Lexical variability as a measure of language complexity in translated and source text

Rogelio González Avilés

KIK405: Final Assignment
rogelio.aviles@helsinki.fi

Abstract

The quantitative analysis of a large corpus of text can be used to obtain relevant information about the diversity of the vocabulary used. By comparing the English original with the French translated versions it is possible to characterize translations as less complex than their originals from a lexical point of view. This is best achieved by an analysis of the lemmatized versions, since lemmatization reduces different occurrences of inflected words to one single form.

1 Introduction

The field of corpus linguistics has provided important tools for the quantification of linguistic phenomena occurring in large collections of text. This quantitative analysis in turn allows for a statistical inquiry into linguistic phenomena. One of this measurable parameters is lexical variability. The analysis of lexical variability in a large collection of text or speech can reveal a significant amount of information about language usage, for example, what level of complexity the author is expected to deliver and how this correlates to the subject dealt with, the context of the text or who the intended reader or listener is. Such a large collection of text is commonly referred to as *corpus*. Jurafsky and Martin (119) define a corpus as "a computer-readable collection of text or speech".

This paper specifically aims at analyzing lexical variation within a bilingual corpus of text. Some observations can be made about the relationship between sample size and lexical variation, or about lexical features that are specific to translated texts. For this purpose this research will make use of the quantitative tools provided by computer-based corpus linguistics. The use of different computer tools will make it possible to observe patterns of distribution related to lexical usage in this corpus. An important research question will be: how are these patterns of distribution affected by focusing the research on different versions of the same text, for example in the original or translated text, or in the raw or lemmatized versions?

This paper will refer briefly to background research on the field of lexical variation through computer-based tools ans to a few important concepts that explain how the research has been carried out.

2 Background

To answer the question of how many words there are in a given text we cannot simply rely on the word processor's word count function, since this will only reveal the amount of individual words present, what is referred to as tokens (Dickinson et al., 76). An alternative way of measuring the amount of words present in a text will be by counting the amount of types, or "number of distinct words" (Jurafsky and Martin., 120). By counting the word types we then ensure that words that are repeated throughout a text will be counted only once; by counting the tokens, instead, a different kind of information is obtained, for example the length of the text.

By using both parameters at the same time, we can see what types or different words are more frequent within a given corpus. More importantly, a relevant measure of the degree of lexical variability in a text is the type-token ratio, a mathematical expression that is calculated as the number of types divided by the number of tokens, and is usually expressed as a percentage (by multiplying by 100). A hypothetical type-token ratio of 1 would mean that the sample has no repeated words at all and that the variation is thorough. The farther this ratio is from one the more repetitive the vocabulary usage is in the sample; a very low type-token ratio will be indicative of scarce lexical variation. As an index this has been used, for example, for determining the complexity of vocabulary a teacher uses with different audiences (see Thomas) or the amount of words children are able to use, with vast amounts

of developmental and diagnostic application possibilities (see Richards). It has also been used to analyze specific features of translated text as opposed to text originally written in a language (see Lapshinova-Koltunski).

Considering that "a single word in a language can show up in different forms" (Dickinson et al., 75) it seems relevant that, in order to obtain accurate results, the counting of types and tokens is carried out on a version that has reduced such different forms to a one single canonical one or *lemma* (ibid).

Since this is a statistical research it is important to consider the effect of the size of the sample used. It is expected that shorter samples will exhibit more variation than larger samples and certainly more than the corpus considered as a whole, since longer texts have fewer chances of introducing new words as the token count increases, thus reducing lexical variability.

3 Data and Method

The corpus that will be analyzed is a bilingual set of Ted Talks texts, first transcribed in English and then translated into French. This set has also another version in each language, which contains only the citation form of the words, or lemmata. This lemmatized version is not without problems, as we will see, but it still delivers a more accurate approximation to the quantification of lexical items used, and therefore serves well the purpose of measuring the type-token ratio of the corpus, even better than the non-lemmatized or raw version.

