psychological and introspective process, as well as a process of problem solving and decision-making. *Revista Argentina de Clínica Psicológica*, *29*(1), 1413-1424.

Williams, D. (2005). *Recurrent features of translation in Canada: A corpus-based study* [Doctoral thesis, University of Ottawa]. Canada. https://doi.org/10.20381/ruor-12864

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

Zipf, G. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.