

Filozofická fakulta Univerzity Palackého

Czech as source and target language  
in a comparable corpus of journalistic texts

(Diplomová práce)

2017

Pavλίna Zagolová

Filozofická fakulta Univerzity Palackého

Katedra anglistiky a amerikanistiky

Čeština jako zdrojový a cílový jazyk ve  
srovnatelném korpusu žurnalistických textů

Czech as source and target language  
in a comparable corpus of journalistic texts

(diplomová práce)

Autor: Bc. Pavlína Zagolová

Studijní obor: Angličtina se zaměřením na tlumočení a překlad

Vedoucí práce: Mgr. Michaela Martinková, PhD.

Olomouc 2017

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně a uvedla úplný seznam citované a použité literatury.

V Olomouci dne 22. srpna 2017

.....

*Bc. Pavlína Zagolová*

### **Poděkování**

Děkuji vedoucí diplomové práce Mgr. Michaele Martinkové, PhD. za její cenné rady, trpělivost a podporu, Mgr. Andree Ryšavé za exkurz do korpusové statistiky a Mgr. Lucii Chlumské, Ph.D. za užitečnou metodickou pomoc při zkoumání korpusu.

## **ABBREVIATIONS**

CET – Czech Translations from The Economist

CRO – Czech Originals from Respekt

MCC – Monolingual comparable corpus

MLC – multi-lingual corpus

PoS-grams – part of speech grams

TTR – Type-token ratio

TU – Translation universals

# TABLE OF CONTENTS

<b>ABBREVIATIONS .....</b>	<b>5</b>
<b>TABLE OF CONTENTS.....</b>	<b>6</b>
<b>1 INTRODUCTION .....</b>	<b>8</b>
<b>2 THEORETICAL FRAMEWORK.....</b>	<b>12</b>
<b>2.1 Corpus linguistics .....</b>	<b>12</b>
2.1.1 Corpus linguistics and translatology .....	12
2.1.2 Monolingual comparable corpus .....	13
2.1.2.1 Representability and comparability .....	13
2.1.3 Corpus-based vs corpus-driven research .....	15
<b>2.2 Journalism and translation .....</b>	<b>16</b>
2.2.1 Journalistic discourse.....	16
2.2.2 Translating news: Journalists as translators .....	16
2.2.2.1 From the 18th century till now .....	16
2.2.2.2 Different roles of translators vs. journalists .....	17
2.2.2.3 Aims, methods and strategies .....	19
2.2.2.4 Summary.....	20
2.2.3 Respekt and The Economist .....	20
<b>2.3 The search for translation universals .....</b>	<b>22</b>
2.3.1 Definition and history .....	22
2.3.1.1 The first stage .....	22
2.3.1.2 The second stage.....	23
2.3.1.3 The third stage .....	23
2.3.2 Current research of Translation universals.....	26
2.3.3 Research of translated Czech.....	29
2.3.4 Critical views.....	30
2.3.5 Summary and hypotheses .....	31
<b>2.4 Lexical bundles .....</b>	<b>33</b>
2.4.1 Definition and basic features .....	33
2.4.2 Typology of bundles.....	35
2.4.3 Usage and application.....	36
<b>2.5 Lexical richness .....</b>	<b>38</b>
<b>2.6 Average sentence length.....</b>	<b>40</b>
<b>3 ANALYTICAL PART .....</b>	<b>41</b>
<b>3.1 Sketch Engine .....</b>	<b>41</b>
<b>3.2 Corpus design .....</b>	<b>42</b>

3.2.1	Time span, corpus size and text length.....	42
3.2.2	Metadata .....	46
3.2.2.1	Authorship .....	46
3.2.2.2	Date of publication .....	49
3.2.3	Summary of key characteristics.....	51
<b>3.3</b>	<b>Corpus compilation .....</b>	<b>52</b>
<b>3.4</b>	<b>Data analysis .....</b>	<b>54</b>
3.4.1	Methods and terminology.....	54
3.4.1.1	N-gram extraction criteria and terminology .....	54
3.4.1.2	Statistical significance .....	54
3.4.2	Lexical bundles.....	57
3.4.2.1	3-grams .....	57
3.4.2.2	4-grams .....	63
3.4.3	Lexical richness .....	67
3.4.4	Average sentence length.....	68
<b>4</b>	<b>CONCLUSION .....</b>	<b>70</b>
<b>5</b>	<b>APPENDICES.....</b>	<b>75</b>
<b>6</b>	<b>SHRNUTÍ.....</b>	<b>76</b>
<b>7</b>	<b>LIST OF FIGURES, GRAPHS AND TABLES.....</b>	<b>80</b>
<b>8</b>	<b>WORKS CITED .....</b>	<b>82</b>
<b>9</b>	<b>ABSTRACT .....</b>	<b>89</b>
<b>10</b>	<b>ANOTACE .....</b>	<b>90</b>

# 1 INTRODUCTION

Over the years, corpus-based research has secured a prominent place among numerous tools of linguistic analysis. This powerful methodology, if used properly, allows us to explore many aspects of language in use and provides valuable insight not only into language structure but also into language teaching and second language acquisition. Corpus linguistics has been helping to shape our knowledge of translation and languages for nearly three decades. According to Zanettin, its contribution to translation theory and practice is invaluable and its impact far-reaching (2013, 20). Tymoczko goes as far as to claim that corpus translation studies will remain the central approach within the entire discipline of translation studies in the foreseeable future (1998, 1). Such a claim seems to have been appropriate as the increasing availability of information technologies makes this approach more prominent each day. Furthermore, corpus-based research contributed to the descriptive branch of translation studies in general by exploring translation norms, the notion of equivalence or translators' strategies. And it has been especially productive in research focusing on the differences between translations and their respective source texts or non-translated texts in the target language.

This paper deals with the latter; it aims to examine the language of Czech translations and compare it to the language of texts originally written in Czech. With the help of computerised tools and the corpus methodology, it focuses on the relationship of translated and non-translated language (so-called T-universals theory as presented by Chesterman 2003, 218). This theory assumes that translated texts share certain features which set them apart from non-translated texts. According to Chlumská and Richterová, leading Czech corpus translationologists, studies of translated language and the examination of possible translation laws, observable regularities and possible translation universals remain to be one of the most prominent research topics of corpus-based translation studies (Chlumská and Richterová 2014a, 17)<sup>1</sup>.

---

<sup>1</sup> "Překládový jazyk a jeho zákonitosti jsou už přes dvacet let jednou z hlavních výzkumných oblastí korpusové translatologie (*corpus-based translation studies*), která v sobě propojuje translatologické poznatky, deskriptivní přístup a metodologii korpusové lingvistiky. Ústředním bodem zkoumání se stala myšlenka, že přeložené texty vykazují určité společné rysy, jež je odlišují od textů nepřeložených, napsaných v původním jazyce." (Chlumská and Richterová 2014a, 17)



This paper does not aim to prove or disprove the existence of translation universals but rather to explore the possibility of identifying certain language features which might help us to distinguish between the original and translated Czech. The underlying hypothesis, advocated by many researchers in the field of translation studies (Koppel and Ordan 2011; Baroni and Bernardini 2006; Ilisei et al. 2010; Baroni and Bernardini 2005), is that translations (as opposed to non-translations in the same language) can be identified with the use of corpus analysis tools. So far, such claims have been mostly backed by native speakers' intuition.

This particular study will compare the language of Czech journalistic texts translated from English with the language of Czech non-translated texts of the same text type. First, it aims to investigate the presence and frequency of occurring lexical bundles (also called clusters or n-grams), i.e., frequent combinations of words that are register specific and relate to language proficiency and fluency of the author of the text and his mastery of writing in the register (Allen 2009, 105). Lexical bundles in both corpora will be examined to draw conclusions about tendencies of Czech translations from English compared to similar Czech texts which are not a result of a translation process. The analysis aims to explore a possible asymmetry in the use and variability of lexical bundles and to test the hypothesis that n-grams are more frequent in translated Czech (based on Baroni and Bernardini 2003, 377; Xiao 2011, 145)<sup>2</sup>. The overall usage of unique lexical bundles along with the total frequency of occurrence will be examined for 3-grams and 4-grams and checked for statistical significance of any differences in distribution.

Second, two more features, deemed essential by various researchers<sup>3</sup>, namely lexical richness and average sentence length, will be investigated as possible tests for establishing whether a text is a translation or a piece of original writing. According to our hypotheses, justified in section 2.3.2, both of these parameters are expected to be lower for translated Czech. Because we are concerned with numerical data which can be transformed into simple statistics, the research at hand is quantitative in nature.

This study will analyse journalistic texts published in the weekly newspaper *Respekt* which reports on the issues of domestic and foreign policy, economic, cultural and science topics. The contrastive analysis will compare the language of Czech

---

<sup>2</sup> See section 2.3.2 and for more details.

<sup>3</sup> See Ilisei et al. 2010, 504.

translations from English (articles originally published in the British newspaper *The Economist*) and the language of Czech originals on similar topics. Both the translations and the Czech originals were published in *Respekt*, only in different sections of the weekly newspaper. These two bodies of texts will represent separate subcorpora designed to fulfil the criteria of a monolingual comparable corpus.

Similar studies of translation language have already been conducted. To name just a few, Richard Xiao (2011) worked with corpora of English and Chinese texts and Sara Laviosa (1998a) examined an English corpus of newspaper articles and prose. Studies of the Czech language, for example, include a case study of simplification on a monolingual comparable corpus Jerome by Lucie Chlumská and Olga Richterová (2014a). Still, studies of translation universals usually work with multi-lingual corpora (MLC) rather than with monolingual comparable corpora (MCC). Studies using MLC are useful for exploring universal features of translation known as S-universals. These S-universals (S stands for source) cover potential features of translated texts in comparison with the respective source texts (Chesterman 2011, 176). This study aims to advance the understanding of language specific translation universals (T-universals) and to make a valuable contribution to corpus linguistics and translatology by building a unique corpus of texts that might help researchers in the future.

The first part will lay the groundwork necessary for the subsequent work on the corpus itself. A definition of a linguistic corpus and of a monolingual comparable corpus will be introduced in Section 2.1 which deals with corpus linguistics. The following section 2.2 will outline the basic features of journalistic texts and journalistic discourse in general focusing on the intersection of translatology and journalism. Section 2.3 will provide an overview of the search for translation universals from a historical point of view. It focuses on the current research and various approaches adopted by different researchers with the aim to identify possible features of translated texts. The following section focuses on three such features which will be analysed in this study: section 2.4 investigates how lexical bundles can be used as a method for exploring and comparing linguistic production; sections 2.5 and 2.6 examine the lexical richness and the average sentence length.

Section 3 will then describe the procedure of building the corpus. First, the software Sketch Engine which allows the creation of custom corpora will be introduced in section 3.1. Section 3.2 will focus on the corpus design and its basic

characteristics (time span, size, text length, authorship and date of publication). Consequently, the compilation of the corpus will follow (section 3.3) along with the data analysis (section 0). Finally, conclusions will be drawn.

## **2 THEORETICAL FRAMEWORK**

### **2.1 Corpus linguistics**

#### **2.1.1 Corpus linguistics and translatology**

The contribution of corpus linguistics to translation studies is indisputable. The use of computerised tools to examine large bodies of texts opens new possibilities for linguistic analyses and provides almost unlimited processing capacity. Heralded by the first computerised study of translations conducted by Gellerstam (1986), a whole new discipline called “corpus-based translation studies” emerged. Among the most prominent lines of research is the search for universal features of translations, a task often tackled with contrastive linguistic methods. These, according to Zanettin, are very well fitted to explore and assess translation-specific and language-specific constraints (2013, 21–22).

Mona Baker points out that translated texts have been treated as second hand distorted material and therefore excluded from many representative corpora studies. She is of the opinion that translations play a very important role in our lives and they are worth studying: “...translated texts record genuine communicative events and as such are neither inferior nor superior to other communicative events in any language. They are however different, and the nature of this difference needs to be explored and recorded.” (1993, 234). Jiménez-Crespo comments on the unavoidable intersection of information technology and of the search for universals which might go far beyond the corpus methods that have become so prominent in the recent decades. He expresses his view that research into universal tendencies of translation might not only help us understand generalities in the translation process but might also shed light on the effect and the impact of technology on translatology (Jiménez-Crespo 2010). Among others, also Chesterman considers this line of research to be highly beneficial in terms of methodological advancement. “Corpus-based research into translation universals has been one of the most important methodological advances in Translation Studies during the past decade or so, in that it has encouraged researchers to adopt standard scientific methods of hypothesis-testing.” (2003, 226).

## 2.1.2 Monolingual comparable corpus

As stated above, the corpus compiled for the purposes of this study fits the label “monolingual comparable corpus”. An MCC is a computerised body of texts comprising translated texts and non-translated (original) texts in the same language. These are collected and compiled in a way that ensures their comparability (e.g. texts of the same text-type, from the same setting, texts on a certain topic or from a given period etc.). MCCs are especially suitable for identifying typical features of translated language; in Bernardini’s terms “a quantitative analysis carried out on an MCC has a number of advantages for corpus-based research in translation studies” (2011, 11). Along similar lines, McEnery and Xiao observe that comparable corpora offer numerous possibilities for translation studies. They allow us to investigate typological, cultural and universal language-specific features and increases our knowledge of differences concerning source texts and translations (2008, 1). Laviosa highlights the use of MCCs for the investigation of possible explicitation, simplification, normalization, the law of interference<sup>4</sup> and also the unique items hypothesis<sup>5</sup> (2010, 83).

### 2.1.2.1 Representability and comparability

Chesterman stresses that one of the weak points of studies based on an MCC is the assumed representativeness of texts included in the corpora. In order to ensure that the corpus in question is representative and truly comparable, it is of the utmost importance to mind the criteria for selection of such “comparable texts” (2003, 214–215). The author mentions “an awareness of the need for a text typology which would allow valid comparisons to be made between representative sets of texts...” (ibid.). Chlumská shares this view that representative selection of data is a major concern and suggests as many selection criteria as possible, namely text-type, genre, time period and size (2014, 228). Baroni and Bernardini propose to aim for broad comparability of the corpora (MCC) involved to minimise the risk that results are invalidated by methodological problems (2003, 367–368). Laviosa, who conducted research on an

---

<sup>4</sup> See sections 2.3.1.3 and 2.3.2 for more details concerning these terms.

<sup>5</sup> The “unique items hypothesis” developed by the Finnish researcher Sonja Tirkkonen-Condit aims to describe general tendencies of translated language. It suggests that translations contain fewer unique items than the language of originals. The label “unique items” describes elements which are common or frequent in the original language, but manifest differently (lack their linguistic counterparts) in different languages (Chesterman 2007, 4).

MCC of narrative works in English, mentions aspects such as genre, time span, distribution of male/female authors and team/single authorship along with the overall size of each component (total word count). In addition, she includes the basic characteristics of the target readers, namely their literacy. On the other hand, she admits that while including such additional criteria contributes to the overall comparability, it also restricts the use of such corpora for other purposes or studies (Laviosa, 1998a, 4–5).

In a section dedicated to planning the construction of a corpus, Meyer also mentions the overall size of the corpus, types of genres, the length of the individual text samples, the range of speakers or the time-frame. He also includes sociolinguistic variables such as gender balance, age, the level of education, dialect variation, social contexts and social relationships. He concludes that planning of a valid and representative corpus mostly depends on its intended use which is a shared responsibility of both the corpus creator and the subsequent users (Meyer 2004, 30–53). This view is further supported by Kenny who comments on the corpora design in general: “Design criteria crucially depend on the envisaged use of the corpus and centre on the idea that corpora should be somehow ‘representative’ of a particular type of language production and/or reception.” (1998, 50).<sup>6</sup>

We can conclude that the more criteria for selection of texts applied, the more common ground providing support for the claim of comparability. Nevertheless, the rising number of restricting criteria imposes further limitations on the data selection. In general, too many criteria might considerably hinder the corpus creation, not to mention time and money spent to satisfy it. The main issue in this thesis, as we will see later, is the disproportionate number of texts published as originals or translations in a given language combination (Czech originals vs Czech translations from English). Further complications arise when confronted with the fact that not all the desired metadata (in this case the author(s) name and the translator’s name) is available. Sometimes publishers do not provide it or it is not accessible to the researcher. Taking all this into account, a certain balance between the number of selection criteria and the practical implications for the researchers should be achieved, aiming for the best optimal comparability.

---

<sup>6</sup> In this respect, Chlumská (2014) remarks that the issue of representability needs to be related to a specific type of corpus in question; representability of parallel corpora is achieved differently than in comparable corpora (17).

### 2.1.3 Corpus-based vs corpus-driven research

There are two basic approaches to corpus-based studies which reflect the researcher's mode of work, his/her relative preoccupation with a certain hypothesis and its position within the research. If we see corpus linguistics as a method or as a tool for exploring certain preconceived hypotheses or theories, we speak of "corpus-based research". "Corpus-driven linguistics", on the other hand, "rejects the characterisation of corpus linguistics as a method and claims instead that the corpus itself should be the sole source of our hypotheses about language. It is thus claimed that the corpus itself embodies its own theory of language." (Tognini-Bonelli 2001, cited in McEnery and Hardie, 2012a, 6). The main distinction is thus whether we see corpus linguistics as a mere method or as a theory<sup>7</sup>. The corpus-driven label is often perceived as synonymous with a "bottom-up approach" whereas the corpus-based approach represents a "top-down approach" (ibid., 151). However, the authors also admit that this binary distinction is not always so clear-cut and it is not unusual for a researcher to adopt a more fluid approach—such is the case of this particular study.

---

<sup>7</sup> This latter view is common for a group of scholars referred to as the neo-Firthians.

## **2.2 Journalism and translation**

### **2.2.1 Journalistic discourse**

The primary function of journalism is to inform, in other words, “[to] supply[...] citizens with the information they need to make decisions about their lives” (Detrani 2011, 87). A hallmark of a journalistic product is thus information quality, an elusive concept hard to define yet at the same time sought by the majority of journalists who wish to adhere to the ethical codex and professional standards. The relevant pieces of information are shaped and manipulated according to specific needs and circumstances under which they are being produced. In this respect, Apostol et al. (2015, 146–147) mention the nature of the information, the transmission channel and the type of journalistic text (journalistic genre) but there are many other factors which determine the final journalistic product. For the purposes of defining this particular discourse, the authors propose features such as the institutional nature, the ability to respond to expectations, the audience’s segmentation level and the acceptability (ibid. 148). Bielsa and Bassnett conclude that “It could be argued that the main objective of news translation is the fast transmission of information in a clear way so that it can be communicated effectively to readers.” (2009, 63)

When it comes to linguistic features, typical features of journalistic texts are their narrative and referential nature which manifest in reporting on short stories or incidents, descriptive statements, and the use of current language (Apostol et al. 2015, 148). The authors also highlight the presence of redundancy (which is apparent throughout the entire text, from the title and the headline to its body) and text coherence (149–150).

### **2.2.2 Translating news: Journalists as translators**

#### **2.2.2.1 From the 18th century till now**

Nowadays, due to the globalised nature of media production, it is not unusual to encounter journalistic texts originally published elsewhere and translated into the language of the (new) target audience. On the contrary—translation of news is quite common, if not omnipresent. Time and space constraints, the constant pressure to publish as fast as possible, cover diverse topics and suit the needs of the prospective readers have transformed the journalistic profession to a great extent. Nevertheless,



as reported by Valdeón, the role of translation in journalism is greatly underestimated. The author believes that translation has always been part of this profession ever since its conception in the 18th century (Valdeón 2012, 851). To the author's best knowledge, very few studies dealing with journalistic production with respect to translation can be found: "In fact, it was not until the first decade of the 21st century that a number of translation scholars gradually became concerned with the work of news translators." (ibid). In his article, he puts forward the idea that the rise of the internet brought forth an increased demand for translation of news and helped to ensure its current status. Valdeón supplies a comprehensive overview of the historical development of the role of translation within the journalistic profession and concludes that "irrespective of the changes that have characterised the evolution of journalism as a profession, translation has remained central..." (2012, 862–863).