The freeware corpus analysis toolkit Antconc has been of crucial help in order to establish a type and token count and from this to calculate the appropriate ratios. It was also used to single out lemmatization decisions that might affect those numbers. An empty search of word lists for each of the directories gives the type and token amounts, while a search for different regular expressions proved useful to understand how some lemmatization decisions were made, and how these might affect the type and token count. Several problems were thus detected in the lemmatization of words in the French version. Although they do not significantly affect the total type-token ratio, due to the large size of the sample, I think this is worth mentioning for the sake of accuracy.

An insufficient reduction of the French contractions is the main problem; occurrences of the raw version singular contraction *du* are consistently lem-

matized to two separate words, *de* and *le*, on the one hand. But, on the other, the base form of the plural contraction *des* should also be the two words *de* and *le*, but it is instead lemmatized as simply *du*, which is still a contraction. For example "la plupart des abeilles" (TED-talks; French raw, file 1185) is lemmatized as "la plupart du abeille (TED-talks; French lemmatized, file 1185), instead of the more thorough *la pluspart de le abeille*. This will result in the loss of some occurrences of the lemmatized tokens *de* and *le*, and in the unnecessary addition of the wrongly lemmatized type *du*. A similar problem is observed with the contractions *au* and *aux*: the first one is lemmatized to *à le* and the second one to *au*. Another problem is related to the French articles and how they overlap with some pronominal forms: all definite articles *le, la, les* are systematically lemmatized to the singular masculine form *le*, but when the same tokens are used as pronouns the result of lemmatization is either *le* or *la/le*; the apostrophe form of the singular pronoun *l'* also becomes *la/le* in the lemmatized version. So, for example, "les libérer et les apporter" (TED-talks; French raw, file 685) is lemmatized to "la/le libérer et la/le apporter" (TED talk; French lemmatized, file 685). These results create new tokens that are not present in the raw version, since for AntConc the expression *la/le* consists of two tokens. There are 1610 occurences of *la/le* in the French lemmatized corpus, and therefore 3220 tokens where the raw version has only half that number. Overall, this has a very small effect considering that the total size of the French raw corpus is of 449 688 tokens. It does therefore not noticeably affect this study whose goal is to establish the total type-token ratio of the corpus.

In the English lemmatized version the oblique case forms of the personal pronouns (such as *me, him, her, them*) are not reduced to a single nominative form (*I, he, she, they*). This might be an acceptable approach to the complex problem or lemmatization of pronouns, it creates however a disparity in the type count in the lemmatized version, since *I* and *me* will be different types, but there will only be one type for all cases of *you*.

And lastly, when comparing the different versions I decided to leave out one of the shorter files from the total count, since its text consists almost entirely of a French language rendering of the song "La vie en rose" (TED-talks; English raw, file 115). This had unnecessarily added 36 false English word types to the total count, which were in reality French words. Again, this might not be too noticeable for the goals of this study, since the type-token ratio in this case changed slightly only in the second decimal place of the percentage expression after discarding file 115. Still, I consider it relevant

not to include false types in the type and token amounts. So as to compare equal size corpora, I have also disregarded this file for the French language word count, although this action only eliminates three types.

4 Results

A word list elaborated with the help of the software AntConc gives the following results for the entire corpus.

- English raw version:
  - 18 494 word types
  - 414 827 word tokens
  - Type-token ratio: 4,5%
- English lemmatized version:
  - 14 328 word types
  - 414 826 word tokens
  - Type-token ratio: 3,5%
- French raw version:
  - 23 878 word types
  - 449 582 word types
  - Type-token ratio: 5,3%
- French lemmatized version:
  - 14 620 word types
  - 454 237 word tokens
  - Type-token ratio: 3,2%

In order to compare, I also performed word counts based on one of the shorter samples (TED-talks, file 1382). It turned out to have a total token number of 94 for 63 word types in the English raw version and 88 tokens for 64 types in the French raw version. This gives a type-token ratio of 67% in

English and 72,7% in French. The ratio remained identical in the English lemmatized version of this file, whereas it descended noticeably in the French lemmatized version, to 67,1%. A somewhat larger sample of three files from the English raw version (TED-talks; English raw, files 7, 28 and 39) gives a ratio of 16,7%, a sample of ten files results in a ratio of 13,9% and a sample of 15 files results in a ratio of 10,2%.

