Russ Linguist (2015) 39:309–325
DOI 10.1007/s11185-015-9151-8

Simplification in translated Czech: a new approach to type-token ratio

Упрощение в чешских переводных текстах: новый подход к отношению словоформа / словоупотребление (type-token ratio)

Václav Cvrček[1] · **Lucie Chlumská**[1]

Published online: 15 September 2015

Abstract The main objective of the paper is to examine whether simplification can be demonstrated to exist in Czech translated texts. In general, simplification as one of the so-called translation universals, is defined as a translators' tendency to create simpler texts. According to research of English texts, simplification may be manifested e.g. by a lower level of lexical richness. To describe lexical richness, a simple type-token ratio (TTR) is widely used; however, it is very sensitive to text size. To overcome this disadvantage, a standardized type-token ratio (sTTR) has been introduced, which is calculated for every 1000 words in the text. Nevertheless, it also has certain drawbacks. Our method for standardizing type-token ratio (zTTR) is based on comparing the observed TTR with the referential TTR values representing texts of identical size. Inspired by the z-score, this metric is capable of comparing the lexical richness of texts regardless of their length. The analysis carried out on a large comparable corpus of translated and non-translated Czech proved that the non-translated texts tend be lexically richer, although the difference is not as striking as some studies have predicted.

Аннотация Основной целью работы является выяснение вопроса, содержатся ли упрощения в чешских переводных текстах. В общем случае упрощение, как одна из так называемых универсалий перевода, определяется как тенденция переводчиков порождать более простые тексты по сравнению с оригиналом. Исследования текстов на английском языке, показывают, что упрощения могут проявляться, например, в более низком уровне лексического богатства. Для описания лексического богатства широко используется простое отношение словоформа / словоупотребление (type-token ratio, TTR); однако оно очень чувствительно к размеру текста. Чтобы преодолеть этот недостаток, было введено стандартизированное отношение словоформа / словоупотребление (sTTR), которое вычисляется для каждой тысячи слов

This paper was written within the Programme for the Development of Fields of Study at Charles University, No. P11 'Czech National Corpus', sub-programme 'Czech National Corpus'.

✉ V. Cvrček
vaclav.cvrcek@ff.cuni.cz

L. Chlumská
lucie.chlumska@ff.cuni.cz

[1] Institute of the Czech National Corpus, Charles University in Prague, Prague, Czech Republic

в тексте. Тем не менее, и этот метод имеет определенные недостатки. Наш метод стандартизации отношения словоформа / словоупотребление (zTTR) основан на сравнении наблюдаемой величины TTR со значениями эталонного TTR, представляющими тексты идентичного размера. Эта метрика, родившаяся под влиянием меры z-score, способна сравнивать лексическое богатство текстов безотносительно к их длине. Наш анализ, выполненный на основе большого корпуса чешских переводных и оригинальных текстов, показал, что оригинальные тексты являются, как правило, лексически богаче, хотя разница не столь значительна, как это предсказывали некоторые исследования.

1 Introduction

As the title suggests, this paper has two main objectives. First, it strives to contribute to the study of translated Czech from the quantitative perspective by testing the simplification hypothesis (see Sect. 2.2). In general, simplification is defined as a translators' tendency to create simpler texts, which are easier to understand. According to research conducted on English, simplification may be manifested e.g. by a lower level of lexical richness (Laviosa 1998; Mihăilă 2010).

To describe lexical richness in texts, a simple statistical measure type-token ratio (TTR) is widely used; however, it is very sensitive to text size. Although its adjusted version (standardized TTR or sTTR, see Sect. 4.2) successfully addresses this issue, it introduces a new one related to intratextual variability. This leads us to the second methodological objective of the paper. We propose a new approach to the type-token ratio based on referential values, called zTTR (see Sect. 4.3), which is applicable to texts of differing sizes and respects intratextual variability.

2 Translated language under scrutiny

Empirical research into the language of translation and its characteristic features has been a focus of attention for both linguists and translation scholars for more than twenty years now. With the boom in corpus linguistics and its methods in the 1990s and the subsequent birth of corpus-based translation studies, new research possibilities and questions emerged, such as what language in translations looks like, whether and how it differs from non-translated language, and how it can be studied quantitatively. Ever since, researchers have concentrated more on descriptive than prescriptive studies of translated language, trying to characterize and explain the specifics of this so-called third code.

Translations make up a proportion of published literature in almost every language—the smaller the target language audience, the more significant this proportion is (in Czech, translations represent more than one-third of all published fiction and professional texts). As Baker (1993) notes, given that translations play an important role in shaping both our cultural experience and our knowledge, it seems surprising that they were viewed for so long as 'second-hand texts', as somehow distorted versions of 'real', original texts. They were not regarded as worthy of serious academic enquiry, especially from a linguistic perspective, and if they were studied at all, they were traditionally analyzed as a mere derivative of the original text, not as independent texts (Baker 1993, p. 234). Only with the arrival of large corpus data did translated language become a popular subject of linguistic research aimed primarily at the quantitative analysis of its characteristic features, often referred to as translation universals.

2.1 Translation universals

The term ‘translation universals’ originated in Mona Baker’s (1993) seminal study *Corpus linguistics and translation studies—implications and applications*. This paper laid the foundations for many subsequent studies on the properties of translated language. Baker (ibid., p. 243) defines translation universals as “universal features of translation, that is features which typically occur in translated texts rather than original utterances and which are not the result of interference from specific linguistic systems”. Based on small-scale studies (not corpus-based) as well as theoretical statements, Baker introduces several hypotheses about translated language and later four potential translation universals (Baker 1996, pp. 176–177):

1. simplification (tendency to simplify the language or message or both);
2. explicitation (tendency to spell things out in translation, including the practice of adding background information);
3. normalization or conservatism (tendency to conform to patterns and practices that are typical of the target language, even to the point of exaggerating them);
4. levelling out (tendency of translated text to gravitate towards the centre of any continuum rather than move towards the fringes).