Van Doorslaer argues that because newsrooms all around the world do not employ translators, translation becomes just one of many tasks undertaken by the professional journalists whose job description includes "information gathering, translating, selecting, reinterpreting, contextualizing and editing" (2010, 181). In this respect, Bielsa and Bassnett mention the newly established term "transediting"<sup>8</sup>, which denotes the common practice halfway between editing and translating (2009, 63). This term "describes the form that translation takes when it has become integrated in news production within the journalistic field" (Bielsa and Bassnett 2009, 64). Apart from transediting, Valdeón introduces one more coined term: "'Tradaptation' or transadaptation, "where translation [goes] beyond word-for-word replacements and suggest[s] more fundamental transformations." (Valdeón 2014 53). Generally speaking, journalistic translations are rarely considered to be "proper translations", they are rather viewed as rewritings, edited versions and variations on the source text(s).

#### **2.2.2.2 Different roles of translators vs. journalists**

When dealing with the nature of journalistic translations, it is vital to realise that the roles of translators and of journalists are essentially different. The translator's loyalty lies with the author of the original and any distortion of the sense of the original text is as a rule undesirable. On the other hand, a journalist's main responsibility is determined by the code of ethics and the standards of his journalistic profession, which

---

<sup>8</sup> Proposed by Karen Stetting in 1989.

above all value information quality, objectivity and responsibility. To quote from a code of ethics available on the website of the Society of Professional Journalists, their task is to “seek truth and report it” (‘SPJ Code of Ethics’ 2014). Some other more specific requirements state that journalists should:

- *Take responsibility for the accuracy of their work. Verify information before releasing it. Use original sources whenever possible.*
- *Remember that neither speed nor format excuses inaccuracy.*
- *Provide context. Take special care not to misrepresent or oversimplify in promoting, previewing or summarizing a story.*
- *Gather, update and correct information throughout the life of a news story.*
- *Recognize a special obligation to serve as watchdogs over public affairs and government.* (‘SPJ Code of Ethics’ 2014)

To what extent these requirements clash with the ethics of a translator’s work is beyond the scope of this paper. However, as the examples above seek to illustrate, there are substantial differences between—if not objectives—certainly priorities of these two professions. For example, the act of providing context, updating and correcting information would certainly be problematic from the perspective of a translator. Van Doorslaer also recognises the uncertain status of translation in journalistic production: “...because translation is everywhere, there are no formal translator positions. The relativity of both the status of source text and authorship creates a situation that is opposite in many respects to the position of translation in traditional research on literary translation, for example, where the author and the ‘sacred original’ are of central importance.” (2010, 183).

In Valdeón’s view, the author of the original assumes a peripheral role. The translation—in this case rather an adaptation—is seen as a valuable procedure necessary to produce a good target text. And yet, even though the translator’s role is crucial, an interesting paradox arises: the person behind the translation remains hidden. “Additionally, in news production, the process remains far more invisible than in the case of canonical texts. News consumers are rarely aware of any translation processes, let alone of any ideological shifts aimed at infusing the target versions with new meaning.” (2014, 53). Bani (2006) also comments on the absence of the translator’s name: “The indication of a translator’s identity is not always available in newspapers; on the contrary, there are many cases in which the translator is completely invisible from the graphic point of view, where the name is missing or only the initials are

indicated or it is difficult to find the name inside the newspaper.” (35). She believes that this practice not only makes it very difficult (sometimes impossible) to trace the author/translator of the text, but it also contributes to the general confusion about the text’s origin.

### **2.2.2.3 Aims, methods and strategies**

When adapting a news article, the journalist must deal with several issues caused by the fact that a different medium aims to attract readers from a different country. These prospective readers not only speak different languages, but they come from different socio-cultural backgrounds. They have different values, interests and needs. Among numerous changes which occur in order to satisfy these differences, the authors mention a possible change of titles and leads<sup>9</sup>, elimination of unnecessary information, the addition of important background information, changing the order of paragraphs or summarising information (Bielsa and Bassnett 2009, 64). The journalist is not a mere translator but rather an independent re-creator of the media contents and his dominating translation strategy is domestication and linguistic adaptation (ibid. 72, 104). These strategies aim to facilitate comprehension and compensate for the lack of background knowledge among the new target readers. “The purpose of news translation is to adapt texts to the needs of different publics, which requires not only reorganizing and contextualizing information, but also an exercise of subtle rewriting in order to heighten the effectiveness of the original text in the new context,” conclude Bielsa and Bassnett (2009, 104). Gambier (2006, 14) identified four main translation strategies employed in news translation; apart from previously mentioned re-organization and addition, he proposes deletion and substitution.

Valdeón points out that there has been very little research into transformations of journalistic translations. He observes that apart from domestication, journalistic translations to a large extent utilise framing (2014, 51). Above all, this strategy takes into account the target readers and aims to produce a text which the new audience can identify with. Framing manifests both on a linguistic and paralinguistic level: “In news translation, this entails the adaptation of a text for the target readership, a process can lead [sic] to appropriation of source material.” (ibid).<sup>10</sup>

---

<sup>9</sup> A lead is the opening paragraph of an article.

<sup>10</sup> According to Zelizer and Allan, it employs linguistic tools such as metaphors, examples and so-called catch-phrases but also images and visual material in general (2010, 48).

Gambier also calls for raising awareness about the context of news translation as it largely determines the translation product: “International news communication cannot be analysed merely as a matter of isolated news texts. Translation Studies has emphasized, in recent decades, the importance of context and contextualisation in the translating process, in the decisions made by translators.” (Gambier 2006, 14).

#### **2.2.2.4 Summary**

To sum up, translation of journalistic texts is without a doubt an interesting subfield of translation which deserves further examination. The nature of journalistic texts, the role of translators (journalists), their aims, strategies, specific working conditions and constraints, all of this deserves researchers’ attention. The previous sections offer just a brief overview of some specific features of the journalistic production in order to describe the special nature of the texts included in both corpora compiled for the purposes of this study. Even though this paper does not aim to explore features of journalistic translations in detail, it must be noted that these specific circumstances of news production might significantly influence the final translation product. The extent of this potential influence is yet to be determined not only by a direct comparison of source and target texts in a suitable parallel corpus<sup>11</sup> but also by scrutinising the wider communicative situation.

#### **2.2.3 Respekt and The Economist**

The following paragraphs briefly introduce *Respekt* and *The Economist*, weekly newspapers where the original Czech texts (*Respekt*), Czech translations (*Respekt*), and their English source texts (*The Economist*) were published.

*Respekt* is a Czech weekly magazine founded in 1989. It is distributed in print, on the web and also through a mobile application. To quote from the official website, *Respekt* covers both domestic and foreign affairs, it deals with topics such as politics, economy, history, societal issues and trends and also covers news concerning science, research and culture (‘Respekt’ 2017).

---

<sup>11</sup> A parallel corpus consists of original texts and their respective translations.

According to the publishing house “Economia”, the prospective readership of *Respekt* can be characterised as follows (‘Respekt - Inzerce’ 2017):

- 61% are men
- 44% are people between 30–49 years of age
- 40% are university graduates, every third reader achieved secondary education attested by a diploma (“maturita” exam)
- 25% are entrepreneurs, 17% managers
- 51% live in a household of a higher standard of living (AB classification)

*The Economist* is an English-language weekly newspaper (published in a magazine format) founded in 1843. All of the articles are also published on numerous platforms both printed and web-based. Again, to quote from the official web page, it provides coverage of international news and deals with topics such as politics, business, finance, science and technology (‘The Economist: About Us’ 2017).

## 2.3 The search for translation universals

### 2.3.1 Definition and history

Translation universals (sometimes called universal or general tendencies, translation norms or laws of translation) have always been discussed by many scholars who wanted to deepen their understanding of the relationship between source and target texts and the understanding of the translation process. The term translation universals (TU) covers features that are claimed to be characteristic to all translations regardless of other variables such as language pairs, text-types or different historical periods (Chesterman 2004, 3). These universal features, in turn, provide a basis for the hypothesis that translated texts can be distinguished from non-translated texts. Although some researchers are very sceptical about this hypothesis, others embrace it. It remains uncertain whether such universal features exist, what their nature is and to what extent they are present in translations. Regardless of this discord among translatoologists, the search for TUs is still under way and it has even gained fresh impetus from new research tools available. “Since the emergence of Corpus-Based Translation Studies, research into potential regularities in translational behaviour or ‘general tendencies in translation’ has been at its core.” (Jiménez-Crespo 2010, 1).

The search for these universal features has a long tradition in translation studies even though the name for this notion varies. From a historical perspective, Chesterman (2003) talks of three distinct stages of the search for translation universals: ideal universals, pejorative universals and descriptive universals. These will be briefly dealt with in the following subsections.

#### 2.3.1.1 The first stage

The first stage called “ideal universals” marks the first attempts at formulating what a good translation should look like. According to Munday (2009, 25–27), such a prescriptive approach can be seen in the works of St. Jerome, Etienne Dolet or Alexander Fraser Tytler<sup>12</sup> who attempted to sum up some general laws of translation. However, Dolet’s five “principles of translation” and Tytler’s three “general laws of translation” are obviously bound to serve as criteria for distinguishing good and bad translations, rather than postulating universal features of translation. Needless to say,

---

<sup>12</sup> St. Jerome (4th century), Dolet (16th century), Tytler (18th century).

even these scholars were aware of the need to differentiate between translations of different text-types (e.g. sacred texts vs. other types of texts) so this prescriptive approach is far from being truly universal and suffers from overgeneralization (Chesterman, 2003, 214). In addition, there is another problematic aspect, namely the fact that judging which translation is good or bad is often rather subjective.

### **2.3.1.2 The second stage**

The second stage called “pejorative universals” is similar to the first, as it is also quite general, but there is an important difference in the perception and evaluation of the universals. “Here, all translations (or: all translations of a certain kind) are regarded as being deficient in some way...” (Chesterman 2004, 5). This long-standing pejorative approach compared target texts to the respective source texts and assumed that all translations undergo undesirable changes and shifts which have negative effects on the quality. In fact, translated texts were considered inferior to originals (Chesterman 2010, 38). One of the major proponents of this approach was Antoine Berman who spoke of “universals of deformation” and proposed that it was “the system of textual deformation that operates in every translation and prevents it from being a ‘trial of the foreign’” (1985, 286). Berman identified twelve major deforming tendencies in the domain of literary prose, for example, rationalization, clarification, expansion, popularization or qualitative and quantitative impoverishment. Although the author based his analysis on his own experience as a translator into French, he asserted that these tendencies can be found also in English, Spanish or German translations (286–288).

### **2.3.1.3 The third stage**

The third stage called “descriptive universals” differs from the previous stage as it is marked by a different approach: “Where the pejorative approach would be critical of translationese or interference, then, this descriptive approach simply accepts that translations will be inevitably influenced by formal features of the source text” (Chesterman 2003, 2018).

The hypothesis that the language of translated texts is essentially different even gave rise to the idea that it constitutes a hybrid code, so-called “third code” distinct from both the source and target language (Chesterman 2003, 218). This term was introduced by William Frawley (1984); Swedish scholar Gellerstam (1986) on the

other hand preferred the term “translationese” mentioned above. According to Lind, a third code (or “translationese”) can be described as “a product of negotiation of the translator between the first code of the source text, language, and culture, a product that differs not just in obvious ways from its source, but also from native texts of the ‘second code’” (2007, 1). For example, Wollin’s definition of translationese compares the features of such texts to the translator’s fingerprints<sup>13</sup> caused by his contact with the source text and the language of the original (2005, 1508).

Gellerstam explored the differences between original Swedish fiction texts and translations from English (Santos 1995, 59). He arrived at a conclusion that aside from cultural differences, the texts differ because of different styles of the original languages. For example, translations contained fewer instances of Swedish colloquialisms, a higher number of English loanwords, words described as international false friends, evaluative adjectives or verbs of feeling (Santos 1995, 60). He concluded that these lexical differences, observable by comparing relative frequencies of certain words (overrepresented in translated texts), might work as a trigger which helps the reader to identify “translationese” (Wollin 2005, 1509).

Among others, Chesterman highlights Blum-Kulka’s contribution to the search for descriptive universals (Chesterman 2004, 7). Blum-Kulka (1985) introduced the “explicitation hypothesis” by describing shifts in cohesion and coherence. She argues that translated texts, as a rule, exhibit a higher degree of explicitness and redundancy (Blum-Kulka 1985, 299–300). However, the rise of this line of research is usually connected to Mona Baker’s seminal paper “Corpus linguistics and translation studies—implications and applications” (1993). Baker advocates for the existence of “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” (243). The features as proposed by Baker include explicitation, normalization, simplification and levelling out.

- **Explicitation** “is an overall tendency to spell things out rather than leave them implicit in translation” claims Baker (1996, 180) and proposes that this universal can be observed in textual phenomena such as text length (purportedly longer than in STs) and punctuation (TTs prefer punctuation marks of weaker rank, such as using semicolons and or periods instead of

---

<sup>13</sup> This is probably a simile borrowed from Gellerstam (1986).



commas, and avoid what Baker calls the “experimental use”), or in its simplest form by adding background information (*ibid.*, 176, 180–182). According to Xiao, a higher degree of explicitation is achieved also by the use of reformulation markers (such as *that is*) and more frequent use of connectives (2011, 146–147).

- **Normalization** is defined as “a tendency to exaggerate features of the target language and to conform to its typical patterns” (Baker 1996, 182). Zanettin proposes that distribution of collocations, the use of colloquial words or the use of creative collocations might serve as an indicator of the normalization tendency (2013, 24)
- **Simplification** is “the idea that translators subconsciously simplify the language or message or both” (Baker 1996, 176). Simplification thus involves facilitating decoding and processing on the side of the recipient (Baker 1996, 182). Simplification manifests itself at the lexical, syntactic and also stylistic level (Xiao 2011, 146). For simplification, Zanettin proposes indicators such as a type-token ratio<sup>14</sup>, a ratio of function to content words and average sentence length (2013, 23).
- **Levelling out** (also convergence) is defined as “...the tendency of translated text to gravitate towards the centre of a continuum” (Baker 1996, 184). The author also points out that this feature is independent of both target and source language and can be observed through indicators such as lexical density, type-token ratio and sentence length.

Malmkjær, on the other hand, considers Gideon Toury (1995) to be the forerunner of the search for descriptive universals (Malmkjær 2008, 6). But Toury preferred to think of any such laws in terms of their non-absolute nature and thus proposed “the law of standardisation” and “the law of interference”. The law of standardisation states that “textual relations obtaining in the original are often modified [...] in favour of (more) habitual options offered by a target culture” (Toury 1995, 268). The law of interference describes „phenomena pertaining to the make-up of the source text tend to be transferred to the target text“ (275). Toury later explained why he avoided the term “universals”: “The reason why I prefer ‘laws’ is not merely because, unlike ‘universals’, this notion has the possibility of *exception* built into it...

---

<sup>14</sup> See section 2.5 for more details concerning the type-token ratio.

but mainly because it should always be possible to explain away [seeming] exceptions to a law with the help of *another* law, operation on another level.”<sup>15</sup> (Toury 2004, 29). This reluctance to speak of “universals” of translation is typical not only of Toury but also of many of his contemporaries and successors (who will be introduced in the following sections).

### 2.3.2 Current research of Translation universals

Current research in the descriptive universalist approach according to Chesterman falls into two categories. Some researchers look for S-universals (S stands for “source”) and aim to formulate universal statements about the differences between translations and their respective source texts. The second approach focuses on T-universals (T stands for “target”): researchers attempt to formulate a hypothesis about translated and non-translated texts in the same language (2011, 176). It deals with the way translators process the text and it assumes the possible existence of features that are common to all translated texts. These might include simplification, conventionalization, atypical lexical patterning or under-representation of target language specific items (Chesterman 2004, 7–8).

The current research (from the late 1990s onwards) is marked by the introduction of corpus studies which greatly contributed to translation studies and especially to the search of translation universals. Laviosa’s influential research on T-universals was conducted on an English comparable corpus (consisting of journalistic texts and narrative prose from multiple source languages) in 1996-1998. The author discovered noticeable differences between translated and original texts: “the translated articles use a relatively lower proportion of lexical versus grammatical words independently of the source language, as well as a higher proportion of frequent versus less frequent words” (1998a, 1). Laviosa thus proposes two additional features of translated texts: relatively greater repetition of the most frequent words and less variety in the words most frequently used (1998b, 4). This led her to formulate a

---

<sup>15</sup> “For instance, an expected phonetic change that does not occur (which is always a possibility) is often justified as evidence of having been created at a later period, when the law had stopped being active, or as an evidence of having been imported from without, in a situation of language contact, or as a result of a combination of the two.” (Toury 2004, 30)

hypothesis that translations exhibit a lower lexical density and mean sentence length<sup>16</sup> (1998a, 4).

However, Laviosa does not speak of translation universals, she uses the term “patterns of lexical use”. Even though her MCC and the respective subcorpus of translated texts included material from multiple source languages, she does not explicitly state that the observed features constitute universal features of translation in general. Nevertheless, she proposes that the above-mentioned features, the core patterns of lexical use, “may prove typical of translational English in general” (1998b, 4). All the same, she proposes to use the outcome of her investigation as a hypothesis to be further tested on various text genres and types of translation including interpreting (1998b, 9).

Some researchers such as Koppel and Ordan (2011, section 1) claim that specific features of translated texts—some of which might be caused by language interference—and the knowledge of their existence might in itself be sufficient to determine if a given text is an original text or a product of a translation process. They go as far as to propose that these features are sufficient to identify the source language (*ibid.*). This view is supported by Baroni and Bernardini who refer to an experiment which shows that both humans and computer algorithms are very successful in telling translated texts from original texts of the same genre and dealing with the same topic (Baroni and Bernardini 2006, 3–4). “From the point of view of translation studies, our results are of interest because they bring clear evidence of the existence of translationese features even in high quality translations.” (2006, 4).

An experiment which tested human subjects’ success rate in telling translations from non-translations was conducted in 2005. When faced with the same task as a computer, all of the participants (translators as well as non-translators) were able to identify translation language above chance level (the average success rate being 70.61%); there were no significant differences between translators and non-translators (Baroni and Bernardini 2006). The human subjects were outperformed by the computer, which reached the success rate of 74.4%. The authors report that “at least for the particular data-set we considered, it is indeed possible to speak of a translationese dialect on objective grounds, given that an algorithm is able to identify

---

<sup>16</sup> This hypothesis was not confirmed because the translational texts in her corpus proved to have higher sentence length. The author proposed a study on a bigger and more varied corpus to explain this unexpected tendency (Laviosa 1998, 8).

translated text with good accuracy” (Baroni and Bernardini 2005, sections 7+8) In addition to the appearance of function words, Baroni and Bernardini advocate for exploring features such as part of speech grams (PoS-grams)<sup>17</sup>, the use of specific pronouns and the number of adverbs (2005, section 8).