5 Conclusion and Discussion

As was expected, sample size is inversely proportional to the type-token ratio, since larger files have less opportunities for repetitions. The ratios were noticeably higher within the smaller samples.

A comparison of the raw version TED-talks in both languages suggests that the French translations, with a type-token ratio of 5,3%, have a slightly more varied use of vocabulary than their English counterparts (ratio of 4,5%). One possible explanation for this lies within specific morphological features of these languages: since verb conjugations, pronouns and articles have many more different forms in French than in English, a higher lexical variation is observable in the French raw version. But this situation is reversed when we consider only the lemmatized versions, where the ratios were 3,5% for English and a much lower 3,2% for French). After reducing different word forms to a single base form or *lemma* we get a different result; it is the original English (lemmatized) text which turns out to have a slightly higher lexical variation rate. This result is in agreement with the research presented by Lapshinkova-Koltunski, according to whom in translated texts "the most common words are repeated more often" in translations than in their originals (p. 96). Consequently, type-token ratio will be lower in translation than in its source text. On the other hand, translated texts "have a relatively low percentage of content words"(ibid) and a higher ratio of functional words, which tend to be repeated; not surprisingly, the most common words in English (lemmatized) was the verb *be* which has many auxiliary uses, followed by the articles, prepositions and pronouns, the other common auxiliary *do* in the 15th place and the first content word (*people*) only in the 36th place; in French the most common words were also functional words, the article *le*, the preposition *de* and the auxiliary *être* in the third place, followed by more prepositions and auxiliaries, with the first possible content word *chose* only in the 40th place.

Another relevant result is that the type-token ratio gives information

about the intended audience of the text. TED is a nonprofit foundation whose goal is to spread ideas, "usually in the form of short, powerful talks" (TED Foundation). Since the subjects can be highly technical and complex, yet they should reach a general public, it is understandable that lexical variation should be restricted and not reach such high levels as in the case of the works of literature that were analyzed for the course, in which we found type-token ratios as high as 10,2% in the case of Barrie's *Peter Pan*. Lexical density is also generally lower in text intended to be heard by a public than in written form.

6 List of tools and methods used

6.1 Tools and software

- AntConc

6.2 Regular expressions

- To count all types and tokens and rank frequency of words: perform an empty search
- To find all cases of *du* in French: \bdu\b
- To find all cases of *des* in French: \bdes\b
- To find all cases of *au* in French: \bau\b
- To find all cases of *aux* in French: \baux\b
- To find all cases of *les* in French: \bles\b
- To find all cases of *la/le* in French: \bla/le\b
- To find all cases of *him, her, them, me* or *it* in English: (\bhim\b)|(\bher\b)|(\bthem\b)|(\bme\b)|(\bit\b)
- To find all cases of *he, she, they, I* or *it* in English: (\bhe\b)|(\bshe\b)|(\bthey\b)|(\bI\b)|(\bit\b)

References

Primary Sources

TED-Talks, retrieved from the course's Moodle page 19 Nov 2019, https://moodle.helsinki.fi/mod/resource/view.php?id=1653096

Secondary Sources

M. Dickinson, C. Brew, and D. Meurers. Language and computers. John Wiley Sons, 2012.

D. Jurafsky and J. Martin. *Speech and language processing, an introduction to natural language processing, computational linguistics, and speech recognition*, Pearson Education International, London 2009.

Lapshinova-Koltunski, Ekaterina. "Variation in translation: evidence from corpora". In Claudio Fantinuoli Federico Zanettin (eds.), *New directions in corpus-based translation studies*, 93–114. Language Science Press, Berlin.

Richards, Brian. "Type/Token Ratios: what do they really tell us?". In *Journal of Child Language*, Volume 14, Issue 02. June 1987, pp 201-209.

TED Foundation. *Our Organization.* TED Foundation, https://www.ted.com/about/our-organization. Accessed 20 Nov 2019

Thomas, Dax, *Type-token Ratios in One Teacher's Classroom Talk: An Investigation of Lexical Complexity*, 2005, https://www.birmingham.ac.uk/Documents/college-artslaw/cels/essays/languageteaching/DaxThomas2005a.pdf. Accessed 19 Nov 2019.