In recent years, these hypotheses have been heavily tested on many languages, not only English, but also Finnish (Tirkkonen-Condit 2004), German (Neumann 2006,[1] 2014[2]), Dutch (Delaere, De Sutter and Plevoet 2012) or Chinese (Xiao 2010). The latest corpus-based studies have mostly disproved the universal status of the suggested translation features to the point that some scholars refuse to call them universals and instead prefer neutral expressions, such as properties or tendencies (Lind 2007; Neumann 2014; even Baker herself in her presentation at the EST Congress 2001[3]). The reason is that there are many factors that influence the features of translated texts, especially genre or text-type differences and the source language effect. The so-called translation universals thus may not be universal in all types of translation or language pairs, but they certainly provide researchers with many inspiring hypotheses and they have provoked new studies and methods for testing translated language. Since Baker’s (1993) study, several new candidates for translation properties were discovered, such as ‘sanitization’ (Kenny 1998), the ‘unique items hypothesis’ (Tirkkonen-Condit 2004) and ‘shining through’ (Teich 2003).

Although there are several small-scale studies and theses on translation universals in Czech (mostly qualitative and based on small data sets, e.g. one original and several translations), none of the aforementioned features have been properly tested on large corpus data from a quantitative perspective. Our motivation was to introduce this type of research into Czech linguistics / corpus-based translation studies both by providing suitable data and methodology and by testing the first of the hypotheses, simplification in translation.

2.2 Simplification

Simplification is usually quite vaguely described as the tendency of translators to simplify the target text in terms of lexical, syntactical or stylistic features. Before we summarize possible

[1] Neumann, S. CroCo: A multiply annotated and aligned corpus for the investigation of translation properties. Invited talk, Language Technology Group Seminars, Macquarie University, Sydney, 15 May 2006.

[2] Neumann, S. Beyond translation properties: the contribution of corpus studies to empirical translation theory. Plenary talk, UCCTS4, Lancaster, UK, 25th July 2014.

[3] Baker, M. Patterns of idiomaticity in translated vs. original English. Paper given at the Third EST Congress Translation Studies: Claims, Changes and Challenges, August 30–Sept. 1, 2001, Copenhagen.

manifestations of simplification in texts, it is necessary to first distinguish between so-called 'S-universals' and 'T-universals', terms coined by Chesterman (2004). The first type, source or S-universal, concerns the differences between translations and their source texts (e.g. in parallel corpora), whereas target or T-universals apply to differences between translations and comparable non-translated texts in the same language (e.g. in a monolingual comparable corpus).

Based on the distinction, simplification may be regarded as S-universal or T-universal, depending on the research focus and data available. We can either examine whether and how translators simplify the language in translated texts compared to their originals, or how the translated language differs from non-translated texts in terms of lexical richness, lexical density, sentence length etc. Given our data—a monolingual comparable corpus of translated and non-translated Czech (see Sect. 3)—we have focused on the T-universal characteristics of simplification.

In her influential research, Laviosa (1998) tried to define core patterns of lexical use in terms of simplification. In her study of translated and non-translated newspaper articles, she came to the conclusion that translated texts have a relatively lower percentage of content words versus grammatical words (i.e. their lexical density is lower) and the most frequent words are repeated more often in translated texts. She further noticed that one of the translational corpora she used (Guardian Translational English Corpus) had a lower type-token ratio (i.e. there was a higher repetition of words in translated texts).

Corpas Pastor, Mitkov, Afzal and Pekar (2008, p. 4) adopted an NLP approach to simplification and among other measures used readability tests. Similarly to Laviosa, they also expected "translated corpora [. . .] to be characterized by less varied and more familiar vocabulary, [. . .] to contain shorter sentences than sentences of original text". Mihăilă (2010, p. 3) summarized the findings of other scholars and presented similar hypotheses: "[T]he translated texts contain a lower level of lexical richness and density."

Despite the fact that simplification, as a general tendency to simplify the language and / or message, can be operationalized in several ways, the underlying hypothesis is usually based on the reduction of lexical variability in translated texts. In our study we focused on lexical richness, namely on the type-token ratio measure (see Sect. 4). In order to analyze simplification as a T-universal, it is necessary to have a comparable corpus of translated and non-translated texts. We used the Jerome corpus as described in detail in Sect. 3.

3 Data—the Jerome corpus

The Jerome corpus (see Chlumská 2013) is a monolingual comparable corpus (according to the corpus typology by Laviosa 2002, p. 36 or Fernandes 2006, p. 91). It was compiled[4] at the Institute of the Czech National Corpus and made available to the public[5] at the end of 2013. It consists of a translational corpus of Czech translations from various languages and a non-translational corpus of Czech originals. It is a synchronic corpus containing texts published in 1992–2009. The corpus is lemmatized, morphologically tagged and annotated in terms of standard text information (author, name, date and place of publication, date of first edition etc.) as well as translation-related information (translator's name and gender, source language). First, we describe the compilation criteria and then summarize its final design, including size, number of texts, authors etc.

[4] As part of grant VG027 2013 FA CU, see Chlumská (2013).

[5] The corpus can be accessed via the KonText interface: http://www.korpus.cz.

3.1 Criteria for compilation

Although most comparable corpora used in corpus-based translation studies do not exceed several million tokens, our objective was to create a large corpus especially suitable for quantitative research, i.e. to include as many texts as possible without violating the desired representativeness. This task proved to be almost impossible; it was necessary to make a compromise (see Zanettin 2011, p. 20), and pragmatically sort the objectives according to their importance, so as to meet the crucial criteria.

With a large size (see Table 1 in Sect. 3.2) being the most desirable feature, all texts from the Czech National Corpus (CNC) database published within the required period were included in the Jerome corpus, provided that:

- they were complete texts (no partial texts or volumes);
- the same author did not have more than three publications in the corpus;
- the same translator did not have more than three translations in the corpus (each one must be by a different author).

Another important objective was to include more than one text type:[6] both fiction and professional literature. Further divisions of fiction (such as novels, short stories, poems etc.) were not taken into account; however, they are included in the text annotation to enable the users to create their own subcorpora. The CNC texts from the professional domain are further divided into a wide range of genres or disciplines, such as law, medicine, history, music, chemistry etc. and can also be filtered accordingly.

It is crucial for a translational corpus to be balanced in terms of the source languages of translations. However, in Czech, as in many smaller or medium-size languages, translations from English are three times more common than from any other language. To include the same amount of texts from all available languages would considerably affect the desired corpus size and also would ignore the real situation of translations in Czech, so a pragmatic approach was adopted. The Jerome corpus as a whole therefore reflects the reality of Czech translated literature in the given period;[7] English is by far the most prevalent language. However, to make up for the possible interference effect, a balanced subcorpus was created within the Jerome corpus. This subcorpus of 5 million tokens includes an equal amount of texts translated from 14 different languages in fiction and 6 in professional literature. It can be later used to validate the findings in terms of their universality across source languages.