A further proof of this claim that translations can be identified was provided by Ilisei et al. (2010) who worked with a corpus of Spanish texts in medical and technical domains. The researchers’ aim was to train a computer to distinguish between translated and non-translated language using various features (classifiers). “The outstanding accuracy provided by several classifiers is evidence that translations can indeed be identified,” they conclude (Ilisei et al. 2010, 510). Based on their experiment, the authors claim that “translated texts exhibit lower lexical density and richness, seem to be more readable, have a smaller proportion of simple sentences and appear to be significantly shorter, and discourse markers were used significantly less often.” (Ilisei et al. 2010, 504). They state that the most useful features are lexical richness, sentence length, and the proportion of grammatical and lexical words respectively (Ibid., 508).

Baroni and Bernardini conducted research on collocational differences in an MCC consisting of official reports submitted by different EU countries (both originals and translations) and in an MCC of Italian texts containing also both originals and translations from various source languages (Baroni and Bernardini 2003). The authors also came to the conclusion that there are noticeable differences between translations and originals concerning repeating certain patterns, namely that bigrams<sup>18</sup> are more frequent in translations than in non-translations (2003, 377). However, a closer analysis revealed that this is true only when topic-dependent bigrams are considered; topic-independent bigrams were as common in the subcorpus of original texts: “It does seem that translated language is repetitive, possibly more repetitive than original language. Yet the two differ in what they tend to repeat: translations show a tendency to repeat structural patterns and strongly topic-dependent sequences, whereas originals show a higher incidence of topic-independent sequences...” (Baroni and Bernardini 2003, 379)<sup>19</sup>. Among others, Biel (2009), who investigated the potential

---

<sup>17</sup> PoS-grams are n-grams viewed as strings of part of speech categories.

<sup>18</sup> A bigram is a string of two uninterrupted word-forms.

<sup>19</sup> The “strongly topic-dependent” bigrams include expressions referring to a specific language, minority or to a geographical location, while the “topic independent” include more general deictic expressions or metadiscoursal items.

of corpus tools for exploring the language of legal translations<sup>20</sup>, suggests examining possible over-representation and under-representation of linguistic features and the presence of untypical collocations (Ibid.).

Further evidence supporting Baroni and Bernardini's claim may lie in the findings of Xiao (2011), who explored word clusters in translated and non-translated Chinese. The author reasons that translators tend to use recurring patterns because they want to achieve improved fluency in their writing (145). "[T]his aim for greater fluency, according to Baker (2004, 173), is caused by the "social pressure to produce fluent (and hence unmarked) language." Bisiada (2015, 24), on the other hand, links this tendency to the translator's aim at improving the text's readability.

### **2.3.3 Research of translated Czech**

Studies of Czech translated language first appeared in 2007. The explicitation hypothesis as introduced by Blum-Kulka was tested by Konšalová (2007), who focused on its manifestation on the morpho-syntactic level (e.g. frequencies of finite verbs or infinitival constructions). Also Kamenická (2007) dealt with the explicitation hypothesis but rather than examining its manifestation she aimed to redefine it, arguing that explicitation is rather an umbrella term for phenomena which bear resemblance to one another, claiming that it is "a prototype category" (Kamenická 2007, 55). Kubáčková (2009) presented her research on generalization and specification of lexical meaning as potential translation universals, stressing that any claims of universality should be based on quantitative analysis of a larger corpus (Kubáčková 2009, 47–48). What she worked with was an English-Czech corpus comprising monolingual, multilingual and parallel subcorpora, Her study, which combined quantitative and qualitative methods, concludes that "[g]eneralization is observed as a weak but universal tendency of translated texts in monolingual comparable corpora." (47).

Chlumská (2015) explored the basic features of translated Czech as exemplified in the MCC corpus Jerome. This corpus, which includes both fiction and non-fiction texts, was created to allow the study of general frequencies, parts-of-speech distribution, and n-gram analyses. Its design aimed to reflect the proportion of

---

<sup>20</sup> Biel speaks of "the textual fit hypotheses", that is "how the translated language (translationese) differs from the non-translated language." (2009, 11)

translated and original literature available at the time. Chlumská stated that certain differences which point toward features of simplification and convergence can be found, along with examples of unusual lexical patterning (2015, 150). Nevertheless, she remained cautious as to claim that these features are truly universal because the observed trends and differences in frequency distribution were not so prominent as some previous studies suggest<sup>21</sup>. She concluded that the corpus had its limitations and an analysis of the respective source texts would be needed to fully understand the observed tendencies (Chlumská 2015, 151).

### 2.3.4 Critical views

To conclude this section, it is necessary to mention that there are also numerous translation scholars who strenuously oppose the claim of translation universals. A recent paper by Evans and Levinson (2009) fittingly named “The Myth of Language Universals” argues against the existence of any universal tendencies. The authors claim that all such tendencies along with the proof of their existence remain unconvincing and very sparse. “Although there are significant recurrent patterns in organization, these are better explained as stable engineering solutions satisfying multiple design constraints, reflecting both cultural-historical factors and the constraints of human cognition.” (2009, 429).

Among others, House is an outspoken adversary to this line of research, claiming that “the quest for translation universals is in essence futile, i.e. that there are no, and there can be no, translation universals” (2008, 11). House argues that researchers who look for TU disregard issues such as the directionality in translation, the specification of language-pair, genre and diachronic language development (2008, 11–12). Also Chesterman acknowledges that the search for TU is far from being conclusive: “If a hypothesis is found to hold only for a subset of translations, we cannot call it a universal.” (2003, 220). He proposes several conditions which should always be considered before making any strong claims:

---

<sup>21</sup> “S vědomím všech výše zmíněných faktorů je možné konstatovat, že překladová čeština se od nepřekladové češtiny skutečně liší, ale hned vzápětí je třeba dodat, že odhalené rozdíly zdaleka nejsou tak výrazné a zásadní, jak by se na základě formulovaných hypotéz a předchozích translátologických prací mohlo zdát.” (Chlumská 2015, 150)

- **Language-bound condition:** features typical of a given pair of languages and translation direction.
- **Time-bound condition:** features of a particular period or a culture.
- **Type-bound condition:** features typical of a particular text-type, genre or skopos type.
- **Translator-bound condition:** features pertaining to a specific time period of a particular translator.
- **Situation bound condition:** possible publishing house policies and editorial conventions (Chesterman 2003, 220–221).

Some authors criticise the individual labels (e.g. explicitation and simplification) and consider Baker's account of TU to be too simplistic and repetitive. For example, House thinks that the labels are far too general and imprecise (House 2008, 11). Along the same lines, Pym points out that "all four [Baker's universals] appear to be saying much the same thing" (Pym 2008, 10). He believes that Toury's "law of interference" covers all of the proposed TU and criticises an obvious overlap of explicitation, simplification and normalization (Ibid.). Also Chesterman recognises the problem of the operationalization of TU; he calls for an explicit description of the used methodology and for more research replication (Chesterman 2003, 223). Pym further criticises the fact that Baker's account does not include language interference (Pym 2008, 14–15). As reported by Koppel and Ordan, differences between translations from different source languages reflect general differences between the languages in question. The authors share Pym's view that such translations "can be distinguished from each other and that closely related source languages manifest similar forms of interference" (2011, section 1).

### **2.3.5 Summary and hypotheses**

When studying features of translated language, it should be noted that calling any linguistic phenomena truly universal would be a strong claim. It would require extensive studies on large bodies of texts from different discourses and different language pairs and combinations. All prospective results and observed tendencies should be related to specific language material and its basic characteristics (domain, language pair, directionality of translation etc.). As can be seen in the previous paragraphs, further research will be required to either validate or disprove Baker's hypothesis, which even today, almost twenty-five years later, stirs passionate debates

among translators and linguists. Redefinition and specification of the proposed tendencies along with a more precise methodology are required to produce valid results.

This study is not so ambitious as to try to solve this issue. But it aims to contribute to the ongoing debate with a small pilot study on Czech data and to explore some possibilities for studying features of translated language. Based on the available literature dealing with possible differences between translated and non-translated language, three main criteria or tests (applicable for the corpus at hand) for distinguishing translated and original texts were established: the distribution of n-grams, the lexical richness and the average sentence length.

Three hypotheses as mentioned in the introduction were formulated:

1. N-grams are more frequent in translated texts than in non-translated texts in the same language (based on Baroni and Bernardini 2003, 377; Xiao 2011, 145).
2. Translated texts exhibit lower lexical richness than non-translated texts in the same language (based on Ilisei et al. 2010, 504; Zanettin 2013, 23; Laviosa 1998a, 4).
3. Translated texts exhibit lower average sentence length than non-translated texts in the same language (based on Zanettin 2013, 23; Laviosa 1998a, 4).

Based on these three specific hypotheses, the general hypothesis that translated texts can be distinguished from non-translated texts in the same language with the use of corpus analysis tools will be tested. The following sections provide further information about the examined textual features, focusing on its application and possible limitations.



## 2.4 Lexical bundles

### 2.4.1 Definition and basic features

The first part of the analysis will focus on the presence of frequently occurring combinations of words, so-called lexical bundles (also n-grams, word clusters or formulaic language). For example Stubbs (2004) provides the following definition of such recurrent phrases: “The simplest definition of a phrase is a string of two or more uninterrupted word-forms which occur more than once in a text or corpus...” (118). Bibet et al. define n-grams as structural units “identified empirically, as combinations of words that in fact recur most commonly in a given register, derived formulaic units of language which are register-specific and perform a variety of discourse functions” (Biber et al. 1999, 992). Hyland calls them “extended collocations which appear more frequently than expected by chance, helping to shape meanings in specific contexts and contributing to our sense of coherence in a text.” (2008, 4). He agrees that their occurrence is important for establishing a distinctive register and for differentiation between texts from different disciplines and genres (Ibid., 4–5).

The most relevant feature for the identification of lexical bundles is the frequency (Biber et al. 1999, 990–991, Allen 2009, 106): “A combination of words must recur frequently in order to be considered a lexical bundle.” (Biber et al. 1999, 990). According to Biber et al. (1999, 992) the threshold of minimum occurrence of a given sequence in order to be considered a lexical bundle is ten times per million words and to satisfy the second criterion, it must be spread across at least five different texts. However, they admit that as the frequency significantly drops with the length of the examined sequences, a lower threshold (at least five instances per million words) is allowed for five-word and six-word bundles (1999, 992–993).

Salazar mentions that these criteria of extraction based on the minimum frequency of occurrence account for one of the distinguishing characteristics of lexical bundles, that is their fixedness (2014, 14). She also highlights their compositional nature which (in contrast with idioms) allows to derive their meaning from the individual words and also permits variation and “positional flexibility” of such sequences of words (ibid., 14–15). “As for their structure, the large majority of lexical bundles are not complete structural units, but rather parts of phrases or clauses with embedded fragments,” concludes Salazar (2014, 15).

The presence of appropriate lexical bundles in writing is an important feature which indicates that the language user (writer) is skilled. i.e., he/she is a fluent and competent user with sufficient communicative competence (Hyland 2008, 5). Pawley calls attention to lexical bundles' capacity of establishing a style and highlights their importance as "the main building blocks" which "play a key role in linguistic competence" (Pawley 2009, xiv, xvi). Also Allen connects lexical bundles' presence to the language user's communicative competence: "the knowledge and use of a wide range of formulaic language helps [language learners] to achieve naturalness in language use" (2009, 106).

Some studies suggest that the presence of formulaic language facilitates the processing of text on the side of recipients: "those [collocations] which are divergent from native speaker norms, take longer to process when reading" (Allen 2009, 106)." It seems only natural that the same is true from the point of view of the speaker/writer: "Essentially, the frequent occurrence of these formulaic expressions is an aid both at the point of production and reception; on the one hand, it minimizes the decoding and encoding load of both parts in producing and receiving a fluent spoken and written discourse." (Rafiee, Tavakoli, and Amirian 2011, 138). The authors also mention that not everyone sees formulaic expressions simply as signs of fluent and native-like production. For example, according to Haswel (1991), a frequent use of these expressions might also be a mark of an "apprentice writer" (cited in Rafiee, Tavakoli, and Amirian 2011, 138).

Lexical bundles essentially restrict our freedom of expression by narrowing our choices of words in a given setting. In this context, Pawley mentions the "idiom principle" established by Sinclair who claims that there is a large pool of semi-preconstructed phrases we choose from rather than using chains of unrelated items (Pawley 2009, xii). In this respect, Chlumská raises an important question concerning typological differences between English and Czech. While the idiom principle might very well hold true for English (analytical language), the same necessarily does not have to be valid for a flecive language such as Czech (Chlumská 2016, 235). But at the same time, she stays confident that even though Sinclair's principle primarily applies to English, exploration of multi-word units (N-grams and PoS-grams) shows great potential for further research on the language of Czech translations (*ibid.*).

## 2.4.2 Typology of bundles

The basic categorization (one followed in this thesis) takes into account the number of words of the string. We distinguish 1-grams (unigrams) and 2-grams (bi-grams), which consist of one or two separate words respectively, 3-grams (trigrams) consisting of three, 4-grams consisting of four and so forth. However, the longer the chain, the lower the incidence and frequency which makes the bundles less useful and significant for an analysis. Gries states that most n-gram studies focus on structures consisting of three to four words above a particular threshold of occurrences per million<sup>22</sup> (Gries 2011, 2).

Lexical bundles can be further categorised in two ways: according to their primary function, and according to their structure. As to their function bundles can be 1. stance expressions (e.g. *I don't know, I thought it was*), 2. discourse organizers (e.g. *if you look at, what to do is*), and 3. referential expressions (e.g. *and this is the, a lot of the*) (Biber, Conrad, Cortes 2004, 384–387). In a recent study of lexical bundles in journalistic writing, Rafiee and Keihaniyan (2012) concluded that in this particular register it is the category of referential bundles that prevails.

The second taxonomy is based on the structural characteristics of the lexical bundles. This taxonomy can also be very complex; it always depends on which particular register is being scrutinised because lexical bundles vary across registers. Biber et al. (1999, 1001–1024), who focused on the classification of lexical bundles in speech and in writing, identified fourteen major structural types of bundles in conversation and twelve major categories in academic prose.

Depending on the type of query used, in a morphologically annotated corpus we can also search for Part-of-Speech-grams (PoS-grams) mentioned earlier. PoS-grams, defined as “strings of part of speech tags” (Stubbs 2007, 4), are very useful for identifying patterns of translated language (Baroni and Bernardini 2005, section 8; Chlumská 2016, 235). According to Brett and Pinna, who examined PoS-grams in a corpus of travel journalism, this type of query is very flexible and holds great potential for discovering sequences of words that would otherwise remain unnoticed (Pinna and Brett 2012, 53).

---

<sup>22</sup> According to Gries, the threshold is variable (eg. 10 or 15 occurrences per million) depending on the study at hand. In contrast with Biber who considers the n-gram frequency a necessary condition to call an n-gram a lexical bundle, Gries does not mention a specific threshold.

Even though the n-grams examined in this study are not classified according to the functional or structural criteria, nor is the PoS-gram query used, this typology is mentioned as a possibility for further research on this topic.

### **2.4.3 Usage and application**

Studies of formulaic language often focus on the presence of lexical bundles in learner writing in contrast with native language use, on features of specific discourses (especially academic) or on possible automatic genre identification, or, as in this study the possible differences between the language of translations and non-translations. Not infrequently, researchers also aim for pedagogical applications of their findings. When looking for these items, we are essentially looking for patterns in two respects: to what extent the language is patterned (comparing the overall frequencies of the occurring lexical bundles) and how much variation is present (categorising bundles and comparing different types of bundles).

Ellis et al. (2008) build upon previous psycholinguistic research which developed a theory that language users are particularly sensitive to the frequencies of occurrence of certain linguistic features and constructions and especially to their formulaic nature and collocability. They performed three experiments with two groups of participants: the first consisted of native speakers of English and the second consisted of international students studying English as their second language. The first experiment tested the subjects' ability to judge whether a presented sequence of words is grammatical or not.<sup>23</sup> During the second experiment, the participants were presented formulaic expressions on a computer screen and they were asked to read them as quickly as possible. The pause between the visual presentation and the beginning of their voiced response was measured. (Ibid., 384–385). During the third experiment, the participants were asked to read the last word of a given string as quickly as possible (the final element was either preceded by words or by a placeholder series of x's). Once again, the pause was measured for both types of strings (Ellis et al. 2008, 387). "These experiments demonstrate sensitivity to formulaicity in native fluent speakers, but we have yet to discover the psycholinguistic and corpus linguistic determinants of this sensitivity..." (Ellis et al. 2008, 375–376).

---

<sup>23</sup> The accuracy was greater than 96% (Ellis et al. 2008, 375-383).

The outcome of the experiments mentioned above would surely be consistent with the claim that native speakers can recognise translated texts from non-translated texts. Even though native speakers do not have the computing power to back their claims, this sensitivity to formulaic language and collocations might account for their high success rates in recognising translations from non-translations. Although an introspection is a useful tool for evaluating language data and it often provides valuable insight into language use and human cognition, intuitive judgement should ideally be accompanied by other, more rigorous methods. For example Stubbs (2004) proposes to pair intuitive claims of language users with corpus studies: “Corpus study does not reject intuition, but gives it a different role. Concordances focus intuition...” (Stubbs 2004, 109). Curiously enough, the author points out that the use of corpora for investigating recurrent patterns—although highly advantageous—has been mostly neglected (*ibid.*) Lexical bundles thus seem to be ideal means for spotting possible differences in the contrasted subcorpora.

Chlumská and Richterová (2014b) share this view and consider the analysis of n-gram frequency to be highly rewarding in terms of spotting differences between translated and non-translated Czech (266). For the purposes of this study, strings of 3-word fragments (3-grams) were chosen to be examined as some authors (Gries 2011, 2; Chlumská and Richterová 2014b, 266) consider 3-grams to be especially suitable for this kind of analysis. This analysis of 3-grams in section 3.4.2.1 will be followed by an analysis of 4-grams (section 3.4.2.2) to determine whether longer strings would be of use as well. Before the two analyses are presented, let me briefly comment on the two other features in which translated texts potentially differ from non-translated texts, namely lexical richness and average sentence length.

## 2.5 Lexical richness

Lexical richness is a feature accounting for the diversity of lexical means. Its degree is usually measured through a type-token ratio (TTR), which represents “the number of different types compared to the total number of tokens in the corpus...” (Kenny 1998, 51) where “types” represent all different words (particular word forms) and “tokens” all running words in a text (any instance of a particular word form). The types are divided by the number of tokens. The result of such calculation is subsequently multiplied by 100 which gives us a percentage (1–100) representing the degree of variation in a given corpus. The higher the percentage, the richer variety of language used. However, numbers close to 100% (each word form occurs only once) are not to be expected as it is unrealistic to find such a variety in a text of any length. The repetition of certain items which naturally occur frequently is unavoidable (and it is in no way detrimental to the quality of the text itself).

As mentioned earlier, the TTR is considered a useful method for distinguishing between translated and non-translated texts. According to our hypothesis, the translated component (the CET corpus) should exhibit lower lexical richness than the non-translated texts (the CRO corpus) in the same language. The relatively lower percentage (if confirmed) might point towards greater standardisation which manifests as the reduction of lexical variability (Cvrček and Chlumská 2015, 312).

A noteworthy drawback of this method is its sensitivity to the number of words of each text included in the corpus. This is caused by the asynchronous increase of types and tokens: “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, cited in Cvrček and Chlumská 2015, 315). In other words, texts of different lengths exhibit different degrees of repetitiveness, so including texts dramatically different in length might skew the results. Ideally, the individual texts should be of similar length in order to produce valid results or the TTR should be calculated separately for all of the individual texts. Other methods include calculating and comparing the TTR using a random sampling technique or using special software-based assessment (Koizumi and In’nami 2012, 523).

