

STTR as a measure of lexical diversity: investigating simplification in EPTIC

Alice Fedotova

LDA Project Report

August 15, 2023

1 Introduction

Lexical simplification, one of the translation universals proposed by Baker (1993), has traditionally been operationalized using parameters such as lexical density, the proportion of lexical to grammatical words, and the type-token ratio (TTR), a measure of lexical variety which is obtained from the ratio of unique words to the total words in a text. Initial investigations adopting an intermodal perspective have mainly focused on small-to-medium sized corpora, predominantly using TTR as the principal measure of simplification (Bernardini et al., 2016). However, despite its widespread use, the TTR is not without its drawbacks, particularly when it is used to compare texts of different length (Nasseri, 2021). Several alternative measures have been proposed to address this problem. One of these alternatives is the standardized type-token ratio (STTR), also known as the mean segmental type-token ratio (Brezina, 2018). Following the approach introduced by Bernardini et al. (2016), the aim of this study is to investigate STTR in the English and Italian subcorpora of EPTIC. One of the main advantages of EPTIC is that it allows for multi-directional comparisons, which are crucial to take into account as “simplification parameters seem to apply differently to different languages” (Bernardini et al., 2016). Source texts and mediated language in general are also examined to further contextualize the patterns of simplification across different dimensions.

2 Research questions and hypotheses

The research questions are the following:

Research Question 1: Is there a significant difference between translated texts, interpreted texts, verbatim reports and original speeches in terms of mean STTR?

Research Question 2: If there is a significant difference between the texts, which text types significantly differ from the others in terms of mean STTR?

RQ1 requires one hypothesis to be tested, i.e. that the means are equal for each text type. If the null hypothesis is rejected, four additional hypotheses in the form of intermodal and monolingual comparable comparisons (intermodal; intermodal, control; comparable, interpreting; comparable, translation) are investigated to answer RQ2 as in the approach introduced in Bernardini et al. (2016). The analysis is conducted on both English and Italian texts.

3 Method

3.1 Variables

In all analyses, the dependent variable is lexical diversity, operationalized using the standardized type-token ratio (STTR). STTR is obtained by dividing the text into standard-size segments (e.g., 1000 words) and then calculating the TTR for each segment. Because most texts do not divide exactly into standard-size segments, the last segment which is shorter than the standard size is excluded from the calculations (Brezina, 2018). Following Xu and Li (2022), the STTR is counted on the basis of 1,000 words after removing truncated words, filled pauses and repairs.

The independent variable is the type of text. There are four possible levels of the independent variable: spoken source text ('sp_st'), written source text ('wr_st'), interpreted target text ('sp_tt') and translated target text ('wr_tt'). This categorization is derived from the structure of the corpus, and applies to both directions involving English and Italian. In the data, English texts are preceded by 'en_', while Italian texts are preceded by 'it_'. However, not all levels are considered in all analyses, as some of the tests only concern specific pairs of texts (see Section 3.4.4).

3.2 Samples

This study is based on the European Parliament Translation and Interpreting Corpus (EPTIC)¹, an intermodal, parallel and multilingual corpus comprising different language combinations and mediation directions. Here, both directions of the English-Italian combination, namely the EN > IT and IT > EN subcorpora, were considered. A description of the extracted data is provided in Table 1. All of the available texts were used for the analysis, even though it should be noted that a fewer number of texts were found compared to Bernardini et al. (2016).

3.3 Data

The data was obtained from Sketch Engine². Considering Sketch Engine's limitation of allowing only a maximum of 100 characters per row to be downloaded, the <text/> tag could not be used. Consequently, the CQL tag <s/> was initially used in Concordance to download the individual sentences. A Python script was then implemented to recombine the sentences

¹<https://corpora.dipintra.it/eptic/?section=documentation>

²<https://bellatrix.sslmit.unibo.it/noske/eptic/>

Subcorpus	N. of segm.	Tokens	Types
en_sp_tt	16	16,000	2,556
en_wr_tt	18	18,000	2,994
en_sp_st	20	20,000	3,158
en_wr_st	19	19,000	3,121
Subtotal	73	73,000	11,829
it_sp_tt	17	17,000	3,169
it_wr_tt	18	18,000	3,763
it_sp_st	17	17,000	3,683
it_wr_st	17	17,000	3,678
Subtotal	69	69,000	14,293
Total	142	142,000	26,122

Table 1: Data obtained from the Sketch Engine version of EPTIC.