3.2 Corpus design

After filtering all the texts according to the above mentioned criteria, the final size of the corpus is approximately 85 million tokens (incl. punctuation), i.e. approximately 69 million text words (see Table 1). Included in the corpus are a total of 1,526 texts written by 1,244 authors (or teams of authors) and translated by 607 translators (or teams of translators). These relatively high numbers should guarantee a sufficient heterogeneity of the corpus, preventing the risk of significant interference from the author's or translator's idiolect.

Both parts of the corpus, translational and non-translational, are comparable as to the text type, date of publication and total size. However, the number of texts is slightly different, as

[6] However important, the issue of text types / genres and their definition far exceeds the limited scope of this paper. In this case, the traditional division available in the CNC was used.

[7] According to the Czech National Library statistics of translated books, available (in Czech) at http://text.nkp.cz/sluzby/sluzby-pro/sluzby-pro-vydavatele/vykazy.

Table 1 Size of the Jerome corpus

Jerome corpus	Tokens incl. punctuation (TRA / non-TRA)	Texts (TRA / non-TRA)
Total	85 065 312	1 526
Fiction	26 551 540 / 26 617 523	394 / 444
Professional	15 949 930 / 15 946 319	382 / 304

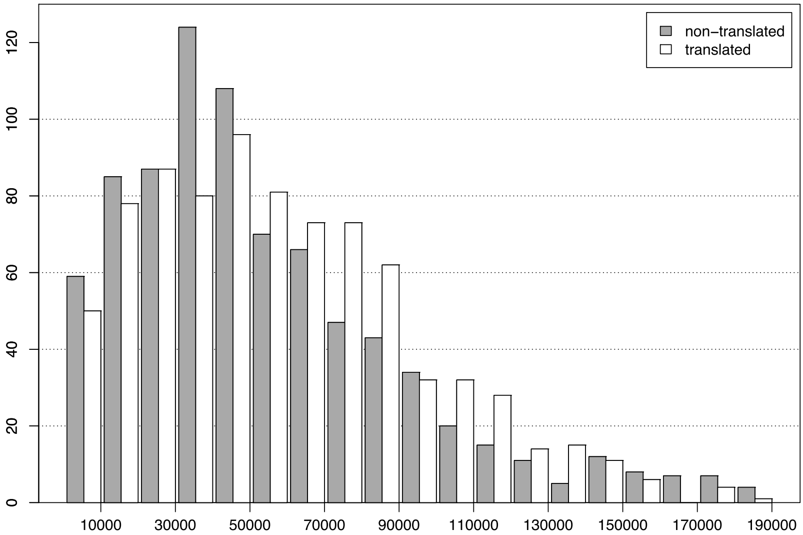

Fig. 1 Distribution of texts according to their size in both parts of the Jerome corpus (the x-axis indicates the size in tokens, while the y-axis shows the number of texts)

they are of different lengths (see Table 1). Despite all efforts to choose similar texts, this issue is practically inevitable when using full texts as opposed to samples. No matter how well-balanced a comparable corpus is, it rarely comprises texts with identical size distribution in both parts (translation and non-translation). This might complicate the calculation of certain statistical tests that are sensitive to text size. This is one of the reasons why we came up with a different approach to one of the measures, namely the type-token ratio (see Fig. 1).

Despite the fact that there is an obvious correlation between the sizes of translated and non-translated texts ($r = 0.9387$), it is obvious that for some sizes these populations significantly differ (30,000–39,000 with the dominance of non-translated texts or 70,000–79,000 with the dominance of translated texts). The relatively high coefficient of correlation can thus be ascribed to the fact that all Czech texts follow a similar distribution of sizes (regardless of their origin), see Fig. 2.

4 Methodology

Type-token ratio (TTR) is one of the most popular measures to quantitatively describe the lexical richness of a given text. It has both advantages and disadvantages; the latter were to be addressed by adjustments found in the standardized TTR (or sTTR, see Sect. 4.2). In this

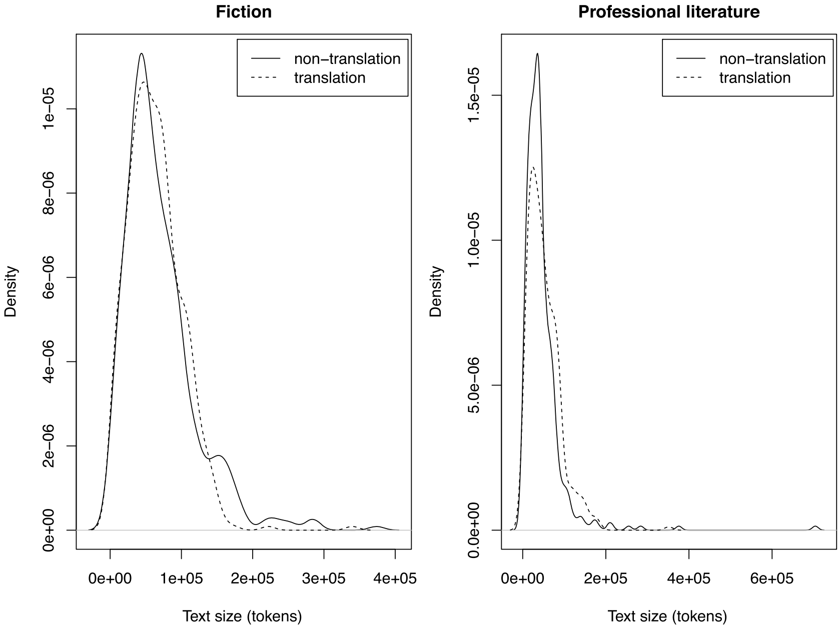

Fig. 2 Kernel density plots of text sizes in fiction and professional literature

section we argue that even the improved sTTR measure cannot overcome certain issues of text size and we suggest a different approach based on referential values.