As will be seen in the following section 3.2.1 (Graph 4), the length of the texts included in the MCC at hand is variable, which might pose a problem. However, because no specialised tool for overcoming this obstacle was at our disposal and

because the text length variability is a feature common to both corpora (the CET corpus and the CRO corpus), this difference, as suggested by Lucie Chlumská<sup>24</sup>, will be disregarded. The TTR of each corpus will be calculated for all the texts as a whole.

The other issue is that the TTR seems to be influenced by the text type. As reported by Torruella and Capsada (2013, 453), for example poetry, as a rule, uses a richer variety of language than scientific prose and different authors might also influence the measure of lexical richness<sup>25</sup> independent of the text type (ibid.). As discussed earlier, both corpora comprise journalistic texts covering similar topics, so the first limitation does not apply here. The second limitation regarding the authorship of the texts is discussed in section 3.2.2.1.

---

<sup>24</sup> Mgr. Lucie Chlumská, Ph.D., personal communication, March 10, 2017.

<sup>25</sup> The authors use the synonym “lexical diversity” (Torruella and Capsada 2013, 453).

## 2.6 Average sentence length

The last examined feature, the average sentence length, is one of the frequently analysed textual features when it comes to translated and non-translated texts (see Baroni and Bernardini 2006, Baker 1996, Laviosa 1998a, Ilisei et al. 2010, Lee 2013, Giannossa 2016). According to our hypothesis, the average sentence length should be lower in the corpus of translated Czech than in the corpus of non-translated texts.

Once again, this parameter will be calculated for each corpus as a whole separately, this time simply by dividing the total number of words by the respective number of sentences. Even though measuring the average sentence length is not as problematic as establishing and comparing the TTR, Xiao et al. mention that this parameter is “sensitive to genres and may not be a reliable indicator of simplification” (2008, 24). The authors further propose that the observed differences in the average sentence length should be related to specific genres and languages in question (Ibid., 8). In this case, it is related to journalistic discourse and to translated (translations from English to Czech) and non-translated Czech.

Laviosa, who tested the reliability of this particular parameter for recognising translated and non-translated texts, advises caution in two respects: “I cautiously hypothesize, pending further evidence from a more varied and larger sample, that the average sentence length may be particularly sensitive, in the narrative subject domain, to the influence of different source languages, as well as the author's particular style.” (Laviosa 1998a, 8). Even though the first observation is not an issue here<sup>26</sup>, the influence of the author's style must not be overlooked not only for the calculation of the average sentence length but for all of the examined features. The authorship of the texts included in the corpora is discussed in detail in section 3.2.2.1 and as such it imposes major limitations on the usefulness of the MCC at hand.

---

<sup>26</sup> All of the translated texts in the CET corpus were translated from English.



## 3 ANALYTICAL PART

### 3.1 Sketch Engine

Sketch Engine is an online corpus analysis tool developed in 2003 by Adam Kilgarriff and Pavel Rychlý<sup>27</sup>. It allows searching and exploring collections of electronic texts. It is specially designed to allow observation of usual and unusual patterns of language and as such it serves linguists, translators, lexicographers, terminologists, but also students and language teachers. It offers about 400 ready-to-use corpora in more than 90 languages to be explored ('Sketch Engine' 2017).

First, it is a concordancer, i.e. a programme for exploration and retrieval of data from a corpus. It shows search results (concordances) by displaying the data in question in the format of a KWIC (key words in context): the keyword is highlighted in the middle and the immediate context to the right and to the left is provided on one line. Generally, concordancers allow searching for collocations, generate frequency lists of words or tags and allow exploration of typical combinations, multi-word phrases, synonyms or translations (when working with parallel corpora).

Second, Sketch engine also serves as a corpus manager, i.e. it offers the possibility of building and searching custom user created corpora (sometimes called DIY corpora, or opportunistic corpora), either directly from the web (Webcrawled corpora) or via uploading specific data regardless of its source. The latter function is the main reason why it was chosen for the purposes of this study. Unlike some other concordancers (e.g. WordSmith Tools), Sketch engine has an inbuilt tagger, which does automatic lemmatization and assigns part of speech tags<sup>28</sup> (including information about grammatical categories such as gender, case, number or stylistic value) to every token<sup>29</sup> in the corpus.

---

<sup>27</sup> According to McEnery and Hardie, Sketch Engine belongs to the fourth generation of corpus analysis tools, which began as webs allowing access to specific corpora but later grew into more generalised systems (2012a, 45). As such, these systems do not run on the user's computer but on a server which can be accessed through the user's web browser. (Ibid.).

<sup>28</sup> The tagset used is Majka.

<sup>29</sup> A token is the smallest unit in a corpus. There are always more tokens than words in a corpus, because tokens include punctuation.

## 3.2 Corpus design

For the purposes of this study, two journalistic corpora were compiled on the basis of texts from *Respekt*: one which includes texts originally written and published in Czech (the subcorpus is henceforth referred to as CRO – “Respekt: Czech Original texts”), and the other which includes translated Czech, namely translations from the English-language weekly newspaper *The Economist*, regularly published in “The Economist” section of *Respekt*. From now on, this subcorpus is referred to as CET, i.e. “Respekt: Translations into Czech from The Economist”). Texts published in this section were provided with the copyright statement that the English original can be found on “www.economist.com”—this served as the second criterion for classifying the texts as translations.

The conclusion that both corpora can be considered comparable was drawn upon the fact that all the Czech texts were published in the same magazine (of the same periodicity) and that all of the originals of the Czech translations come from a magazine covering similar topics.<sup>30</sup> This ensures the same text type—all of the texts belong to the journalistic discourse. As there are numerous journalistic genres and sub-genres, further classification is not specified, however upon closer inspection, the majority of the texts belong to a category of publicist writings which mix reporting on current events along with presenting the writer’s personal opinions and viewpoints.

### 3.2.1 Time span, corpus size and text length

The CET corpus contains all of the texts found in the category “The Economist” which were accessible through the basic subscription programme; there are 418 texts in total covering an eleven-year time-span (2007 to 2017). However, certain years (2011 to 2013) are underrepresented because the vast majority of the texts published during this period remained inaccessible through the electronic subscription. Similarly, the year 2017 contains only 10 articles because the data collection and the corpus compilation finished in early March 2017. Graph 1 below shows the number of texts in each corpus across the years.

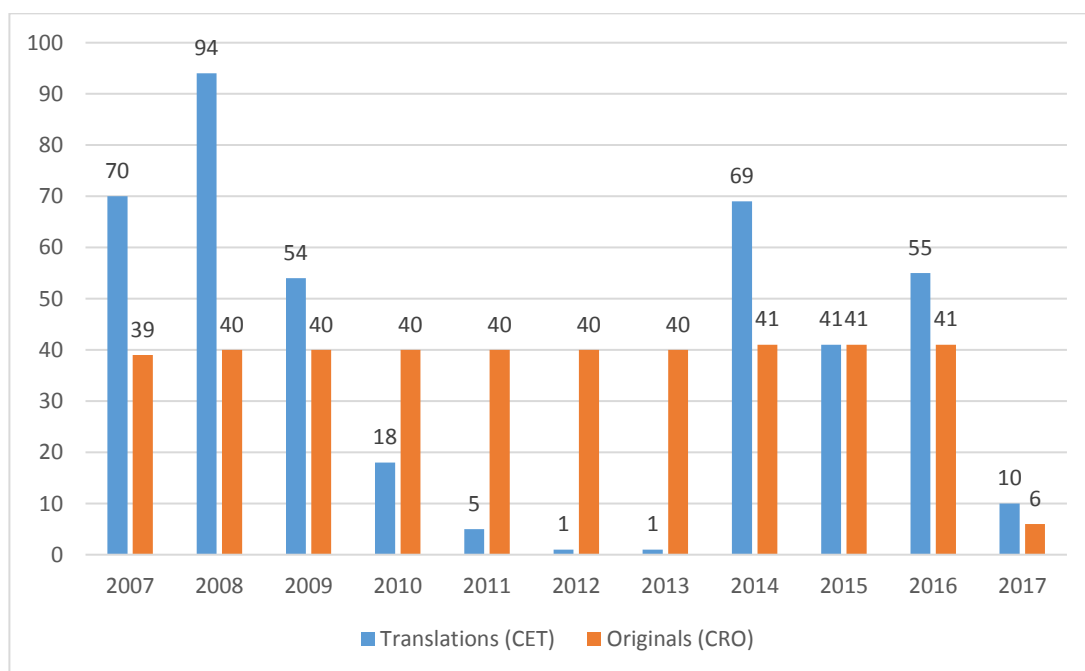
---

<sup>30</sup> Again, to quote from the official web page, it provides coverage of international news and deals with topics such as politics, business, finance, science and technology (‘The Economist: About Us’ 2017).

The CRO corpus contains texts from the same time span (2007 to 2017). As there were many Czech originals to choose from, some additional selection criteria were applied. The articles included in the CRO corpus were taken from different weekly issues spread across the whole year. The aim was to include a range of different themes and topics (both domestic and international) and texts of various lengths (to match the variable text length of the CET corpus). Certain genres such as interviews, weekly outlines and invitations were excluded due to their specific structure which would deviate from the CET texts.

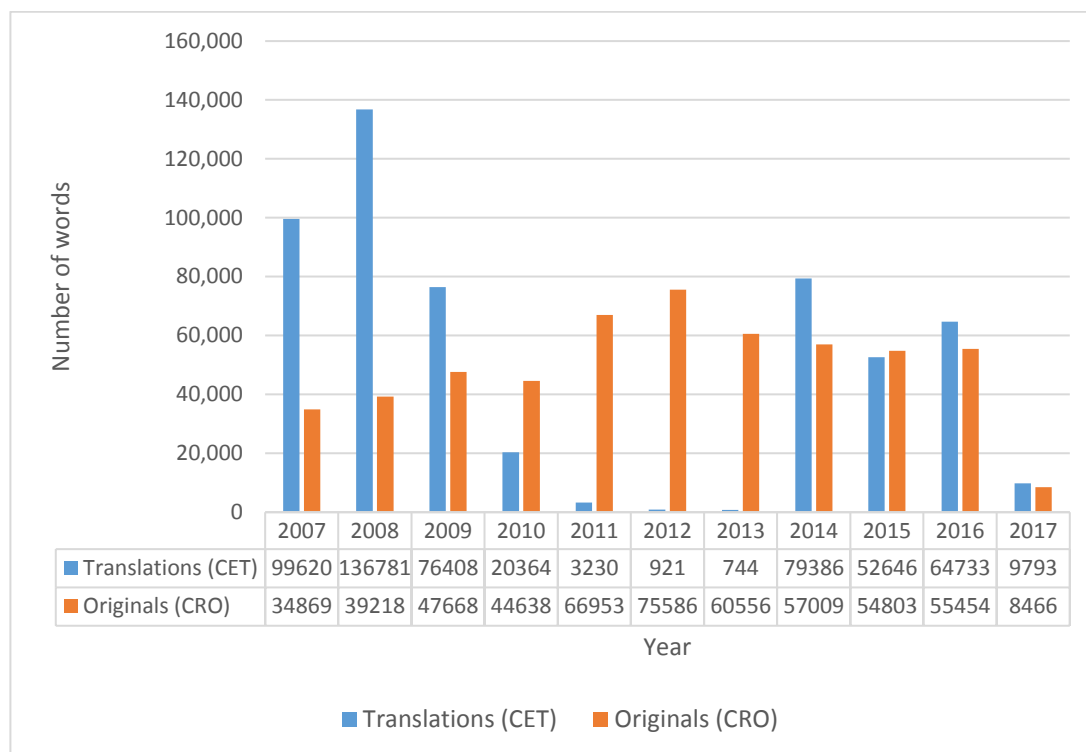
The headings and subheadings, if present, were preserved as this is a common feature of journalistic texts representative of this particular discourse in both corpora. However, the inserted tables, pictures, picture descriptions and additional excerpts linking to Twitter were excluded because they interrupted the natural flow of text and might skew the subsequent automated analysis. There are 408 articles in total, ca 40 articles per year, with the exception of the year 2017 which remains underrepresented (Graph 1).

**Graph 1: Number of texts in each corpus across the years**



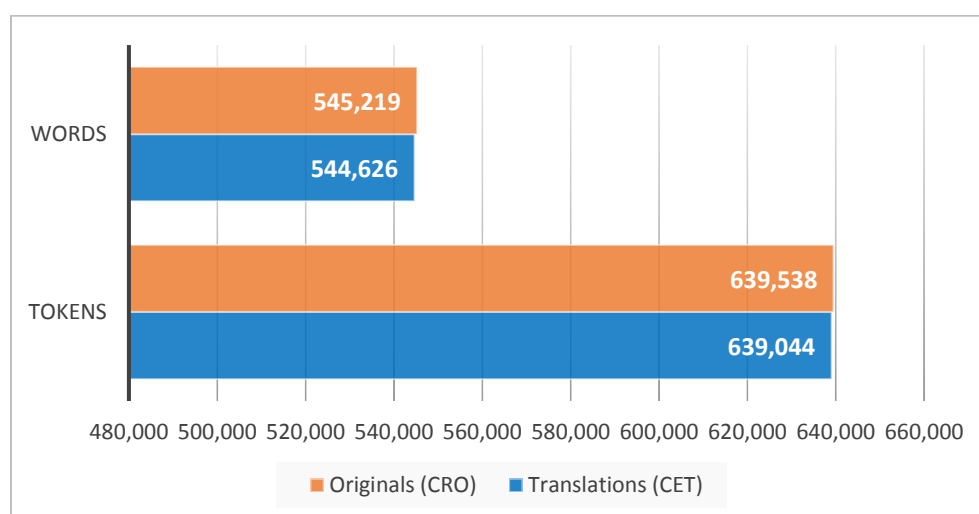
The number of words in each corpus across the years is shown in the next graph (Graph 2).

**Graph 2: Number of words in each corpus across the years**



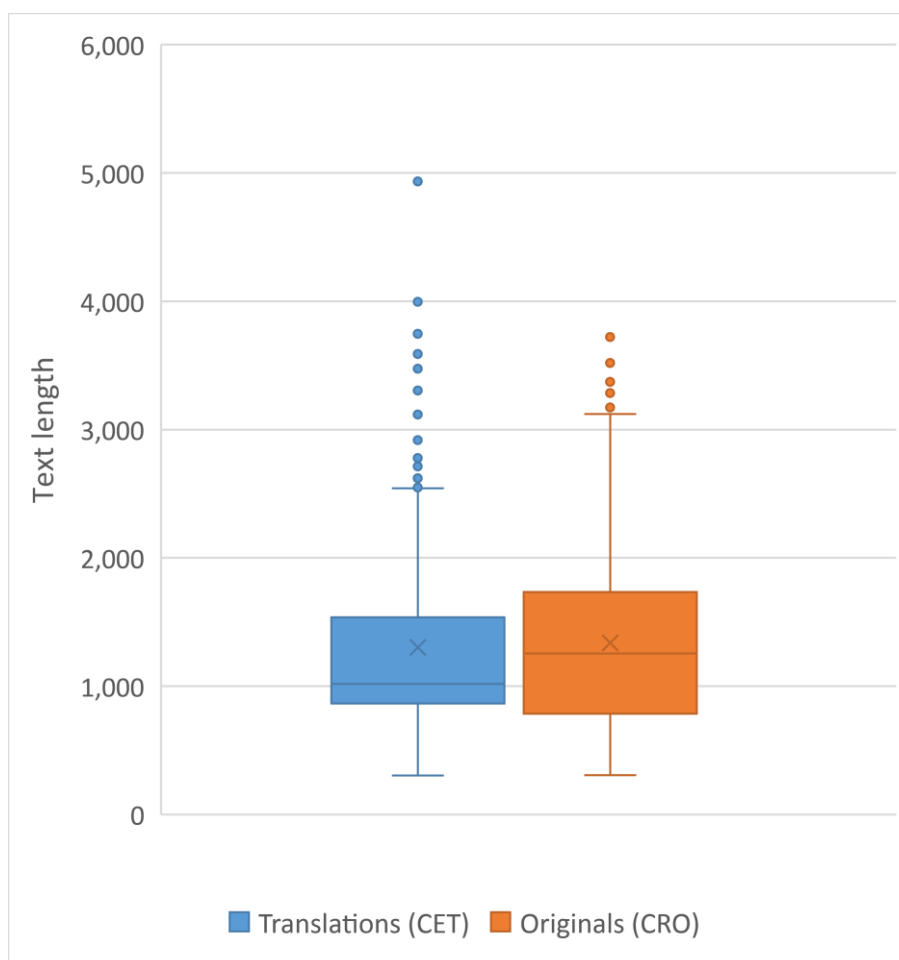
The main aim was to achieve a balanced representation with a comparable number of words/tokens in both corpora so that the CRO corpus and the CET corpus are comparable in terms of corpus size. The CRO corpus has 545,219 words and 639,538 tokens and the CET corpus has 544,626 words and 639,044 tokens (Graph 3).

**Graph 3: Number of words and tokens in each corpus**



Yet another important feature, which has to be taken into account, is the length of the individual texts included in both corpora. Ideally, both the translated and non-translated texts should be of similar length (as mentioned in Section 2.5). Unfortunately, because the texts translated from *the Economist* had variable length, the same variability was allowed for the texts included in the CRO corpus in order to reflect this feature of the CET corpus (see Graph 4). Both corpora contain texts of various lengths, namely between 307 and 3,775 words (the CRO corpus) and between 305 and 4,935 words (the CET corpus).

**Graph 4: Text length variability**

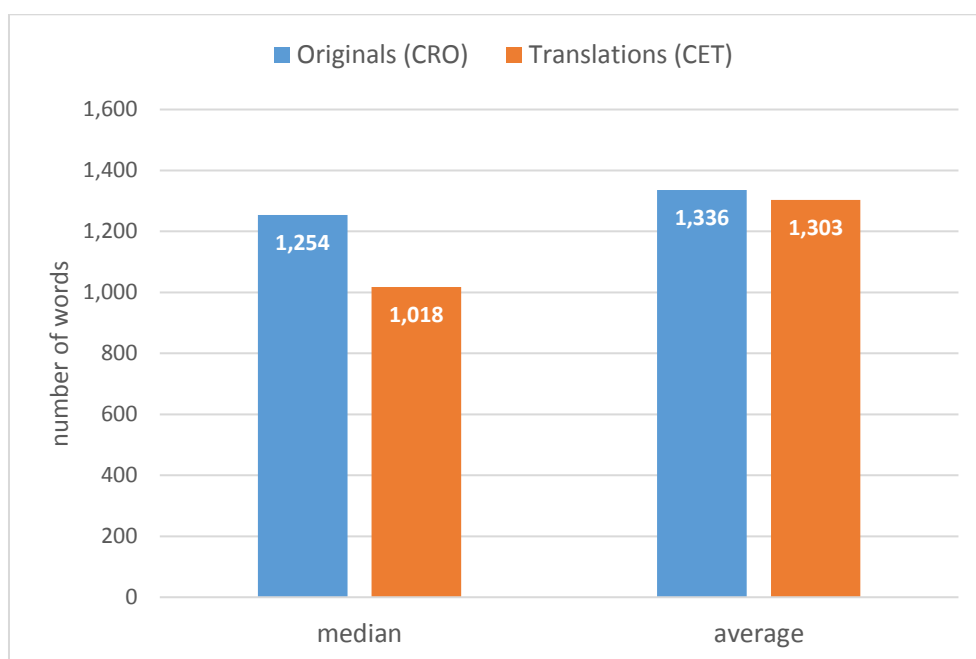


Graph 5 demonstrates the difference in the length of the texts included in both corpora in terms of average and median<sup>31</sup>. The average text length is 1,303 words in the CET corpus and 1,336 words in the CRO corpus. The median text length is 1,018 words in the CET corpus and 1,254 words in the CRO corpus. Both of these values

<sup>31</sup> The median (the middle score) is less affected by possible outliers in a given data set.

are lower for the CET corpus; the difference is more prominent when comparing the median (the middle value). Nevertheless, the differences are not so dramatic as to be considered a serious obstacle for the comparability of the two corpora, as mentioned earlier.