into their original texts, according to the values in the `text.id` field. As discussed in Section 3.1, the texts were then split into segments of 1,000 words and elements such as truncated words, filled pauses, and repairs were excluded prior to the STTR computation. This was accomplished by leveraging the existing tokenization and tags provided on Sketch Engine and then using a combination of regular expressions to exclude tags related to disfluencies and punctuation (SENT, DYSF, EPAUSE, FPAUSE, UNCLEAR, SYM, PUN). To avoid complications with missing values and other potential issues, considering that not all types of texts share the same number of segments for the STTR, the long format was used to organize the data. The data was then analyzed using R version 4.3.1.³

3.4 Analysis

As RQ1 concerns the relationship between STTR, a continuous dependent variable, and the type of text, a categorical independent variable with four levels, one-way analysis of variance (ANOVA) can be used to compare the mean STTRs if the data meets the normality assumption. Given that EPTIC consists of “independently produced translational and interpretational outputs” (Bernardini et al., 2016), we can consider the samples to be independent and conduct one-way independent ANOVA. Other assumptions of ANOVA include equality of variances and absence of outliers, which will be addressed in the next sections.

3.4.1 Outliers

Outliers that do not change the results but affect the assumptions of a test can be discarded (Lim, 2017). As the presence of outliers affects one of the assumptions of ANOVA, outliers

³<https://cran.r-project.org/bin/windows/base/old/4.3.1/>

	English	Italian
sp_st	$W = 0.976, p = 0.876$	$W = 0.942, p = 0.446$
wr_st	$W = 0.940, p = 0.294$	$W = 0.965, p = 0.724$
sp_tt	$W = 0.952, p = 0.564$	$W = 0.920, p = 0.148$
wr_tt	$W = 0.945, p = 0.378$	$W = 0.970, p = 0.811$

Table 2: Results of the Shapiro-Wilk tests for normality.

were dropped using the boxplot method⁴ as this did not change the overall results. The boxplot method is based on the interquartile range (IQR). Any data point falling below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ was considered to be an outlier and excluded from the subsequent analyses.

3.4.2 Normality

STTRs were then checked for normality using the Shapiro-Wilk test. As illustrated in Table 2, all Shapiro-Wilk tests returned p -values over the alpha level of 0.05. As a result, we can conclude that the STTRs are generally normally distributed. One-way ANOVA can therefore be used to compare the mean STTRs.

3.4.3 Homoscedasticity

Levene’s test for equality of variances was also conducted across the EN texts and the IT texts to verify the assumption of homoscedasticity. The variances of the STTRs are similar for both the EN texts, $F(3, 66) = 2.049, n.s.$, and the IT texts, $F(3, 61) = 2.511, n.s.$

3.4.4 Multiple comparisons

As for RQ2, according to Rafter et al. (2002), the type of correction to be applied when the assumption of equal variances is met depends on the number of comparisons in the family. They suggest the following: “for the family of all pairwise comparisons use the Tukey (or Tukey–Kramer) test. For a family containing some, but not all, of the pairwise comparisons, use one of, in order of preference, the GT2 procedure, the Dunn–Sidak test, or the Bonferroni test”. The Bonferroni method will be used as it is the most straightforward to apply and it is suitable for cases where “a family of selected pairwise comparisons is specified prior to data collection” (Rafter et al., 2002). Furthermore, other sources such as Winter (2019) also recommend conducting only the relevant comparisons, and then applying the correction using the `p.adjust()` function. This adjustment of the p -values is based on the actual number of hypotheses being tested.

⁴www.rdocumentation.org/packages/rstatix/versions/0.7.2/topics/identify_outliers

	English	Italian
sp_st	M = 0.426, SD = 0.025	M = 0.496, SD = 0.010
wr_st	M = 0.437, SD = 0.019	M = 0.499, SD = 0.019
sp_tt	M = 0.410, SD = 0.013	M = 0.459, SD = 0.023
wr_tt	M = 0.444, SD = 0.016	M = 0.497, SD = 0.018

Table 3: Mean STTRs (M) and standard deviations (SD).

3.4.5 Effect sizes

Regarding effect size, as suggested in Field et al. (2012), omega squared (ω^2) is calculated for ANOVA and Cohen’s d for the differences between pairs of groups. The formula for omega squared is the following:

$$\omega^2 = \frac{SS_m - df_M \cdot MS_r}{SS_m + SS_r + MS_r} \quad (1)$$

All of these values can be obtained from the output of `summary(aov())`. There is some disagreement on whether both should be reported or not. Some authors recommend that researchers report on either omnibus test effects or contrast effect sizes, but not both, because omnibus effects already include the effect sizes of the contrasts. Others recommend reporting effect sizes for both omnibus effects and post-hoc contrasts (Larson-Hall, 2015). In this study, both will be reported as not all pairwise comparisons will be conducted after the ANOVA.