4.1 Lexical richness and TTR

Despite its obvious drawbacks, TTR is probably the most widely used technique to examine and compare the lexical richness of two or more texts or corpora. Its most appealing advantage is the ease of calculation; almost every text processing tool provides information about the number of types (i.e. all different words)[8] and the number of tokens (i.e. all running words in a text), and the TTR measure is obtained simply by dividing these two numbers.[9]

The greatest disadvantage (disqualifying TTR from many applications) is the fact that TTR is very sensitive to the size of a text or corpus (see Fig. 3). Accounting for the limitation of the vocabulary of any natural language, the number of tokens and types will increase the longer the text becomes, but their increase is asynchronous. When the text reaches a certain length, the increase in new types slows, and the ratio between type and token cannot represent the variability of the use of words (Yang and Wei 2002).

Given that TTR is sensitive to text size, it cannot be used for the comparison of texts of unequal sizes. The larger the text or corpus is, the lower the value of TTR will be. As a corollary, TTR is not an index of lexical richness; it should be treated as a simple function of text

[8] To avoid possible misunderstanding related to the ambiguity of the term: we use the term ‘type’ in this study to denote (different) case-sensitive word-forms (not lemmas). Nevertheless, the algorithm described below will be valid with any kind of types (lemmas, case-insensitive forms etc.).

[9] The simplicity of TTR calculation is not the only reason for its popularity among researchers. Further obvious advantages are its straightforward interpretation and low computational complexity; due to these factors, other metrics, such as Yule’s K (Yule 1944) or Zipf’s Z (Orlov 1982), are used significantly less often.

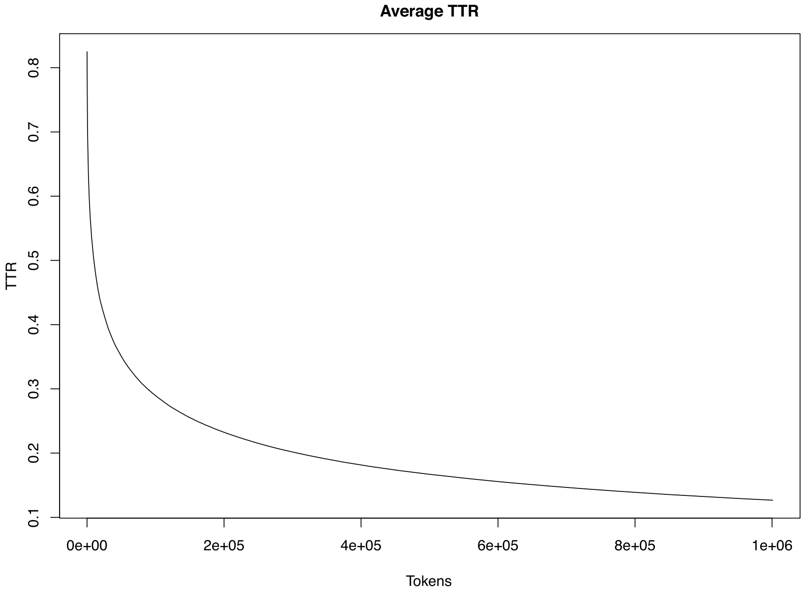

Fig. 3 The effect of text / corpus size in TTR. TTR has been measured for the journalistic part of the corpus SYN2010 (a 100m representative corpus of contemporary written Czech)

size (with more or less precisely predictable results). Moreover, our preliminary experiments have shown that TTR is also influenced by the type of text (e.g. the average TTR for journalistic texts of a given size may differ significantly from the average TTR in fiction texts of the same length); unless we are comparing two texts of the same text type, we cannot rule out this influence.

4.2 Standardized TTR (sTTR)

To overcome the above-mentioned flaws of TTR, another version of this measure was devised—the standardized TTR (sTTR). It is a corrected measure coined by Scott[10] to compensate for text size. sTTR is not based on the total token and type counts in the whole text; instead, it is equal to an average TTR of consecutive chunks of n words (usually 1,000) in the text. By relativizing the TTR value to the same arbitrary level, we obtain sTTR values, which are comparable regardless of the respective text sizes.[11] The sTTR value can thus be interpreted as the expected proportion of types to tokens in a text of exactly n words.

sTTR thus effectively solves the issue surrounding the text size sensitivity of TTR while at the same time introducing other problematic features. Chunking a text and averaging the TTR values does not account for intratextual variability. The sTTR is based on the assumption that chunks are equal or similar with respect to their word frequency distributions (which is usually not true). For example, each chunk of 1,000 words usually contains grammatical words (such as prepositions and conjunctions) but may not contain some rare content words,

[10] WordSmith Tools, version 4 by Mike Scott. More information available at http://www.lexically.net/wordsmith/.

[11] A similar algorithm is used for comparing frequencies of language phenomena in two unequally sized corpora by converting raw frequencies to ipm (instances per million).

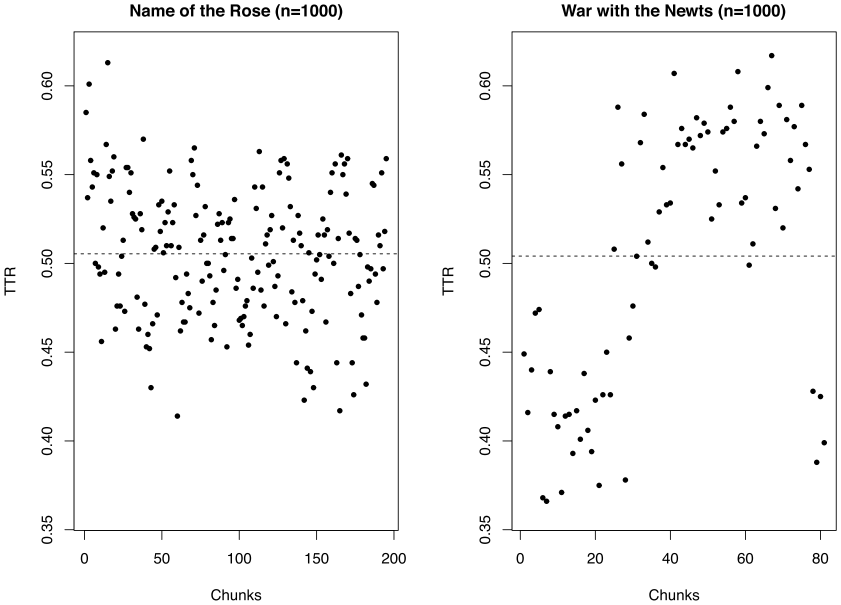

Fig. 4 TTR values per chunk (1,000 tokens) in novels by Eco and Čapek (with the line representing sTTR)

which play a more important role in determining the size of a lexicon (and consequently the text's lexical richness). As a consequence common (grammatical) words are overrepresented with sTTR, whereas content words are underrepresented.