**Graph 5: Average/median text length**



## 3.2.2 Metadata

### 3.2.2.1 Authorship

According to some researchers (Meyer 2004, 53), information about the authorship of the texts included in a corpus (along with the characterization regarding the authors' background and professional status or education) is very important and should be included as metadata.

The name of the Czech translator (the author of the texts in the CET corpus) is not visibly marked; each translation is accompanied by a short phrase “přeloženo týdeníkem Respekt”<sup>32</sup>, which attributes the authorship collectively to the whole editorial staff. Upon further inspection, a short explanation was found (published in September 2015 as an answer to a reader who was interested in the name of the translator), stating that the journalists on staff predominantly translate the articles taken from foreign media. The Czech translations published in “The Economist”

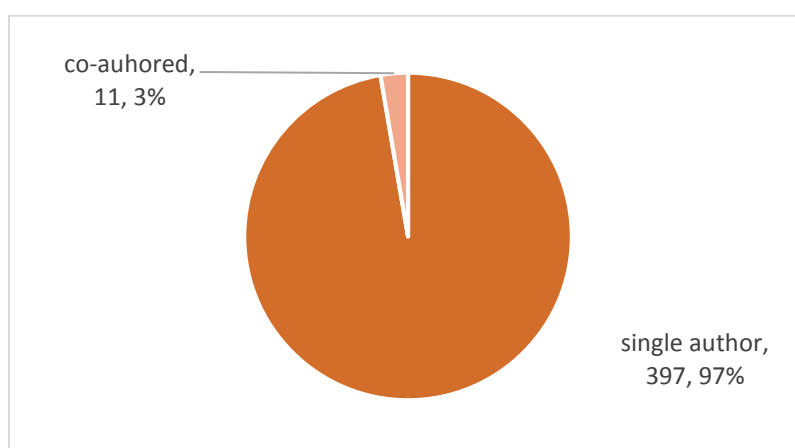
---

<sup>32</sup> “Translated by *Respekt*”

section of *Respekt* (translations from English to Czech) are attributed to Vladimír Fuksa ('Redakce, Dopisy.' 2015)<sup>33</sup>. Vladimír Fuksa is a Czech professional translator of films and TV shows (dubbing) and fiction. Further information about his career or education was not available on the publisher's website. It seems that the claim about the invisibility of news' translators (Valdeón 2014, 53; Bani 2006, 35) is well-founded. In this particular case, the translator's name is as a rule omitted and explicitly mentioned only when the reader enquires about it.

The authorship of the non-translated Czech texts in the CRO corpus was much easier to ascertain as the name of each author was provided. All in all, there are 106 different authors, including 11 instances of collaboration (two people marked as co-authors). Texts written by a single author represent 97% of the corpus while co-authored texts account for only 3% of the articles (Graph 6)<sup>34</sup>.

**Graph 6: Number of authors per article (the CRO corpus)**

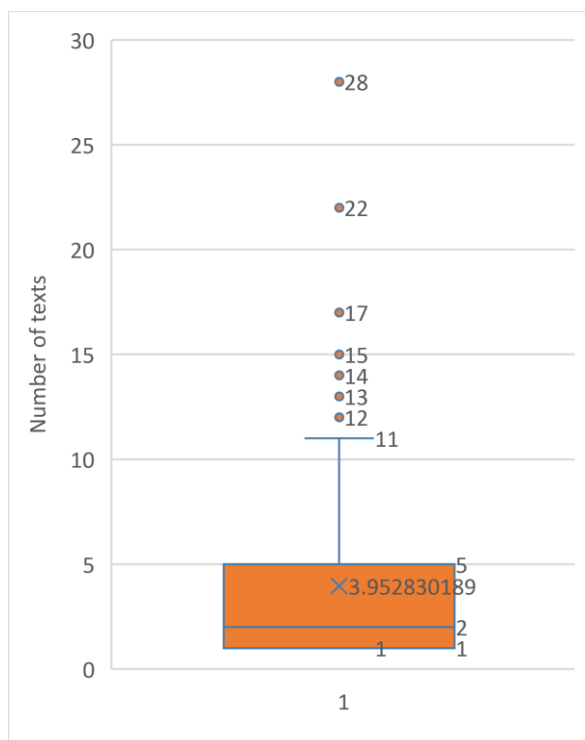


The majority of the authors wrote one or two texts each (65 distinct authors), many contributed with three to fifteen texts (37 distinct authors) and only four wrote more than sixteen pieces each (see Graph 7).

<sup>33</sup> "Cizojazyčné články z valné části překládají redaktoři Respektu – text *Návrat z pekla* připravil Tomáš Lindner, komentáře od Fareeda Zakarii zpracovávají Jiří Sobota nebo Martin M. Šimečka. Články z *The Economist* pro nás překládá pan Vladimír Fuksa." ('Redakce, Dopisy.' 2015)

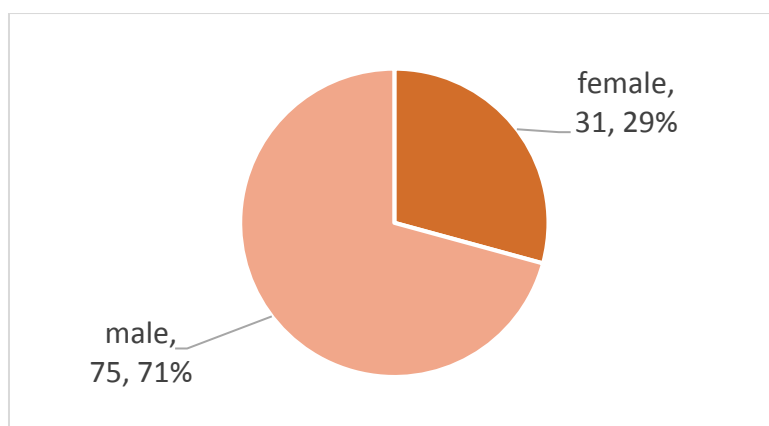
<sup>34</sup> These figures (graphs 6–9) do not take into consideration the length of the individual articles.

**Graph 7: Number of articles per author (the CRO corpus)**



As regarding the gender of the authors, there were 31 women (29%) and 75 (71%) men (Graph 8).

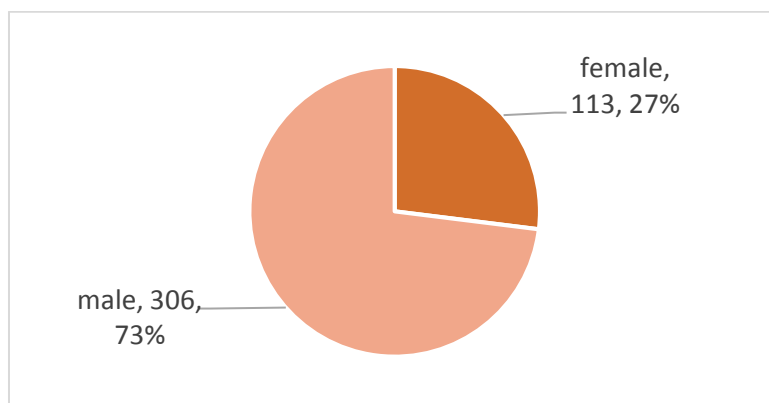
**Graph 8: Authors by gender (the CRO corpus)**



113 articles (27%) were written (or co-written) by women and 306 articles (73%) were written (or co-written) by men (Graph 9).



**Graph 9: Number of articles according to gender (the CRO corpus)**



The authors' background or education is unknown. Some of them are journalists from *Respekt* (usually those who contributed repeatedly), others are long-term external reporters or correspondents, and the rest of the occasional contributors are experts, prominent members of the society or for example travellers. Because additional information regarding the authors of the Czech originals is not available, further claims cannot be made.

Unfortunately, the authorship of the original English texts<sup>35</sup> published in *The Economist* is unknown; the publisher of *The Economist* does not include the names of the authors. The official website states that the authors' names are omitted on purpose, to promote a sense of unity, a so-called "collective voice" of the magazine ('The Economist: About Us' 2017). It further states that the anonymity of the journalists is justified by the fact that the members of the editorial staff often meet and discuss their writing and cooperate or that some of the articles undergo heavy editing. "The main reason for anonymity, however, is a belief that what is written is more important than who writes it." (Ibid.).

As suggested earlier, the fact that there is not a comparable number of different authors for both corpora and that the name(s) of the author(s) of the source texts remain(s) unknown, restricts the usability and representability of the corpus at hand.

### **3.2.2.2 Date of publication**

When discussing the properties of a corpus of translated texts, yet another important feature is the time between publishing the original and publishing the translated version. In this respect, it is must be noted that journalistic texts are very

---

<sup>35</sup> These are the source texts which are not included in either of the corpora at hand.

time sensitive. Translation of news or articles on current affairs should be as quick as possible in order to retain relevance and topicality of the reported issue in accordance with news values<sup>36</sup> (Bielsa and Bassnett 2009, 115).

Ten random samples from the CET corpus were chosen to compare the publishing date of the original and its translation. As Table 1 seeks to illustrate, as a rule, the translations were published quite soon after the publication of the original, in some cases even the same day (2 instances) and very often during the course of the week (7 instances). However, there are still exceptions to the rule (text no. 9), where the delay was 39 days.

**Table 1: Comparing the publication dates (original vs. translation)**

No.	Text status	Date of publication	Title	Delay of translation in days
1	original	6.9.2007	In search of the good company	1
	translation	7.9.2007	Hledání dobré firmy	
2	original	19.6.2008	Another silicon valley?	9
	translation	28.6.2008	Další Silicon Valley?	
3	original	26.3.2009	The nuts and bolts come apart	3
	translation	29.3.2009	Kolo se nám polámalo	
4	original	18.2.2010	Let the Greeks ruin themselves	3
	translation	21.2.2010	Nechte je, ať se zničí	
5	original	12.5.2011	Thrice blessed	3
	translation	15.5.2011	Trojí požehnání	
6	original	16.10.2014	Bolts from the blue	38
	translation	23.11.2014	Blesky z čistého nebe	
7	original	1.11.2014	Good voters, not such good guys	1
	translation	2.11.2014	Dobří voliči a zlobiví hoši	
8	original	5.2.2015	Follow the money	3
	translation	8.2.2015	Jděte po penězích	
9	original	19.3.2016	A hollow superpower	0
	translation	19.3.2016	Dutá supervelmoc	
10	original	11.2.2017	The multi-billion-euro exit charge that could sink Brexit talks	0
	translation	11.2.2017	Účet za brexit <sup>37</sup>	

<sup>36</sup> News values are criteria or rules which determine whether a story, an event or a fact is newsworthy. Bednarek and Caple (2012, 41) distinguish 9 categories of news values: negativity, timeliness, proximity, prominence, consonance, impact, novelty, superlativeness and personalization. As regarding timeliness, they claim that “[m]ore recent events are often more newsworthy” and thus more likely to be registered as news (Ibid. 42).

<sup>37</sup> This table, apart from allowing us to compare the date of publication, provides an opportunity to compare the translation of the titles. Even though this paper is not concerned with the parameter of faithfulness in translation (nor do we have the original English texts at our disposal), based

### 3.2.3 Summary of key characteristics

The aim of this section is to provide a short summary of the key characteristics of the corpora introduced in this section. As stated at the beginning, both corpora compiled for the purposes of this study are monolingual. Both of them are an instance of a sample corpus which means they are finite in terms of size and aim to provide a “static snap-shot” of the language in question. However, it might also be advantageous to use the label “an opportunistic corpus,” which describes corpora which do not fit the traditional categories of a monitor or a snap-shot corpus. “These corpora make no pretension to adhere to a rigorous sampling frame, nor do they aspire to deal with issues of skew by the collection of an ever-larger body of data, as monitor corpora may.” (McEnery and Hardie 2012a, 11).<sup>38</sup>

As regarding the classification on the synchronic-diachronic continuum, it is true that the corpus at hand contains texts published over a period of eleven years (2007–2017). Nevertheless, the synchronic label seems more fitting because the time span is still rather short to be considered synchronic and language development over time is not the concern of this study. Finally, it is a specialised corpus covering a specific domain and genre and it consists of written texts.

To sum up, both corpora are considered to be representative and truly comparable on the basis of the similar corpus size (see Graph 3), the same text type and genre. Furthermore, they cover similar topics and they are intended for the same target readers. While the individual corpora do not comprise texts of the same length, the text-length variability is a feature common to both of them (see Graph 4).

---

on these ten random samples, we can conclude that the translator usually adhered to the original sense of the titles. Titles 1, 2, 8 and 9 follow the original very faithfully, titles 4 and 7 show some minor alterations. Title 3 and 6 provide the Czech equivalent to the respective idiomatic phrase of the original and only tile 10 underwent major changes as a result of shortening and generalization.

<sup>38</sup> An opportunistic corpus is a kind of corpus which makes use of all the available data needed for a specific task. In this respect, Halliday et al. mention, that if we embrace the idea that every corpus is essentially imbalanced, we will be free to approach the issue of representability from a new angle which allows us to utilise all kind of corpora, especially the opportunistic ones (2004, 120).

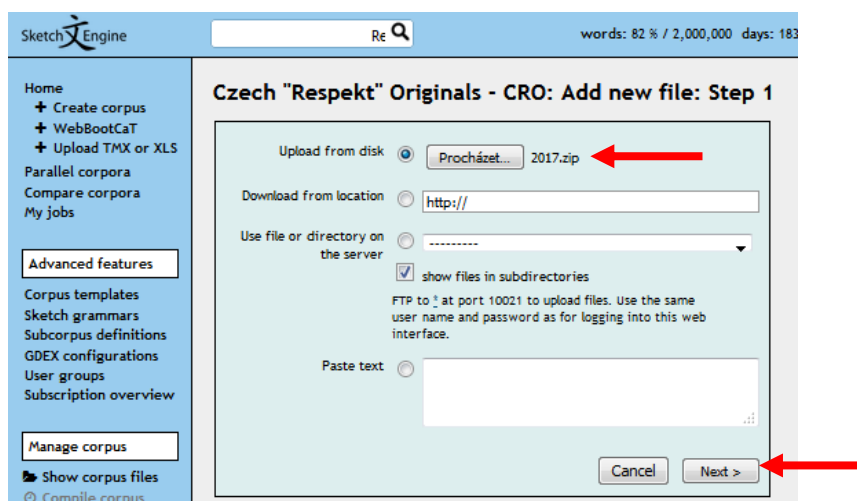
### 3.3 Corpus compilation

Both corpora were created using the web interface of Sketch Engine. The Czech language was entered manually (Figure 1) and the data in zipped archives labelled according to the year of publication were uploaded using the option “upload from disk” (Figure 2). Subsequently, each archive was “expanded” to preserve the metadata in the file name, which reflected the title of the published article (Figure 3).

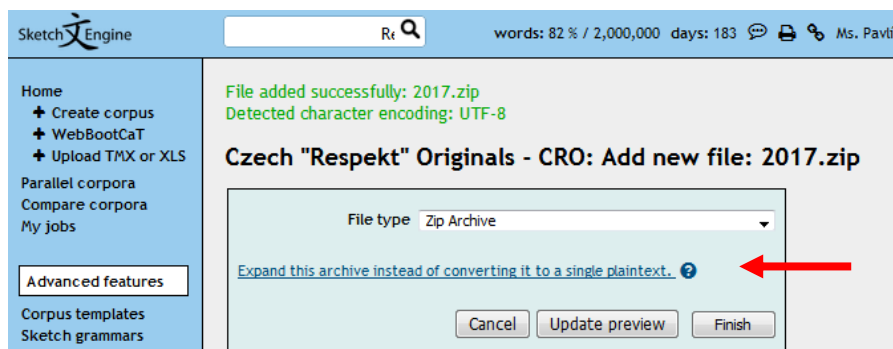
*Figure 1: Creating the corpus*



*Figure 2: Uploading the corpus*

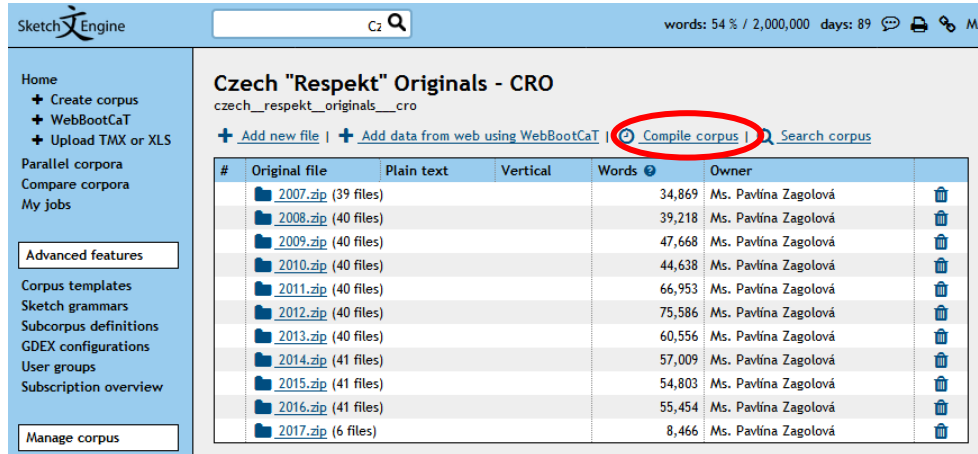


*Figure 3: Expansion of the archives*



Upon completing the upload, the complete corpus was set to be automatically compiled and tagged (Figure 4 and Figure 5). The same procedure was then used for the creation of the CET corpus.

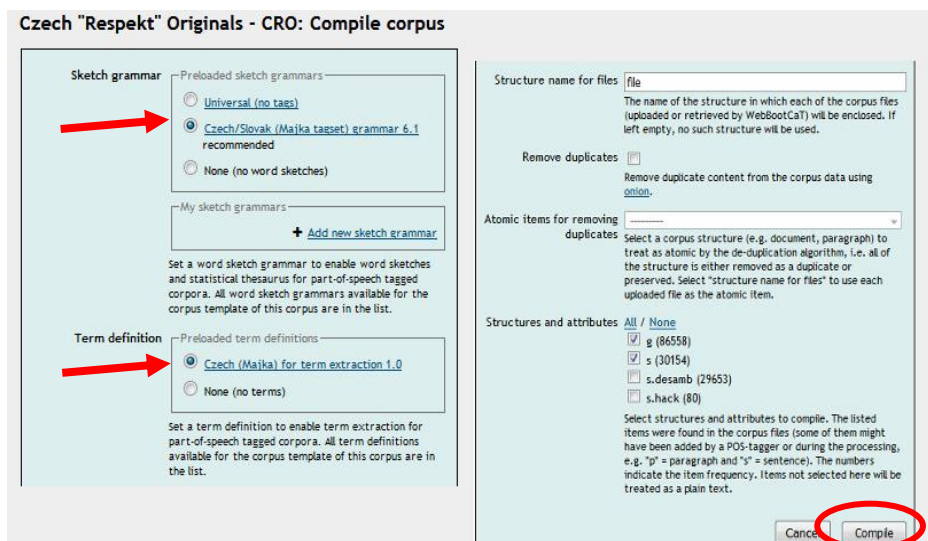
**Figure 4: Compiling the corpus**



The screenshot shows the Sketch Engine interface for a corpus named "Czech 'Respekt' Originals - CRO". The top navigation bar includes "Home", "Create corpus", "WebBootCaT", and "Upload TMX or XLS". The main area displays a table of files with columns for "#", "Original file", "Plain text", "Vertical", "Words", and "Owner". The "Compile corpus" button is highlighted with a red circle.