4 Results

4.1 Descriptives and graphs

Boxplots showing the distribution of the STTRs across the different categories of texts in the English and Italian subcorpora are provided in Figure 1 and Figure 2. As the data is normally distributed, means and SDs are reported in Table 3.

4.2 Test results

4.2.1 ANOVA

The results of the independent one-way ANOVA with 3 degrees of freedom show similar patterns in the two languages. For the English subcorpora, $F(3, 66) = 9.584$, $p < 0.001$, which indicates a statistically significant difference. A strong effect size was found, with omega squared $\omega^2 = 0.182^5$. For the Italian subcorpora, ANOVA showed that the texts differ significantly in terms of STTR, $F(3, 61) = 18.42$, $p < 0.001$. The effect can be judged as very strong based on omega squared, $\omega^2 = 0.397$.

⁵Reference values for omega squared: $\omega^2 = 0.01$ (small), $\omega^2 = 0.06$ (medium), $\omega^2 = 0.14$ (large) (Larson-Hall, 2015).

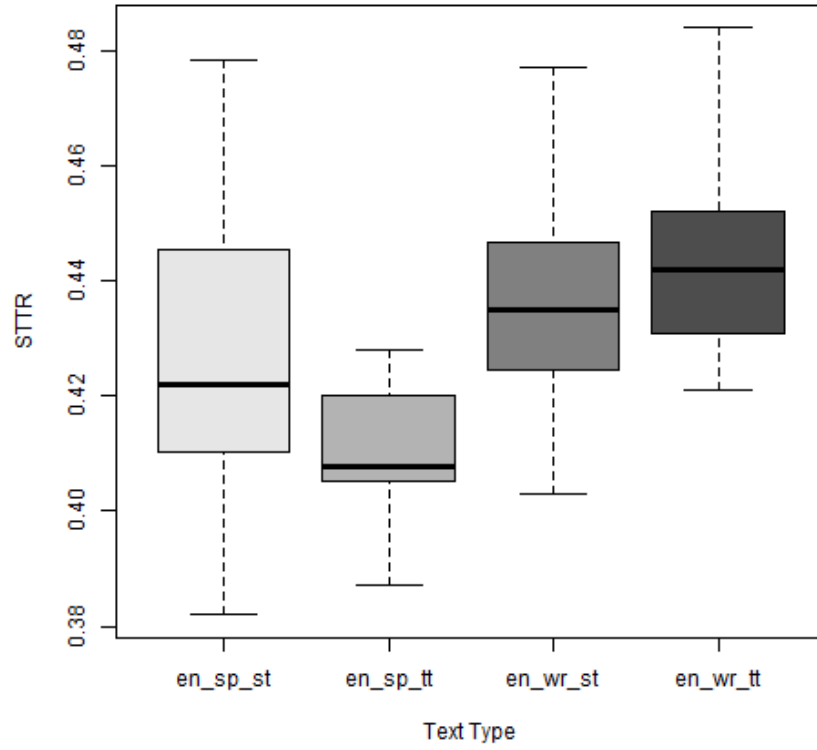


Figure 1: Boxplots of the STTRs in the English subcorpora.