Another important issue presented by sTTR is the function used for its calculation. The arithmetic mean used for sTTR accounts for all TTR values in a set of equally sized chunks into which a text has been split. Therefore all values (however extreme or outlying) contribute to the result. Two texts may therefore have identical sTTR values but the dispersion of chunk TTR may differ to a great extent. As an example of intratextual variability we have used two texts (see Fig. 4): the Czech translation of Umberto Eco's *The Name of the Rose* (*Jméno růže*) and Karel Čapek's novel *Válka s mloky* (*War with the Newts*).

In Fig. 4, single dots represent the TTR value for each chunk (with the size of 1,000 tokens) and the dotted line represents an average TTR value in each text (i.e. sTTR). The TTR values within Eco's novel are distributed randomly around the average, whereas Čapek's text seems to follow a pattern in which the beginning and the end of the novel tend to have lower values of TTR in comparison to its middle parts. This obvious difference in dispersion is visualized using boxplots in Fig. 5 below.

Both texts have almost identical means ($\mathrm{sTTR}_{\mathrm{Eco}} = 0.5054$ and $\mathrm{sTTR}_{\mathrm{Čapek}} = 0.5041$). According to the sTTR they will therefore be evaluated as equally (or similarly) lexically rich. What is neglected in the case of the sTTR is their difference in the dispersion of the TTR within chunks. Eco's novel seems to have a fairly even distribution of types among the whole text and a majority of chunks seem to exhibit more or less the same TTR level (the coefficient of variation of the chunk TTR is 7.3 %). Čapek's novel, on the other hand, is sometimes referred to as a 'novel-feuilleton' as there is no main character and even the narrator's style changes throughout the book several times (ranging from journalistic style to classical fiction). These factors cause the higher coefficient of variation of the chunk TTR in Čapek's novel, i.e. 14.8 %.

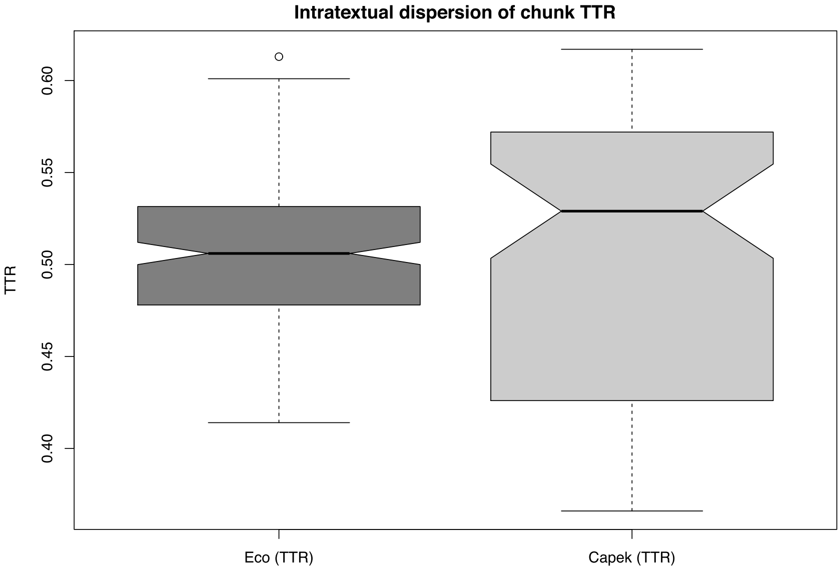

Fig. 5 Intratextual variability of two novels: Umberto Eco's *Jméno růže* and Karel Čapek's *Válka s mloky*

What the sTTR omits is the intratextual variability, which is caused by the internal informational dynamics of a text. These dynamics are derived from the pace by which new topics, characters and themes (and words related to them) are introduced in the course of the text. If a text shows uneven dynamics, it is more likely to have a high level of TTR dispersion (and consequently a more or less unreliable sTTR value).

Moreover, intratextual variability (represented by the chunk TTR dispersion) is not the only factor disqualifying sTTR as a reliable estimator of lexical richness. Another issue is the method of splitting the examined text into parts. Averaging the TTR values for consecutive chunks cannot reveal how similar the lexicon used among them is. Consider the following hypothetical situation with two texts of identical size and sTTR (words are replaced by letters and texts have been split into chunks of three words).

Text A: a b c | a b c | a b c | a b c | a b c | a b c | a b c | a b c
Text B: a b c | d e f | g h i | j k l | m n o | p q r | s t u | v w x

Within both texts, the TTR values of chunks oscillate closely around the average (3 types per 3 tokens); however, the chunks in text A are identical with respect to the inventory of types used (e.g. very repetitive text using the same set of words). Text B, on the other hand, while also having 3 types per 3 tokens, consists of text chunks with different types (e.g. a collection of unrelated short stories introducing new lexical items). Contrary to the intuitive assumption that text B is inevitably lexically richer, sTTR values will be the same for both texts (sTTR = 3). This paradox is caused by confusing the lexical richness of the whole text with the average value of its parts. Having summed up all parts of text A, we would gain no more types in addition to the inventory already used in the first chunk (i.e. 3 for 24 tokens), whereas the lexicon of text B cumulatively grows with each added part (up to 24 types per 24 tokens).