#	Original file	Plain text	Vertical	Words	Owner
	.2007.zip (39 files)			34,869	Ms. Pavlína Zagolová
	.2008.zip (40 files)			39,218	Ms. Pavlína Zagolová
	.2009.zip (40 files)			47,668	Ms. Pavlína Zagolová
	.2010.zip (40 files)			44,638	Ms. Pavlína Zagolová
	.2011.zip (40 files)			66,953	Ms. Pavlína Zagolová
	.2012.zip (40 files)			75,586	Ms. Pavlína Zagolová
	.2013.zip (40 files)			60,556	Ms. Pavlína Zagolová
	.2014.zip (41 files)			57,009	Ms. Pavlína Zagolová
	.2015.zip (41 files)			54,803	Ms. Pavlína Zagolová
	.2016.zip (41 files)			55,454	Ms. Pavlína Zagolová
	.2017.zip (6 files)			8,466	Ms. Pavlína Zagolová

**Figure 5: Tagging the corpus**



The screenshot shows the "Compile corpus" dialog box. It has several sections: "Sketch grammar" with radio buttons for "Universal (no tags)", "Czech/Slovak (Maika dataset) grammar 6.1 recommended", and "None (no word sketches)"; "Term definition" with radio buttons for "Czech (Maika) for term extraction 1.0" and "None (no terms)"; "Structure name for files" with a text input field containing "file"; "Remove duplicates" with a checkbox; "Atomic items for removing duplicates" with a dropdown menu; and "Structures and attributes" with a list of items and checkboxes. The "Compile" button is highlighted with a red circle.

## 3.4 Data analysis

### 3.4.1 Methods and terminology

#### 3.4.1.1 N-gram extraction criteria and terminology

Due to the fact that both corpora at hand are rather modest in size (ca 545 thousand words each), the threshold of minimum occurrence was set to 5 occurrences in the whole corpus (for both 3-grams and 4-grams). One more criterion—analogueous to Biber’s second criterion for defining lexical bundles—was additionally applied to all of the analysed 4-grams and to the 100 most frequent analysed 3-grams. This additional criterion ensures that the n-grams in question are spread across at least 5 different texts in the respective corpus. In this paper, such n-grams are labelled as “filtered”.

When presenting the raw frequencies of different n-grams across the two corpora, the terminology type vs. token will be adopted. In this context, a type is the occurrence of a particular n-gram regardless of its frequency in the corpus. For example, the 3-gram “bez ohledu na” is one n-gram type, even though there are 27 instances of this particular string of words in the CRO corpus. Tokens, on the other hand, represent all the individual occurrences of one n-gram type: for example, there are 27 tokens of the 3-gram “bez ohledu na” in the CRO corpus and 41 tokens in the CET corpus.

#### 3.4.1.2 Statistical significance

Unless stated otherwise, all results described as “statistically significant” were tested by running a test for statistical significance by *Corpus Frequency Wizard tool* (Baroni and Evert 2017).<sup>39</sup> This is an online calculator designed by Marco Baroni and Stefan Evert for the SIGIL<sup>40</sup> project. It allows testing for statistical significance when comparing the frequency of two samples across two different data sets (see Figure 6).

---

<sup>39</sup> This tool is available at <http://sigil.collocations.de/wizard.html>

<sup>40</sup> SIGIL stands for “Statistical Inference: A Gentle Introduction for Linguists”

**Figure 6: Frequency comparison of two samples**

**SIGIL: Corpus Frequency Test Wizard** [back to main page](#)

This site provides some online utilities for the project **Statistical Inference: A Gentle Introduction for Linguists (SIGIL)** by [Marco Baroni](#) and [Stefan Evert](#). The main SIGIL homepage can be found at [purl.org/stefan.evert/SIGIL](http://purl.org/stefan.evert/SIGIL).

**One sample: frequency estimate (confidence interval)** [back to top](#)

Frequency count	Sample size
<input type="text"/>	<input type="text"/>

95% confidence interval  
in automatic format  
with 4 significant digits

extrapolate to  items

---

**Two samples: frequency comparison** [back to top](#)

	Frequency count	Sample size
Sample 1	<input type="text"/>	<input type="text"/>
Sample 2	<input type="text"/>	<input type="text"/>

95% confidence interval  
in automatic format  
with 4 significant digits

The “frequency count” was represented by the individual counts (the respective number of types or tokens) as supplied by the Sketch Engine. The “sample size” for attaining the normalised frequency for each of the corpora was calculated for each corpus separately. Although it is also possible to use the number of tokens as the referential sample size (Bardoel 2012, 27), the total number of n-grams (3-grams or 4-grams respectively) was used for the sake of accuracy as the “sample size”. The formula “number of tokens” - (n-1) was used to calculate the total number of n-grams for each corpus as suggested by the Sketch Engine support.<sup>41</sup> See table Table 2 for an overview of the total number of n-grams for each corpus.

**Table 2: Calculating the “sample size” (the total number of n-grams)**

	Originals (CRO)	Translations (CET)	Formula
number of tokens	639,538	639,044	-
number of 3-grams	639,536	639,042	= number of tokens - (3-1)
number of 4-grams	639,535	639,041	= number of tokens - (4-1)

<sup>41</sup> Ondřej Matuška, personal communication, March 10, 2017.

Figure 7 shows the calculation of the frequency comparison of a 3-gram “na celém světě”. There were 26 tokens of this 3-gram in the CRO corpus and 68 tokens in the CET corpus (both entered as “frequency count”). The respective “sample size” (the total number of 3-grams in each corpus) was entered into the second column.

*Figure 7: Testing for statistical significance (3-gram "na celém světě"): input*

#### Two samples: frequency comparison

	Frequency count	Sample size		
Sample 1	<input type="text" value="26"/>	<input type="text" value="639536"/>	<input type="button" value="Clear fields"/>	95% confidence interval
Sample 2	<input type="text" value="68"/>	<input type="text" value="639042"/>	<input type="button" value="Calculate"/>	in automatic format
				with 4 significant digits

The wizard works with two kinds of statistical tests, chi-square and log-likelihood test, and it automatically chooses a test which is considered to be more accurate for the data entered (Hoffmann et al. 2008, 84–85). The minimum level of significance for both tests is 95% ( $p < .05$ ). Figure 8 shows the result of the frequency test for the 3-gram “na celém světě”: the difference is significant at  $p < .001$ .

*Figure 8: Testing for statistical significance (3-gram "na celém světě"): result*

#### Corpus Frequency Test: Two Samples

```

Test result:  $\chi^2 = 17.91599$  ***
             difference is significant at  $p < .001$  (crit. 10.82757)
Confidence interval: [-98 pmw ... -35 pmw]
                    (two-sided, 95% confidence, Sample 2 > Sample 1)
Sample 1 data: 26 out of 639,536 = 40.65 pmw (relative frequency)
Sample 2 data: 68 out of 639,042 = 106.4 pmw (relative frequency)

```

A similar procedure was followed for the calculation of the statistical significance of the differences in the lexical richness and the average sentence length. See the appropriate sections (3.4.3 and 3.4.4) for a more detailed description.



## 3.4.2 Lexical bundles

### 3.4.2.1 3-grams

A small pilot study<sup>42</sup> was conducted beforehand to ensure that this query produces valid data and that the identified 3-grams are frequent enough to be useful for further processing. Even though the corpora at hand are rather small (and the corpus size significantly influences the number of extracted items and consequently its usefulness in terms of statistical significance), 3-grams proved to be frequent enough to supply enough data for the analysis.

#### 3.4.2.1.1 Corpus query for extraction

The most frequent 3-grams were obtained by entering the particular corpus and clicking the option “Word list”. The search attribute “word (lowercase)” was chosen so that the search algorithm would not differentiate between strings at the beginning of sentences and in the middle. In other words, this option ensured a case insensitive search which did not take into account the n-gram’s position in a sentence. N-value from 3 to 3 was set to search for 3-grams and a filter for the minimum frequency of occurrence was set to 5 instances in the whole corpus<sup>43</sup> (see Figure 9).

**Figure 9: 3-gram search query**

The screenshot displays the 'Word list options' interface. On the left is a navigation menu with 'Word list' highlighted. The main area contains the following settings:

- Subcorpus: create new
- Search attribute: word (lowercase)
- use n-grams: checked, Value of n: from 3 to 3
- hide/nest sub-n-grams: unchecked
- Filter word list by: Regular expression: (empty)
- Minimum frequency: 5
- Maximum frequency: 0 (0 = no maximum frequency)
- Whitelist: Procházet... Soubor nevybrán. Clear
- Blacklist: Procházet... Soubor nevybrán. Clear format
- Include non-words: unchecked
- Output options: Frequency figures: Hit counts (selected), Document counts, ARF; Output type: Simple (selected), Keywords
- Reference (sub)corpus: Czech Web 2012 (czTenTen12 v8) (whole corpus)
- Prefer: rare words, common words 1
- Change output attribute(s): (empty)

Red arrows point to the search attribute, the n-gram settings, and the minimum frequency field. A red circle highlights the 'Make word list' button at the bottom.

<sup>42</sup> First, the search query (see Figure 9) was tested and the total number of 3-grams along with the individual frequencies for each unique 3-gram was shortly examined. Further, we compared the most frequent 3-grams in both corpora to make sure that there were some matches which could provide the basis for a more detailed analysis.

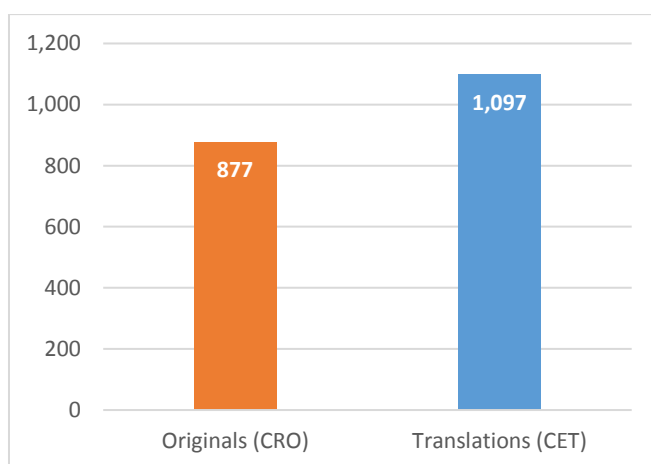
<sup>43</sup> The same procedure was used for both corpora (the CRO corpus and the CET corpus).

### 3.4.2.1.2 The overall number and frequencies

Firstly, the overall frequency of the recurring lexical patterns was examined as suggested by Baker<sup>44</sup>. As mentioned earlier, all of the n-grams presented in this section satisfy the established criterion of the minimum frequency of occurrence (at least 5 instances per corpus). However, the second proposed criterion which states that these units must be spread at least across 5 different texts could not be fulfilled because the high incidence of these patterns did not allow for manual sorting and Sketch Engine itself does not offer automatic filtering of n-grams according to the “document count” criterion. The results presented below (Graph 10 and Graph 11) thus might be influenced by the occurrence of patterns limited to the individual texts. It is therefore possible that, for example, one n-gram type might be present as 5 tokens but only in one text.

There were 1,097 3-gram types in the CRO corpus and 877 3-gram types in the CET corpus (Graph 10). This difference is statistically significant at  $p < .001$ .

**Graph 10: Absolute frequency of 3-gram types**

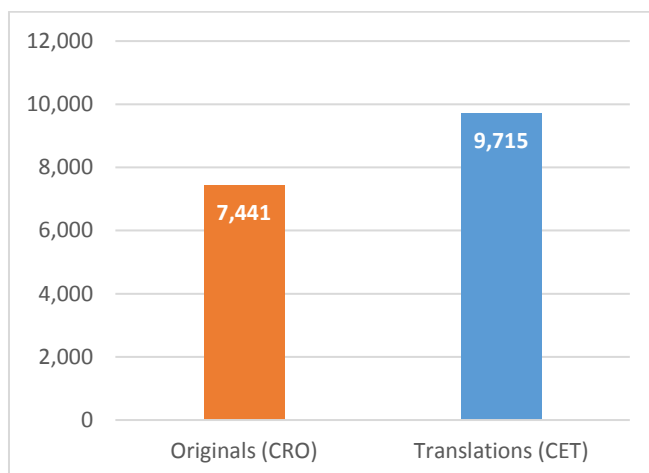


The subsequent analysis revealed 7,441 3-gram tokens in the CRO corpus and 9,715 3-gram tokens in the CET corpus (see Graph 11). This difference also proved to be statistically significant at  $p < .001$ .

---

<sup>44</sup> “As a first step, it seems reasonable to establish whether there is a noticeable difference between the two corpora in terms of the overall number and frequencies of the lexical patterns we have chosen to focus on.” Baker (2004, 175).

**Graph 11: Absolute frequency of 3-gram tokens**



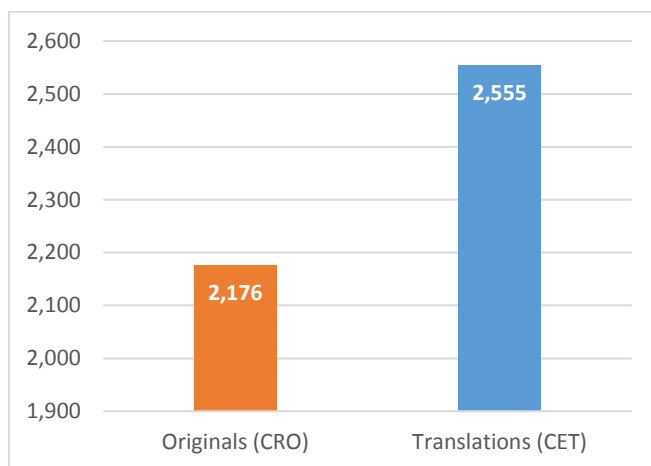
Both graphs show noticeable differences between the usage of unique 3-grams in the two corpora and attest that 3-gram patterns are overall used more frequently in the CET corpus than in the CRO corpus. Both of these differences proved to be statistically significant.

#### 3.4.2.1.3 100 most frequent 3-grams

As a next step, 100 of the most frequent 3-grams (3-gram types) were extracted from both corpora and an analysis in terms of frequency of occurrence and possible overrepresentation or underrepresentation was performed. This time, apart from the criterion of the minimum frequency of occurrence, all of the examined 3-grams were checked for the second criterion, that is to say, all of the filtered 3-grams in both corpora were spread across at least 5 different articles so no further filtering was required.

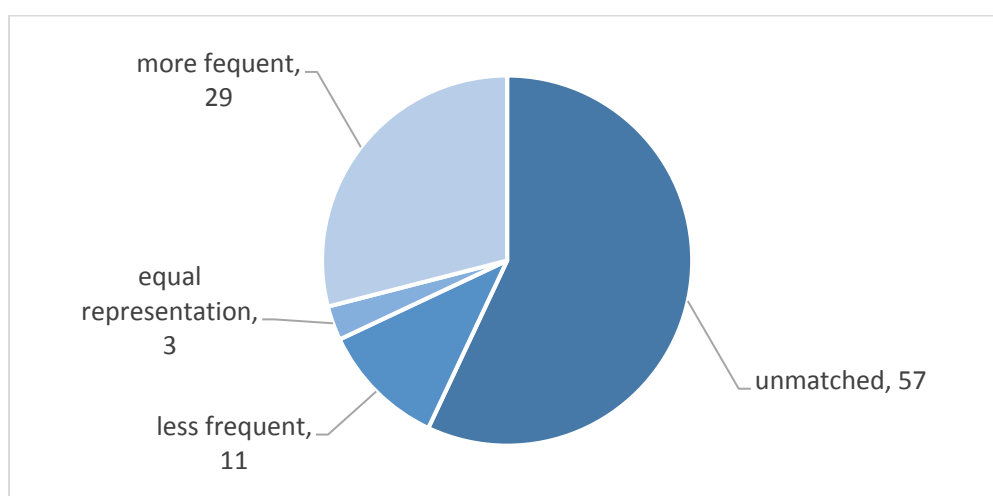
Once again, the absolute frequency of these top 3-grams was inspected. The most frequent one hundred patterns (types) in the CRO corpus appeared 2,176 times while in the CET corpus there were 2,555 tokens (see Graph 12), which is a statistically significant difference at  $p < .001$ .

**Graph 12: Absolute frequency of 100 top 3-grams (filtered)**



In the subsequent analysis, the individual 3-grams were examined in detail in order to establish, whether there are correspondences between the two lists. Among the top 100 most frequent 3-grams, 57 3-grams from the CRO corpus did not have their match in the CET corpus and vice versa. This does not mean that these 3-grams were truly corpus unique; some of them might have their match in the second corpus but its low frequency of occurrence and lower rank might have caused that they were not featured on the list of 100 most frequent 3-grams. The rest of the 3-grams had its match in the second corpus. Out of these 43 matches, 29 expressions were more frequent in the CET corpus, 11 were less frequent and 3 had equal representation when compared with the CRO corpus (see Graph 13).

**Graph 13: Relative frequency of 100 top 3-grams (filtered) in the CET corpus compared to the CRO corpus**



To summarise, the majority of the 3-grams which were featured in both corpora (29 out of 43 matches; that is 67%) were more frequent in the corpus of Czech translations. However, to ensure that this initial impression of overrepresentation was well-founded and such a claim legitimate, a test for statistical significance was administered. See Table 3 where statistically significant instances of overrepresentation (or underrepresentation) across the two corpora are indicated in the last column and highlighted in green.