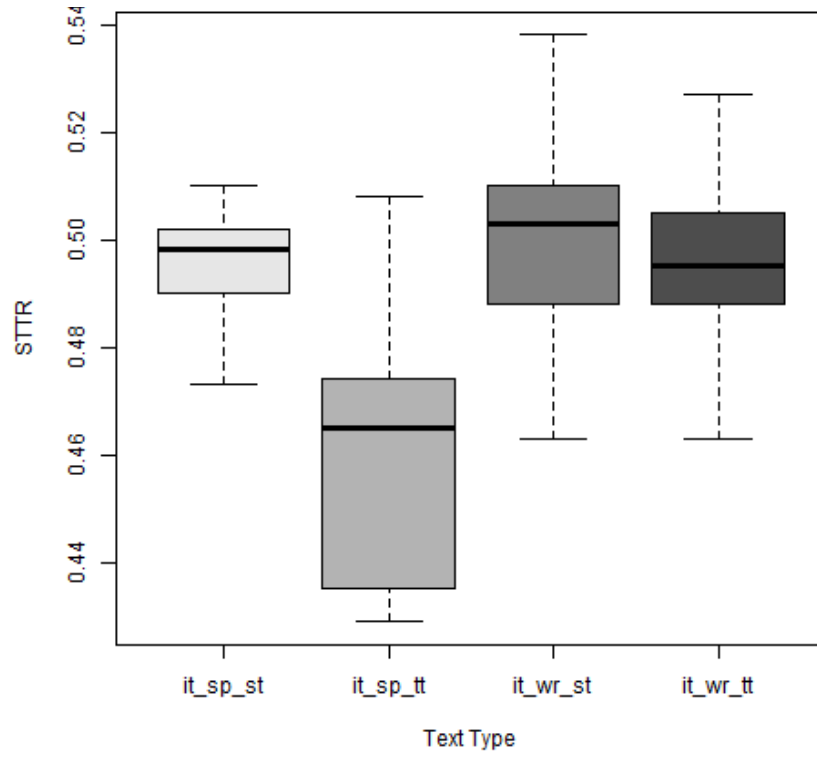


Figure 2: Boxplots of the STTRs in the Italian subcorpora.

	English	Italian
Intermodal (sp_tt vs. wr_tt)	$t(30) = -6.527$, $p < 0.001$, $d = 2.312$	$t(32) = -5.392$, $p < 0.001$, $d = 1.849$
Intermodal, control (sp_st vs. wr_st)	n.s.	n.s.
Comparable, interpreting (sp_tt vs. sp_st)	n.s.	$t(29) = -5.627$, $p < 0.001$, $d = 2.031$
Comparable, translation (wr_tt vs. wr_st)	n.s.	n.s.

Table 4: Results of the pairwise comparisons (t-tests with Bonferroni).

4.2.2 Pairwise comparisons

To identify which groups differ in terms of the mean STTR, pairwise comparisons were conducted using t-tests with Bonferroni correction. Bonferroni tests revealed highly significant differences between translated and interpreted texts in both English, $t(30) = -6.527$, $p < 0.001$, and Italian, $t(32) = -5.382$, $p < 0.001$. The effects can be considered very strong based on Cohen’s $d = 2.312$ for translated and interpreted texts in English, and $d = 1.849$ for translated and interpreted texts in Italian. Additionally, a highly significant difference was found in Italian between the verbatim reports and the transcriptions of the interpretations, $t(29) = -5.627$, $p < 0.001$. Cohen’s $d = 2.031$ indicates a very strong effect size.

5 Conclusions

In this confirmatory study, STTRs were examined as a measure of simplification in the English and Italian subcorpora of EPTIC. As for RQ1, significant ANOVAs provide support for the alternative hypotheses that there are differences in terms of mean STTR across the four text types in both English and Italian. To answer RQ2, pairwise t-tests with Bonferroni correction were conducted on four comparisons of interest that were hypothesized to show a difference in terms of mean STTR. Bonferroni tests revealed highly significant differences in terms of STTR between translated and interpreted texts in both languages, which further substantiates the observation that “interpreted texts in EPTIC are consistently simpler than their translated counterparts” (Bernardini et al., 2016). In examining the monolingual comparable dimension, the data also seems to corroborate the observation that “significant results follow a trend whereby the mediated corpus component is simpler than the corresponding non-mediated one” (Bernardini et al., 2016). The main limitation of this study is however the mismatch in terms of size between the corpora used here and in Bernardini et al. (2016), which means that the results are not directly comparable although they still provide some support for the trends identified in prior research.

References

- 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). Amsterdam, Netherlands: John Benjamins.
- Bernardini, S., Ferraresi, A., & Miličević Petrović, M. (2016). From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies*, 28(1), 61–86.
- Nasseri, M. (2021). *Statistical modelling of lexical and syntactic complexity of academic writing: a genre and corpus-based study of EFL, ESL and English L1 M.A. dissertations*. (PhD diss.). University of Birmingham.
- Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.
- Xu, C., & Li, D. (2022). Exploring genre variation and simplification in interpreted language from comparable and intermodal perspectives. *Babel*, 68(5), 742–770.
- Lim, C. (2017). Re: What is the best way to test for outliers using ANOVA? Retrieved from: https://www.researchgate.net/post/What_is_the_best_way_to_test_for_outliers_using_ANOVA/59ee8ba2ed99e16a240603c0/citation/download
- Grace-Martin, K. (2023). Outliers: To Drop or Not to Drop. *The Analysis Factor*. Retrieved from: <https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/>
- Rafter, J. A., Abell, M. L., & Braselton, J. P. (2002). Multiple comparison methods for means. *Siam Review*, 44(2), 259–278.
- Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.
- Field, Z., Miles, J., & Field, A. (2012). *Discovering statistics using R*. Sage.
- Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R*. Routledge.