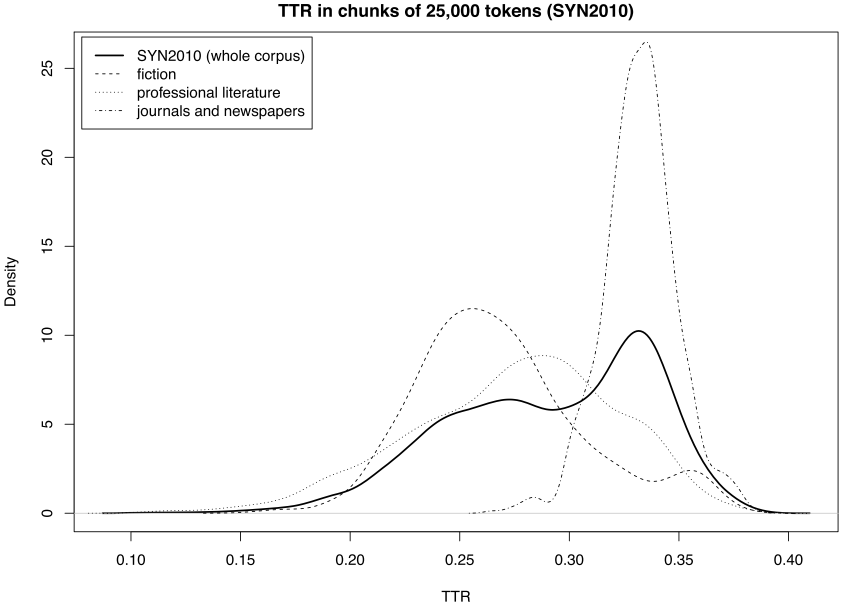

Fig. 6 TTR values in the SYN2010 corpus and in different text types: fiction, professional literature and newspapers

4.3 TTR scaling (zTTR)

The aforementioned disadvantages of TTR and sTTR motivated us to come up with a new measure which we called zTTR. It is based on the comparison of the actual TTR with referential values. In order to compensate for different text sizes, these referential values have to be of two types: the average TTR of a population of texts of a given size; and the standard deviation of the TTR within the same population of texts. By positioning the TTR of a text relative to the TTR distribution in a large sample of texts, we can estimate how extreme (or how average) its value is.

An ideal solution for obtaining referential values would be to have a large and representative sample of texts for every possible text size. This sample would then be used to calculate sample mean and sample standard deviation. The sample mean (as an expected value) could then be used for comparison with the actual value of the TTR of a text; we would be able to assess how (un)expected the TTR value is in comparison with the usual value and usual dispersion (standard deviation) within texts of the same length.

This is, obviously, an impossible task to accomplish, as even the largest available corpora do not include a sufficient amount of texts for any given size. We had to adopt an alternative method, in which we approximate the population of texts of a given size by splitting the whole referential corpus into consecutive chunks of the required size.

However, initial experiments in this field showed that the situation is further complicated by the fact that referential values are influenced not only by text length, but also by text type (or genre). This can be observed in a density graph of TTR values calculated using chunks of 25,000 tokens (see Fig. 6).

The solid line representing the entire SYN2010 corpus shows a tendency towards multimodal distribution. The reason for this lies in the composition of the corpus, consisting as

Table 2 Sample chart for different text lengths for calculating zTTR

Tokens	Fiction		Professional literature	
	Average TTR	Standard deviation (s)	Average TTR	Standard deviation (s)
500	0.5933	0.05566	0.6156	0.05930
600	0.5778	0.05495	0.5990	0.05925
700	0.5647	0.05441	0.5852	0.05930
…	…	…	…	…
199,000	0.1657	0.02667	0.1672	0.02817
199,500	0.1655	0.02680	0.1671	0.02816
200,000	0.1650	0.02684	0.1670	0.02846

it does of three separate populations: fiction, professional literature and journalistic texts. These three populations differ in their mean TTR as well as in their modes and dispersion.

Given that the referential values vary significantly in different text types, we decided to calculate referential values separately for fiction and for professional literature (for the purpose of this study we did not need referential values for journalistic texts as they are not included in the Jerome corpus). For each text size and for each text type (fiction and professional literature) in the representative corpora SYN2000, SYN2005 and SYN2010[12] we have obtained average TTR values and the standard deviation (s). The sampling frequency was 100 tokens for the smaller sizes and 500 tokens for the larger sizes (see Table 2).

The zTTR is calculated as a comparison of the TTR of the examined text with referential values on the basis of the following formula:

$$\mathrm{zTTR} = (\mathrm{TTR} - \text{Average TTR})/s$$

Although this measurement was obviously inspired by the z-score[13] (hence the name: zTTR), it should be emphasized that zTTR calculation is not a normalization *per se*. As the underlying data for referential values do not have a normal distribution, zTTR cannot be interpreted as the z-score (as values corresponding to percentiles of a population, e.g. $z \leq -2$ refers to 2.3 % cases). Nevertheless, zTTR yields comparable results for texts of unequal size as it represents the distance between the raw TTR of a text under examination and the mean TTR (of texts of the same length) in the number of standard deviations.

The interpretation of zTTR values is limited to the comparison of texts (values are not directly comparable to the raw TTR or sTTR). However, we may use the zTTR for comparisons with referential (i.e. expected) values:

$$\mathrm{zTTR} = 0 \ldots \text{ average value}$$
$$\mathrm{zTTR} < 0 \ldots \text{ below average (lexically less rich)}$$
$$\mathrm{zTTR} > 0 \ldots \text{ above average (lexically more rich)}$$

To give a further example of the zTTR calculation, let us imagine a fictional text with 180,357 tokens and 32,995 types. The TTR of the text is 32,995 / 180,357 = 0.1829. In the table of

[12] For the purpose of this study we have excluded texts which were previously included in the Jerome corpus.

[13] A similar approach to normalizing the difference between an actual value and a sample mean using the standard deviation was adopted e.g. for measuring lexical fixedness (Fazly and Stevenson 2006).

Table 3 Table of referential values for the zTTR

Tokens	Average TTR	Standard deviation (s)
180,000	0.1705	0.02741
180,500	0.1702	0.02773

Table 4 Values of TTR, sTTR and zTTR for Eco's *Jméno růže* and Čapek's *Válka s mloky*

Text	Tokens	Types	TTR	sTTR	zTTR
Eco	195,679	28,976	0.1481	0.5054	−0.7011
Čapek	81,758	18,394	0.225	0.5041	0.3523

referential values for fictional texts (see Table 3) we find values closest to our text with respect to its size.

Having interpolated these values, we obtain referential values for the exact size of a text: average TTR = 0.1703 and s = 0.02764. With these figures we can calculate zTTR = $(0.1829 - 0.1703)/0.02764 = 0.45586$. This number can be used for comparison with other texts (regardless of their length and text type).

To demonstrate the differences between the approaches, let us look back at the above-mentioned novels by Eco and Čapek. Table 4 clearly shows the different values of the TTR, sTTR and zTTR. While the sTTR for both novels is almost identical, suggesting a comparable lexical richness, the zTTR indicates differences due to intratextual variability and dynamics; Eco's novel is under average, while Čapek's text is above average.