**Table 3: Comparison of the frequency distribution of the matching 3-grams**

No.	3-gram	Originals (CRO)	Translations (CET)	Statistically significant	p-value
1.	na rozdíl od	108	<b>117</b>	no	
2.	<b>po celém světě</b>	26	<b>68</b>	<b>yes</b>	<b>p &lt; .001</b>
3.	od té doby	47	<b>63</b>	no	
4.	že by se	39	<b>59</b>	no	
5.	v posledních letech	48	<b>58</b>	no	
6.	<b>vzhledem k tomu</b>	17	<b>57</b>	<b>yes</b>	<b>p &lt; .001</b>
7.	v poslední době	39	<b>50</b>	no	
8.	<b>ve srovnání s</b>	23	<b>48</b>	<b>yes</b>	<b>p &lt; .01</b>
9.	<b>pokud jde o</b>	20	<b>46</b>	<b>yes</b>	<b>p &lt; .01</b>
10.	<b>v devadesátých letech</b>	22	<b>43</b>	<b>yes</b>	<b>p &lt; .05</b>
11.	bez ohledu na	27	<b>41</b>	no	
12.	ve spojených státech	<b>38</b>	34	no	
13.	<b>spočívá v tom</b>	15	<b>30</b>	<b>yes</b>	<b>p &lt; .05</b>
14.	i když se	17	<b>30</b>	no	
15.	z nich je	17	<b>29</b>	no	
16.	o více než	18	<b>28</b>	no	
17.	<b>na první pohled</b>	<b>52</b>	28	<b>yes</b>	<b>p &lt; .05</b>
18.	a v roce	16	<b>28</b>	no	
19.	většina z nich	15	<b>27</b>	no	
20.	před deseti lety	26	26	(match)	
21.	se jedná o	15	<b>25</b>	no	
22.	a to je	18	<b>25</b>	no	
23.	v té době	<b>32</b>	24	no	
24.	čím dál víc	18	<b>23</b>	no	
25.	pokud by se	20	<b>22</b>	no	
26.	do značné míry	17	<b>22</b>	no	
27.	v osmdesátých letech	16	<b>21</b>	no	
28.	na druhou stranu	<b>35</b>	21	no	
29.	je v tom	21	21	(match)	
30.	z nich se	17	<b>20</b>	no	
31.	v tomto ohledu	13	<b>20</b>	no	
32.	<b>před dvěma lety</b>	<b>34</b>	20	no	
33.	že se v	19	19	(match)	
34.	v tomto případě	<b>31</b>	18	no	
35.	<b>v tuto chvíli</b>	<b>38</b>	17	<b>yes</b>	<b>p &lt; .01</b>
36.	let minulého století	14	<b>17</b>	no	
37.	je jedním z	16	<b>17</b>	no	
38.	<b>do té doby</b>	<b>38</b>	17	<b>yes</b>	<b>p &lt; .01</b>
39.	že je to	<b>28</b>	16	no	
40.	z velké části	15	<b>16</b>	no	
41.	tváří v tvář	<b>18</b>	16	no	
42.	že se na	13	<b>15</b>	no	
43.	se o to	<b>19</b>	15	no	

As can be seen above, only 6 of the matching 3-grams were significantly overrepresented in the CET corpus (*po celém světě, vzhledem k tomu, ve srovnání s, pokud jde o, v devadesátých letech and spočívá v tom*). On the other hand, there were also 3 overrepresented 3-grams in the CRO corpus (*na první pohled, v tuto chvíli, do té doby*).

#### 3.4.2.1.4 Summary

To conclude this section on 3-grams, it seems that certain differences between translated and non-translated texts can be observed in terms of the overall frequency and the distribution of types and tokens (Graph 10 and Graph 11). The language of translated Czech (the CET corpus) overall not only makes greater use of these formulaic expressions, but the most frequent n-grams are also used more often than the corresponding structures in the corpus of Czech originals (Graph 12). Even though the MCC at hand is rather modest in size, this difference in use of patterns of language seems to be prominent enough to distinguish the language of translations from the language of non-translated Czech using a simple 3-gram analysis.

However, the attempted detailed analysis of the matching lexical bundles encountered two problems: insufficient software and corpus size. Firstly, this research would greatly benefit from more elaborate software for analysis which would enable automatic sorting of the data using more than the criteria available (and applying two criteria at a time, namely the minimum frequency of occurrence and ensuring that the n-grams are spread across at least 5 different texts). Secondly, it must be noted that a bigger corpus would be needed to draw an inescapable conclusion concerning the difference in the use of lexical bundles with respect to translated and non-translated texts.

### 3.4.2.2 4-grams

#### 3.4.2.2.1 Corpus query for extraction

Figure 10 shows the corpus query used for the extraction of 4-grams. The search attribute “word (lowercase)” and the minimum frequency (5 instances in the whole corpus) remained the same as for the extraction of 3-grams. The only difference was the n-gram value “4 to 4”.

**Figure 10: 4-gram search query**

The screenshot shows the 'Word list options' interface. On the left is a navigation menu with 'Word list' highlighted in red. The main panel has the following settings:

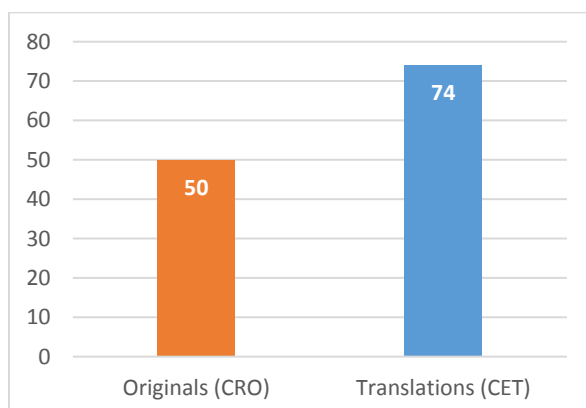
- Subcorpus: create new
- Search attribute: word (lowercase)
- use n-grams. Value of n: from 4 to 4
- hide/nest sub-n-grams
- Filter options:
  - Filter word list by: Regular expression: (empty)
  - Minimum frequency: 5
  - Maximum frequency: 0 (0 = no maximum frequency)
  - Whitelist: Procházet... Soubor nevybrán. Clear
  - Blacklist: Procházet... Soubor nevybrán. Clear format
  - Include non-words
- Output options:
  - Frequency figures:  Hit counts  Document counts  ARF
  - Output type:  Simple  Keywords
  - Reference (sub)corpus: Czech Web 2012 (czTenTen12 v8) (whole corpus)
  - Prefer: rare words  common words 1
  - Change output attribute(s)

A 'Make word list' button is at the bottom left. Red arrows point to the search attribute, the n-gram range, and the minimum frequency field.

#### 3.4.2.2.2 The overall frequency and number

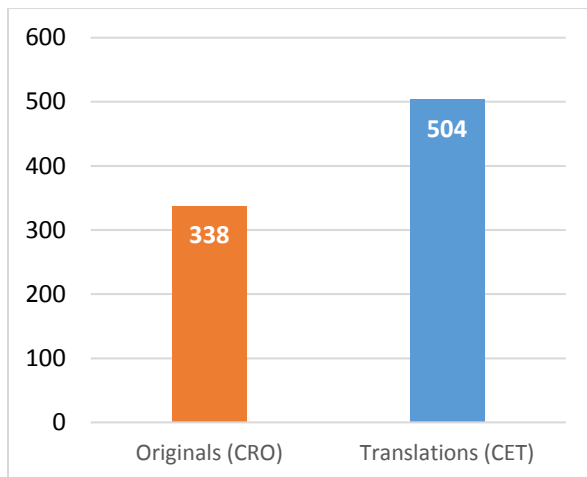
Sketch engine identified 50 4-grams types in the CRO corpus and 74 4-grams types in the CET corpus (see Graph 14). This difference is significant at  $p < .05$ .

**Graph 14: Absolute frequency of 4-gram types**



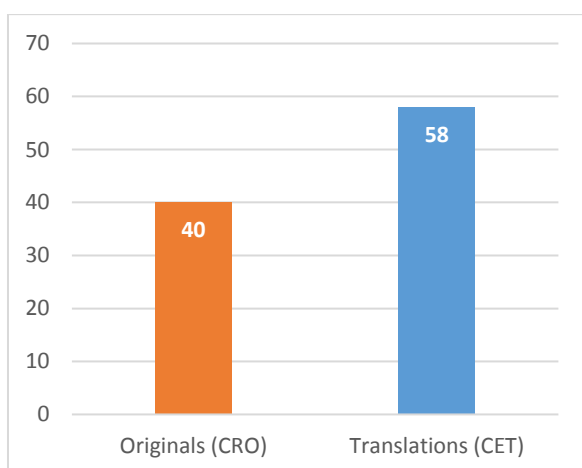
The subsequent analysis of the 4-grams revealed 338 4-gram tokens in the CRO corpus and 504 4-gram tokens in the CET corpus (Graph 15). This difference also satisfies the established criteria of the statistical significance (at  $p < .001$ ).

**Graph 15: Absolute frequency of 4-gram tokens**



However, in order to satisfy the second criterion for the minimum occurrence, all the 4-grams<sup>45</sup> which were not spread at least across 5 different texts were removed from the final list for the subsequent analysis. (10 types from the CRO corpus and 16 types from the CET corpus did not satisfy this criterion). Graph 16 shows the filtered list: there were 40 4-gram types in the CRO corpus and 58 4-gram types in the CET corpus; a difference which is not statistically significant.

**Graph 16: Absolute frequency of 4-gram types (filtered)**

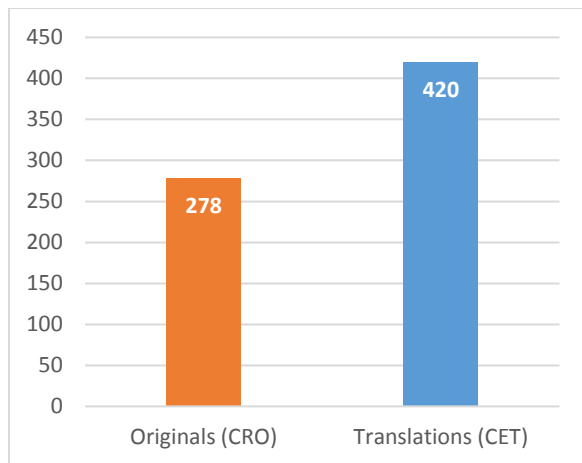


<sup>45</sup> The relatively lower incidence of 4-gram types allowed for manual sorting of the data according to the distribution criterion which could not be satisfied for 3-grams in the previous section (apart from the top one hundred 3-gram types).



Nevertheless, there were 278 4-gram tokens in the CRO corpus and 420 4-gram tokens in the CET corpus (see Graph 17). In this case, the difference is once more statistically significant at  $p < .001$ .

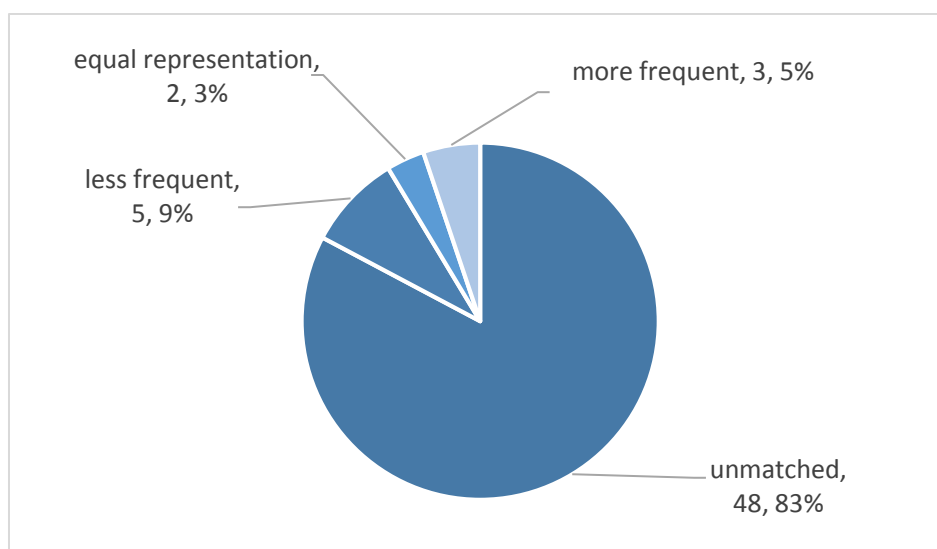
**Graph 17: Absolute frequency of 4-gram tokens (filtered)**



#### 3.4.2.2.3 Analysis of possible overrepresentation

As the next step, a closer analysis of the filtered 4-grams was undertaken. Out of the 58 4-gram types in the CET corpus, only ten had its match in the CRO corpus. Out of these ten, three were more frequent in the CET corpus, five were less frequent and two were represented equally (when compared with the respective frequencies in the CRO corpus) as can be seen in Graph 18.

**Graph 18: Relative frequency of 3-grams (filtered) in the CET corpus compared to the CRO corpus**



Significance testing revealed that the observed differences in the absolute frequencies across the two corpora are not statistically significant, with the exception of the 4-gram “bez ohledu na to”, which was significantly underrepresented in the CET corpus at  $p < .01$  (see Table 4).

**Table 4: Comparison of the distribution of the matching 4-grams**

No.	4-gram	Originals (CRO)	Translations (CET)	Statistically significant	p-value
1.	od té doby se	19	10	no	
2.	<b>bez ohledu na to</b>	<b>19</b>	5	<b>yes</b>	<b>p &lt; .01</b>
3.	po druhé světové válce	12	12	(match)	
4.	a od té doby	10	10	(match)	
5.	a vzhledem k tomu	9	5	no	
6.	jedním z nich je	8	6	no	
7.	se v posledních letech	7	8	no	
8.	ať už jde o	7	5	no	
9.	v posledních deseti letech	6	7	no	
10.	v posledních dvou letech	5	7	no	

This result is inconsistent with the hypothesis that n-grams are more frequent in translated texts than in non-translated Czech, even though it is only one instance of such underrepresentation in the translated corpus. It is also apparent that this kind of 4-gram analysis is not suitable for such a small corpus. The concordances are too scarce to produce any conclusive results when comparing the individual 4-grams.

#### 3.4.2.2.4 Summary

At a first glance, the corpus of Czech translations (CET) seems to be using more 4-grams types and tokens than the corpus of non-translated Czech. Both of the differences presented in Graph 14 and Graph 15 are statistically significant. The differences observed in the frequency distribution of the filtered 4-grams (which are spread across at least 5 different texts) also confirm this tendency (Graph 16 and Graph 17), but only the difference in the frequency of the 4-gram tokens is statistically significant.

The analysis of the possible overrepresentation of the individual 4-grams (Graph 18) is inconclusive. No 4-grams are overrepresented in the CET corpus and there is even evidence to the contrary: one 4-gram (“bez ohledu na to”) is significantly overrepresented in the CRO corpus.

### 3.4.3 Lexical richness

For the analysis of the lexical richness of Czech, the category “lemma” was chosen to represent “types” in the equation<sup>46</sup> for calculating the TTR. This was recommended by Lucie Chlumská in order to account for the diversity of lexemes in both corpora rather than for the diversity of the individual word forms.<sup>47</sup>

The number of lemmas (32,382 in the CET corpus, 34,593 in the CRO corpus) was divided by the respective number of tokens for each corpus, multiplied by 100 and rounded up to the fourth decimal place. The results (Graph 19) show that the lexical richness for the CRO corpus is 5.4091% while only 5.0673% for the CET corpus.

*Graph 19: Comparison of lexical richness*

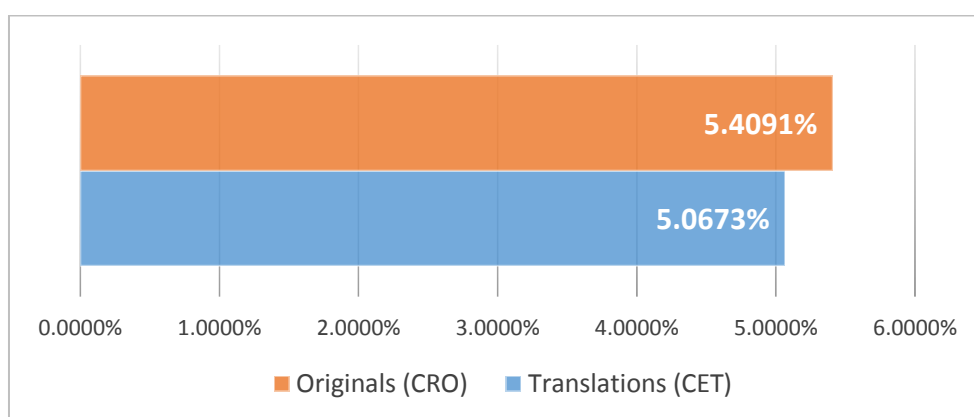
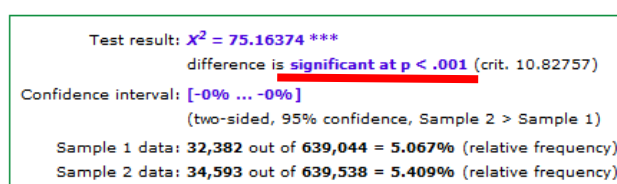


Figure 11 shows the test for the statistical significance of the difference in lexical richness. The total number of lemmas for each of the corpora was entered as the “frequency count” and the respective number of tokens for each corpus as the “sample size”. The difference proved to be significant at  $p < .001$ .

*Figure 11: Statistical significance of the lexical richness*

#### Corpus Frequency Test: Two Samples



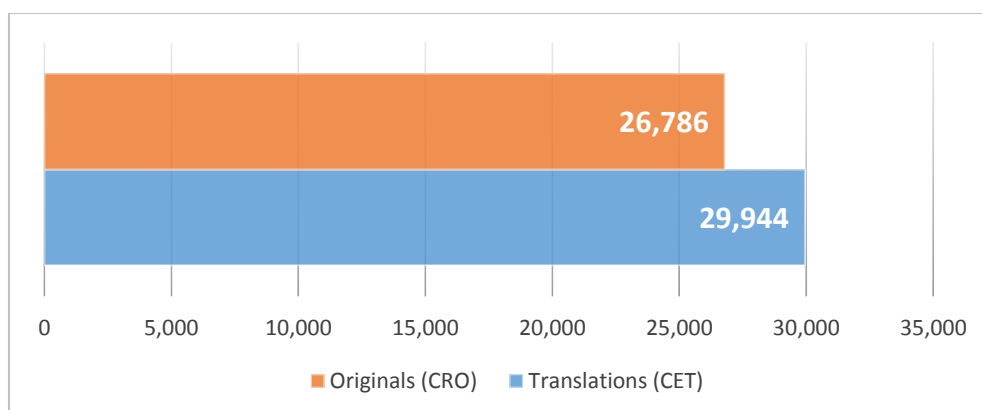
<sup>46</sup> TTR = (number of types/number of tokens) \* 100 .

<sup>47</sup> Mgr. Lucie Chlumská, Ph.D., personal communication, March 10, 2017.

### 3.4.4 Average sentence length

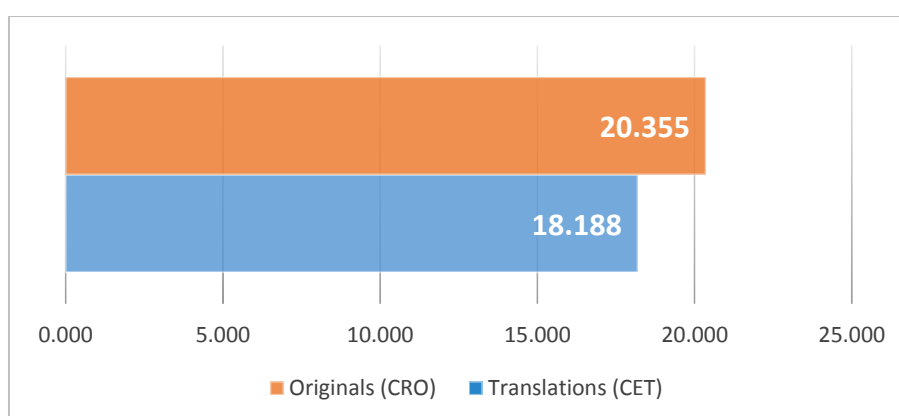
Because the two corpora are comparable with respect to the number of words and tokens (see Graph 3) even the basic graph representing the total number of sentences provides some insight into the basic textual features. The CET corpus contains 29,944 sentences whereas the CRO corpus contains only 26,786 sentences (Graph 20).