To sum up, zTTR combines the advantages of TTR and sTTR: it is not text-size influenced, and it respects intratextual variability and information dynamics (as it does not require the splitting of examined texts into chunks and treats them as a whole). Moreover, it allows for the comparison of texts between text types. The zTTR also has obvious disadvantages, with the most important of them being the demanding process of obtaining referential values (which are language-specific due to typological differences).

5 Results

We calculated both the sTTR and zTTR for the comparable parts of the Jerome corpus, non-translated and translated Czech, separately for each text type, fiction and professional literature. To verify the statistical significance of the differences, we used a Mann-Whitney U test with an alternative hypothesis that the true location shift is greater than 0 (Wilcoxon rank sum test with continuity correction in R[14]). Table 5 shows that the sTTR does not indicate a statistically significant difference between translations and non-translations in professional literature, whereas the zTTR does.

Figures 7 and 8 compare the observed values for the TTR, sTTR and zTTR using boxplots. In fiction, the translated part of the Jerome corpus has a lower type-token ratio suggesting that translations do behave differently in terms of their lexis and repetition, whereas in professional literature, this tendency of the translated texts is indicated only by TTR and zTTR.

[14] R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Table 5 Values for the sTTR and zTTR tested for statistical significance

1st group	2nd group	sTTR		zTTR	
		Wilcox (U-test)	*p*-value	Wilcox (U-test)	*p*-value
Non-translated	Translated	335,859.5	5.38e–8	355,812	1.12e–14
Non-translated fiction	Translated fiction	111,974.5	1.22e–12	115,993	2.20e–16
Non-translated professional	Translated professional	56,876.5	0.6775	67,333	0.0001624

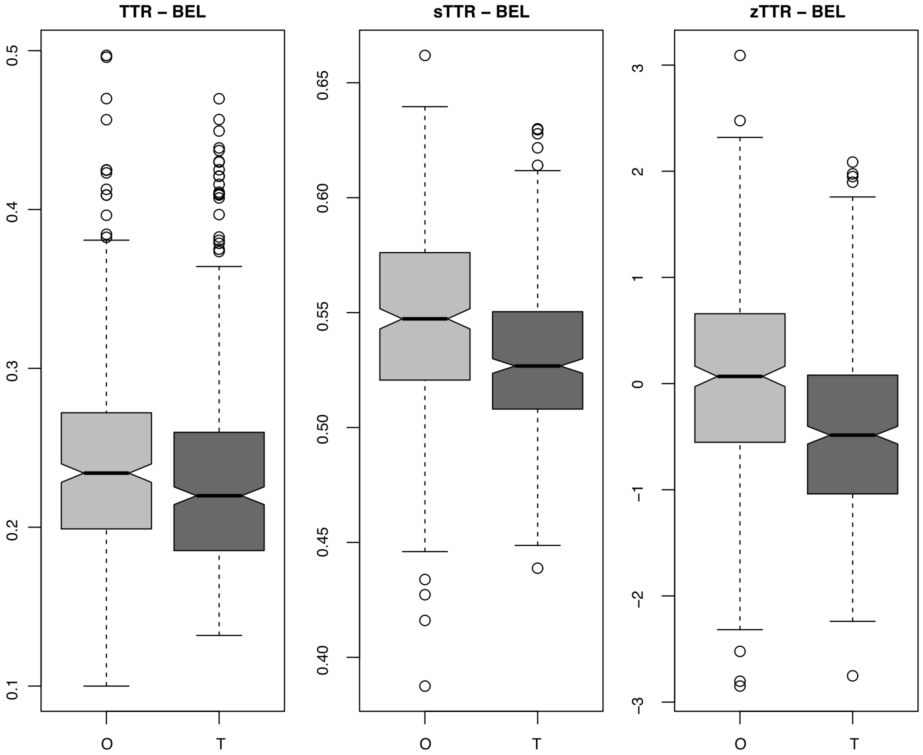

Fig. 7 Comparison of TTR, sTTR and zTTR in fiction (O = original / non-translated Czech, T = translated Czech)

Is there an explanation for similar / different results based on TTR, sTTR and zTTR? As we have demonstrated with Eco's and Čapek's texts, the results based on the TTR can be valid only if the requirement of similar text size distribution in compared texts / parts of the corpus is met (see Fig. 2), whereas sTTR leads to similar conclusions as zTTR in those cases when the informational dynamics (i.e. distribution of types) in texts / parts of a corpus are comparable (which is the case for fiction in the Jerome corpus but not for professional literature, see Figs. 7 and 8).

We have also calculated the effect size for zTTR differences between translated and non-translated texts with the rank-biserial correlation (Wendt formula which is based on the Mann-Whitney U test). The results—oscillating from small to medium effect size (see Ta-

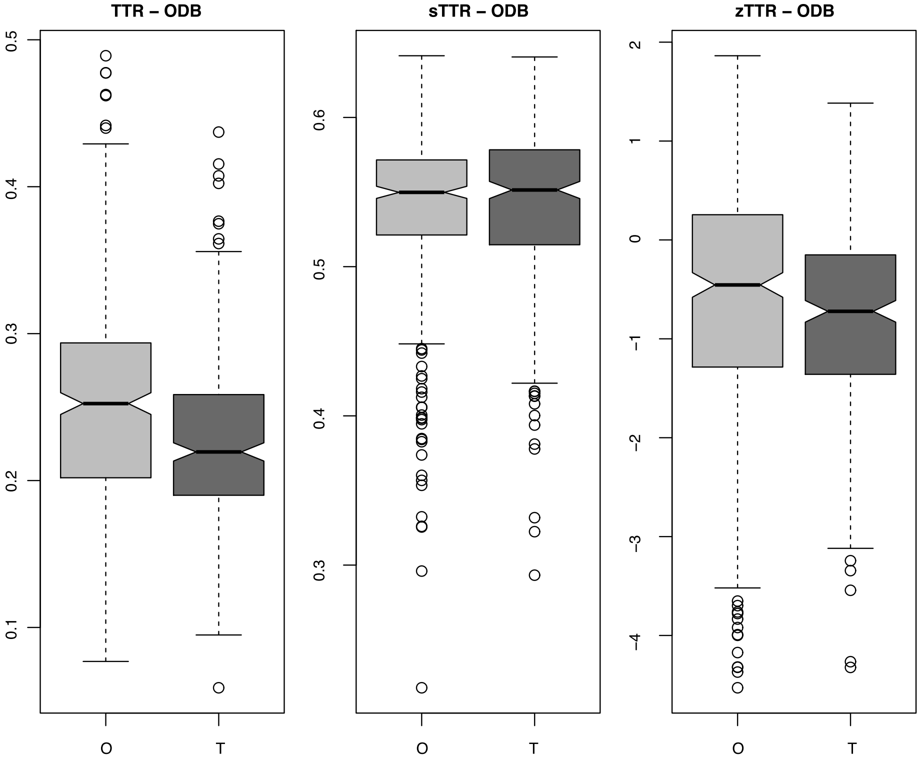

Fig. 8 Comparison of TTR, sTTR and zTTR in professional texts (O = original / non-translated Czech, T = translated Czech)

Table 6 Effect size of the difference in zTTR between translated and non-translated texts according to the Wendt formula for rank-biserial correlation

1st group	2nd group	Rank-biserial correlation (r)
Non-translated	Translated	0.226
Non-translated fiction	Translated fiction	0.326
Non-translated professional	Translated professional	0.16

ble 6)—proved that the non-translated texts tend to have higher zTTR values (and therefore are lexically richer), although the difference is not as striking as some studies have predicted.

However, the type-token ratio measure, no matter how precisely and accurately calculated, serves merely as initial information for further research. The statistical difference between non-translated and translated Czech suggests that there is certainly a potential for additional linguistic analyses, both quantitative (for a general survey of translated language) and qualitative (for concrete case studies based on selected linguistic features).

6 Conclusion

Two separate conclusions can be drawn from the present study. The first is methodological and is related to the algorithm used for calculating lexical richness. Despite the fact that TTR is the most widely used measure of lexical richness, its obvious flaws disqualify it from any serious use (with the most problematic issue being the fact that it is sensitive to text size). The improved version, sTTR, solves this issue but at the same time introduces another one, which is that it ignores intratextual variability and dynamics. We have therefore suggested a further alternative method, zTTR, which takes referential values (drawn from a large reference corpus) into account reflecting not only the size of a text, but also its text type. Confronting the TTR of a text with referential values allows us not only to compare the actual TTR with expected values (for texts of similar length and text type), but also to compare unequally-sized texts with regards to their lexical richness.

The second conclusion is related to the question of simplification as a universal feature of translated versus non-translated texts. With respect to the results presented above, we may conclude that the difference is significant, but the effect size would generally be considered small or medium. This means that with Czech texts (both fiction and professional literature) translated from other languages we might expect a tendency to employ a slightly less diverse lexicon in comparison to non-translated Czech texts. The question as to the extent to which this tendency can be observed in other languages as well (and consequently as to whether we can call it a 'translation universal') remains unanswered and is subject to further research.

Sources

Eco, U. (1988). *Jméno růže* (translation Z. Frýbort). Praha.
Čapek, K. (1981). *Válka s mloky* (1st ed. 1936). Praha.

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, Philadelphia.

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* (Benjamins Translation Library, *18*, pp. 175–186). Amsterdam, Philadelphia.

Chesterman, A. (2004). Hypotheses about translation universals. In G. Hansen, K. Malmkjær, & D. Gile (Eds.), *Claims, changes and challenges in translation studies. Selected contributions from the EST Congress Copenhagen 2001* (Benjamins Translation Library, *50*, EST Subseries, *1*, pp. 1–14). Amsterdam, Philadelphia.

Chlumská, L. (2013). Jerome – a monolingual comparable corpus of translated and non-translated Czech. Available at http://www.korpus.cz.

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 Conference of the Association for Machine Translation in the Americas (AMTA-08)*. Waikiki, Honolulu. Retrieved from http://clg.wlv.ac.uk/papers/AMTA2008.pdf (18 June 2015).

Delaere, I., De Sutter, G., & Plevoet, K. (2012). Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target*, *24*(2), 203–224. doi:10.1075/target.24.2.01del.

Fazly, A., & Stevenson, S. (2006). Automatically constructing a lexicon of verb phrase idiomatic combinations. In *EACL-2006. Proceedings of the 11th Conference of the European chapter of the Accociation for Computational Linguistics. April 3rd–7th, 2006. Trento, Italy*. Retrieved from http://www.aclweb.org/anthology/E06-1043 (18 June 2015).

Fernandes, L. (2006). Corpora in translation studies: revisiting Baker's typology. *Fragmentos*, *30*, 87–95.

Kenny, D. (1998). Creatures of habit? What translators usually do with words. *Meta: Translators' Journal*, *43*(4), 515–523.

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

Laviosa, S. (2002). *Corpus-based translation studies. Theory, findings, applications* (Approaches to Translation Studies, *17*). Amsterdam, New York.

Lind, S. (2007). Translation universals (or laws, or tendencies, or probabilities, or . . .?). *TIC Talk. Newsletter of the United Bible Societies Translation Information Clearinghouse*, *63*. Retrieved from https://www.academia.edu/8696942/Translation_Universals_or_laws_or_tendencies_or_probabilities_or..._ (18 June 2015).

Mihăilă, C. (2010). Translation studies: simplification and explicitation universals. Retrieved from http://www.slideshare.net/claudiumihaila/report-3832657?related=1 (18 June 2015).

Orlov, J. (1982). Ein Modell der Häufigkeitsstruktur des Vokabulars. In H. Guiter & M. V. Arapov (Eds.), *Studies on Zipf's Law* (Quantitative Linguistics *16*, pp. 154–233). Bochum.

Teich, E. (2003). *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts* (Text, Translation, Computational Processing, *5*). Berlin, New York.

Tirkkonen-Condit, S. (2004). Unique items—over- or under-represented in translated language? In A. Mauranen & P. Kujamäki (Eds.), *Translation universals. Do they exist?* (Benjamins Translation Library, *48*, pp. 177–184). Amsterdam, Philadelphia.

Xiao, R. (2010). How different is translated Chinese from native Chinese? A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, *15*(1), 5–35. doi:10.1075/ijcl.15.1.01xia.

Yang, H. Z., & Wei, N. X. (2002). *Yu liao ku yu yan xue dao lun* (transl. 'An introduction to corpus linguistics'). Shanghai.

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge.

Zanettin, F. (2011). Translation and corpus design. *SYNAPS – A Journal of Professional Communication*, *26*, 14–23.