*Graph 20: Number of sentences in both corpora*



The average sentence length for each corpus was calculated by dividing the total word count by the number of sentences and rounded up to the third decimal place. The results (Graph 21) show that translations (the CET corpus) have shorter average sentence length than the original Czech texts (the CRO corpus). The average non-translated sentence has 20.355 words while the average translated sentence has 18.188 words.

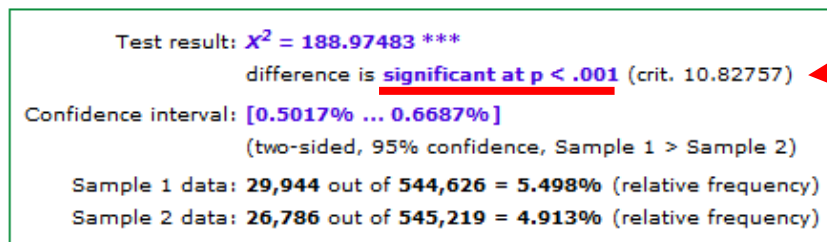
*Graph 21: Average sentence length*



Once again, a test for statistical significance was administered. The “frequency count” was represented by the respective number of sentences in each corpus and the total number of words represented the “sample size”. The difference in the number of sentences in each corpus (and thus also the average sentence length) proved to be significant at  $p < .001$  (see Figure 12).

*Figure 12: Statistical significance of the average sentence length*

### Corpus Frequency Test: Two Samples



```
Test result:  $\chi^2 = 188.97483$  ***  
difference is significant at  $p < .001$  (crit. 10.82757)  
Confidence interval: [0.5017% ... 0.6687%]  
(two-sided, 95% confidence, Sample 1 > Sample 2)  
Sample 1 data: 29,944 out of 544,626 = 5.498% (relative frequency)  
Sample 2 data: 26,786 out of 545,219 = 4.913% (relative frequency)
```

## 4 CONCLUSION

As attested by numerous researchers such as Koppel and Ordan (2011), Baroni and Bernardini (2005, 2006) Ilisei et al. (2010) and Laviosa (1998a, 1998b), it is indeed possible to distinguish translated texts from non-translated texts. Even though the claims of the universal validity of any such features of translated language are often contested on many levels, this study aimed to test this general hypothesis on a small corpus of Czech journalistic texts, comparing Czech originals and Czech translations from English. Based on the available literature dealing with the typical features of translated language and their possible use for distinguishing translated and non-translated language, several key features suitable for a quantitative analysis were identified. The examined features were the frequency distribution of lexical bundles (3-grams and 4-grams), the lexical richness (the comparison of TTR) and the average sentence length.

The comparative analysis of lexical bundles examined the frequency of the occurring 3-grams and 4-grams in terms of the absolute frequency of types and tokens which satisfied the established threshold of the minimum occurrence. The analysis proves that 3-grams are more frequently used in the corpus of translated Czech and all of the results supporting this claim are statistically significant. There were twice as many significantly overrepresented matching 3-grams in comparison with the corpus of Czech originals (6 overrepresented 3-grams in the translated corpus while only 3 in the non-translated corpus).

The subsequent comparative analysis of 4-grams across the two corpora proved that for four of the five examined sets of data, the frequencies were higher for the translated corpus. Out of these four data sets, the observed differences were statistically significant in three cases. Only the examination of the matching 4-grams provided results inconsistent with the hypothesis that 4-grams are more frequent in the translated corpus. Nevertheless, one instance of underrepresentation in the translated corpus is not enough to falsify our hypothesis. Arguably, this result rather attests to the fact that 4-gram analysis is not suitable for such a small corpus. We can conclude that there is not enough data to draw any valid conclusion. But at the same time, we cannot rule out the possibility that our hypothesis could be falsified if we had enough data.

On the whole, the differences demonstrated in the frequency distribution of 3-grams and 4-grams are consistent with our hypothesis that n-grams are more frequent in translated texts (the CET corpus) than in non-translated texts (the CRO corpus). These findings are consistent with Baroni and Bernardini's (2003, 379) claim that there is a higher incidence of repeated patterns in translations. This increased repetitiveness of language might be a result of the translator's effort to achieve increased fluency in the target language as suggested by Xiao (2011, 145), nevertheless, this is just one of the possible interpretations. It might also be a sign of the translator's effort to improve readability of the text as proposed by Bisiada (2015, 24). In turn, the results of the n-gram analysis might be an indicator of the normalization tendency (Zanettin, 2013, 24). At the same time, it is evident that the reliability of the presented differences greatly suffers when we move to the examination of 4-grams, which are considerably less frequent than 3-grams. The 4-gram analysis might be very well suited for the analysis of translated Czech of a much bigger corpus.

The next examined feature, the lexical richness—judged to be the most reliable according to Ilisei et al. (2010)—was supposed to be lower for the corpus of translated texts. The CET corpus indeed exhibits significantly lower lexical richness (5.0673%) when compared with the CRO corpus (5.4091%). The higher lexical richness of the non-translated corpus testifies that there is a relatively richer variety of language used. This proves that translated texts may be recognised on the basis of the relatively lower TTR. The hypothesis that translated texts exhibit lower lexical richness than non-translated texts was confirmed.

The last feature under examination, the average sentence length, was also supposed to be lower for the translated corpus. Once again, this tendency to use shorter sentences holds with the presented data: the average sentence length in the CET corpus (18.188 words) is lower than the average sentence length in the CRO corpus (20.355 words). The difference in the number of sentences in each corpus is statistically significant. This result is consistent with the hypothesis that the average sentence length is lower in translated texts compared to the originals. This result might even suggest that the translator's overall strategy might have included splitting of long sentences. The lower average sentence length along with the lower TTR provide support for the simplification TU (Zanettin 2013, 23) or in Baker's view (1996, 184) for the levelling out (convergence) TU.

Table 5 below provides an overview of the examined features along with the respective numbers for both corpora. The second column from the right provides the evaluation of the data with respect to the hypothesis about that individual feature regarding translated and non-translated texts as stated in the introduction. The last column indicates the statistical significance of that particular observed difference across the two corpora. As can be seen, ten out of the eleven examined features (data sets) support the hypotheses concerning the language of translations. Out of these ten features, nine of the observed differences were proved to be statistically significant.

**Table 5: Overview of the examined features**

No.	Examined feature	Originals		Translations	Consistent with the hypothesis	Statistical significance
		The CRO corpus		The CET corpus		
1	Absolute frequency of 3-gram types	877	<	1,097	yes	p < .001
2	Absolute frequency of 3-gram tokens	7,441	<	9,715	yes	p < .001
3	Absolute frequency of 100 top 3-grams (filtered)	2,176	<	2,555	yes	p < .001
4	Significantly overrepresented matching 3-grams (filtered)	3	<	6	yes	see Table 3
5	Absolute frequency of 4-gram types	50	<	74	yes	p < .05
6	Absolute frequency of 4-gram tokens	338	<	504	yes	p < .001
7	Absolute frequency of 4-gram types (filtered)	40	<	58	yes	not significant
8	Absolute frequency of 4-gram tokens (filtered)	278	<	420	yes	p < .001
9	Significantly overrepresented matching 4-grams (filtered)	1	>	0	no	p < .01
10	Lexical richness (TTR)	5.4091%	>	5.0673%	yes	p < .001
11	Average sentence length	20.355	>	18.188	yes	p < .001

From the outcome of our investigation, it is possible to conclude that it is indeed possible to identify certain textual features that can help us distinguish between the language of Czech translations from English and the language of original untranslated Czech. In this respect, the research into the lexical richness and the



average sentence length produces quite convincing results. The examination of n-grams also proved to be quite useful. However, the size of the corpus at hand is very important, hence the relatively less convincing differences when it comes to the examination of 4-grams.

The corpus size is an issue which needs to be addressed further. While a bigger corpus certainly provides more data, at the same time, it poses a problem in terms of data analysis. In this case, Sketch Engine did not display concordances of n-grams below a certain frequency (5 occurrences). It also did not allow automatic sorting of the n-grams according to multiple criteria at once which considerably hindered the analysis. More elaborate software which would meet all the researcher's needs and allowed automatic processing of a large amount of data (in this particular case of n-grams) would be very beneficial.

Concerning the average sentence length, it would definitely be profitable to have the parallel corpus of the English originals (the source texts) at our disposal. The possibility to compare the originals with their respective translations might tell us more about the translators' strategies and could further strengthen the claim that translations undergo the process of simplification as the lower average sentence length suggests. A combination of an MMC and a parallel corpus might be very well suited for the research into translated language. It is also clear that a combination of quantitative and qualitative methods would certainly be advantageous. It would allow us to deeper, below the surface structures, to explain the language tendencies more thoroughly and with a higher degree of certainty. For example Baroni and Bernardini's (2003, 737-739) distinction between topic-dependent and topic-independent n-grams might shed some light on the distribution of lexical bundles. A structural analysis of n-grams along with the PoS-gram examination could also be highly beneficial.

The findings of our research are quite convincing, and thus the following conclusion can be drawn: the majority of the examined features provide tangible proof for the hypothesis that the language of translation has features distinct from the language of non-translations. At the same time, it must be noted that the results cannot serve as confirmation of any truly universal language tendency due to the research's limitations. Apart from the limited size of the corpus mentioned earlier, the languages and language directionality must be taken into account. In this case, possible tendencies of Czech translations from English came under scrutiny. The next

important limitation is the domain of the texts included in the corpus; the specific nature of journalistic translation and journalistic texts (as discussed in section 2.2.1) in general is yet another important factor which must be accounted for. It is not unusual for journalists to work with numerous sources some of which might be in a different language, thus the status of the original Czech texts might not be the “pure” non-translated language after all.

Last, but not the least, the authorship of the Czech translations and the English originals poses a serious problem. Even though it is reasonable to assume that the English originals come from numerous authors (maybe as a result of a collaborative effort), the publisher of *Respekt* indicated that there is only one translator of “The Economist” section altogether. If this is really the case, the results of the analysis could be skewed and most likely limited to the tendency of one particular translator (if the translator has a distinct style).

Bearing all these limitations in mind, we can conclude that further research into the features of translated Czech using corpus linguistics tools would certainly greatly benefit our understanding of the translation processes and of the possible T-universals. Corpus linguistics offers numerous possibilities for linguistic analysis which go well beyond the scope of this paper which explored just a fraction of the possible utility of this approach. We can conclude with Biel’s words which are still relevant: “Research on translation universals and patterns in translated language is still at an early stage and it remains to be seen where it will take us,” (2009, 12).

## **5 APPENDICES**

1. Originals Corpus CRO 3-gram list.xml
2. Originals Corpus CRO 4-gram list.xml
3. Translations Corpus CET 3-gram list.xml
4. Translations Corpus CET 4-gram list.xml

## 6 SHRNU TÍ

Tato práce se zabývá překladovou češtinou ve srovnání s češtinou nepřekladovou. Klade si za cíl identifikovat možné charakteristiky češtiny jakožto cílového jazyka v překladech z angličtiny ve srovnání s češtinou původní, tj. nepřekladovou. Pracuje s teorií tzv. T-univerzálií, která předkládá hypotézu, že překladový jazyk vykazuje jisté rysy, na základě kterých je možné jej rozeznat od jazyka překladového – jedná se o takzvané T-univerzálie vztažené ke konkrétnímu jazyku.

Pro účely této práce byl sestaven jednojazyčný srovnatelný korpus českých žurnalistických textů publikovaných na webu týdeníku Respekt. První korpus (CET) obsahuje české překlady textů původně publikovaných týdeníkem *The Economist* v anglickém jazyce. Druhý korpus (CRO) obsahuje české originální texty původně publikované taktéž týdeníkem Respekt a byl sestaven tak, aby žánrově i tematicky odpovídal druhému korpusu a oba tedy jako celek splnily kritéria korpusu srovnatelného. Následná komparativní analýza zkoumá, zda je na základě identifikovaných charakteristik možné zjistit, který z korpusů je produktem překladatelského procesu a který nikoliv.

Teoretická část práce se věnuje vymezení základních pojmů a konceptů, na kterých analýza překladového jazyka staví. Nejprve je stručně představena korpusová lingvistika a její místo v translatologii. Následuje definice monolingválního srovnatelného korpusu a kritéria srovnatelnosti. Druhá část pak osvětluje specifika žurnalistických textů a představuje základní problémy a výzvy, které jejich překlad představuje. Zvláštní pozornost je věnována vnímání role překladatele a novináře (mnohdy v jedné osobě), jeho základním strategiím a metodám a viditelnosti pro cílového čtenáře.

Další část představuje překladové univerzálie jakožto základní premisu, z nichž autoři zkoumající možné odlišnosti překladového jazyka vycházejí. Popisuje vnímání překladových univerzálií od dob prvních teoretických úvah o překladu (období tzv. ideálních univerzálií), přes pejorativní pojetí až k období deskriptivnímu. Následuje přehled současného vědeckého bádání na toto téma se zaměřením na možnou identifikaci rysů překladového jazyka. Ve prospěch jejich existence hovoří například Koppel a Ordan (2011), Baroni a Bernardini (2006), Ilisei a kol. (2010) a také Baker (1996). Další autoři včetně Laviosy (1998) a Xiao (2011) pak pojednávají

o rozdílech, které spatřují v opakování určitých struktur, poměru gramatických a lexikálních slov a n-gramů. Další sekce je věnována stěžejním pracím zabývajícím se konkrétně rysy překladové češtiny. Stručné shrnutí kritických pohledů na problematiku překladových univerzálií následuje formulace základních hypotéz, stanovených na základě výše zmíněných autorů: 1) Výskyt n-gramů je častější v překladovém jazyce (n-gramy jsou ve srovnání s nepřekladovým jazykem nadužívány), 2) Překladový jazyk je méně lexikálně bohatý než nepřekladové texty v témže jazyce, 3) Překladové texty mají nižší průměrnou délku vět ve srovnání s nepřekladovými texty v témže jazyce. V závěru teoretické části jsou pak tyto zkoumané parametry blíže představeny – důraz je kladen především na jejich využití a možná omezení při zkoumání překladového jazyka.

Analytická část práce pak popisuje samotnou kompilaci korpusu a analýzu zmiňovaných parametrů. Oba subkorpusy byly vytvořeny skrze webové rozhraní konkordanceru Sketch Engine, který krom vyhledávání v korpusech již zkompilovaných umožňuje sestavení vlastního uživatelského korpusu. Po krátkém představení Sketch Enginu následuje popis návrhu korpusu a základních kritérií, na jejichž základě byly texty vybírány. Samotný korpus byl sestaven tak, aby co nejlépe odpovídal kritériím srovnatelného jednojazyčného korpusu. Subkorpus CET obsahuje překlady publikované mezi lety 2007 až 2017 v sekci „The Economist“, která sdružuje články převzaté/přeložené z anglického časopisu The Economist. Má celkem 639 044 tokenů (544 626 slov) a obsahuje 418 jednotlivých textů. Subkorpus CRO byl navržen tak, aby velikostí (počtem slov/tokenů) a tematickým zastoupením překladovému subkorpusu odpovídal. Obsahuje celkem 639 538 tokenů (545 219 slov) a 408 jednotlivých textů. Podobně jako překladový subkorpus zastřešuje témata jako ekonomika, domácí a světová politika, kultura, vzdělávání, historie, věda a technika. Obsažené texty pokrývají stejné období, bohužel však u obou subkorpusů nebylo možné dosáhnout rovnoměrného zastoupení v jednotlivých letech ani jednotné délky konkrétních textů. Oba subkorpusy se však příliš neliší průměrnou délkou zahrnutých textů a variabilita délek jednotlivých textů v obou korpusech je srovnatelná. Oba korpusy jsou srovnatelné také na základě skutečnost, že všechny texty byly publikovány ve stejném médiu a jsou tak určeny stejnému okruhu čtenářů.

Druhá část analytické sekce pak obsahuje samotnou analýzu výše zmíněných sledovaných parametrů za využití nástroje *Corpus Frequency Wizard*, který umožňuje

zjistit míru statistické signifikance při porovnávání dvou vzorků ze dvou korpusů. Přehled výsledků sledovaných kategorií je uveden v tabulce níže (Table 6).

Komparativní analýza n-gramů odhalila, že 3-gramy (dle počtu typů celkového počtu typů a tokenů) jsou v překladovém korpusu zastoupeny signifikantně častěji. Stejná tendence převládala i při porovnání absolutního zastoupení 100 nejčastějších 3-gramů. Tato tendence opět potvrdila hypotézu, že překladové texty lze rozeznat na základě relativního nadužívání 3-gramů, což by mohla být známka překladatelovy snahy o dosažení idiomatičtějšího vyjadřování v cílovém jazyce.

Analýza 4-gramů odhalila podobné tendence tyto struktury nadužívat, a to ve čtyřech z pěti sledovaných kategorií. Rozdíl v zastoupení unikátních filtrovaných 4-gramů však již nebyl statisticky signifikantní a poslední sledovaná kategorie odhalila tendenci opačnou, byť jen v jednom případě. Komparativní analýza 4-gramů svědčí o podobných tendencích jejich nadužívání v překladech, výsledky však již nejsou tak přesvědčivé jako u 3-gramů. Toto dokládá, že analýza 4-gramů pro takto malý korpus není příliš vhodnou metodou, mohla by však být užitečná pro podobný výzkum ve větším měřítku.

Druhé sledované kritérium, lexikální bohatost, prokázalo hypotézu, že nepřekladové texty jsou lexikálně bohatší než texty překladové, což by mohlo svědčit o větší standardizaci a repetitivnosti překladové češtiny. Poslední sledované kritérium, průměrná délka vět, rovněž potvrdilo hypotézu, že překladové texty obsahují ve srovnání s nepřekladovým korpusem relativně kratší věty, což by mohlo poukazovat na překladatelovu strategii dělit delší věty na kratší úseky. Tato strategie by se mohla projevit jako simplifikace.

Představená kontrastivní analýza potvrzuje základní hypotézu, že překladovou češtinu je skutečně možné identifikovat na základě kvantitativní korpusové analýzy. Jako nejvhodnější se jeví především srovnání lexikální bohatosti, průměrné délky vět a analýza zastoupení 3-gramů. Tyto závěry jsou však vztaženy ke konkrétnímu korpusu v dané jazykové kombinaci a typu textů.

Je třeba zmínit, že chybějící metadata k jednotlivým subkorpusům, konkrétně jednoznačné určení překladatele nebo překladatelů a absence jsem autorů původních anglicky psaných textů, jsou jistou překážkou pro možné zobecnění výsledovaných tendencí a formulování jednoznačných závěrů. Dalším omezením je pak relativně skromná velikost obou korpusů, a tak je vhodné tento výzkum spíše vnímat jako

malou pilotní studii, která poukazuje na možnosti využití analýzy n-gramů a dalších zmiňovaných kritérií pro další studium charakteristik překladové češtiny. Pro další směřování výzkumu překladové češtiny je možné navrhnout analýzu n-gramů s důrazem na jejich strukturu (viz. Rafiee a Keihaniyan 2012), jako kombinaci slovních druhů (tzv. PoS-gramy) případně ve smyslu specifčnosti pro dané téma (viz. Baroni a Bernardini, 2003). Právě spojení kvantitativních a kvalitativních metod by mohlo odhalit zákonitosti, které tyto přístupy samostatně neobsáhnou. Velmi přínosné by pro vysvětlení sledovaných tendencí jazyka bylo současné využití jednojazyčného korpusu a korpusu paralelního, který by umožnil porovnat překladové texty s texty zdrojovými.

**Table 6: Přehled výsledků**

Č.	Sledované parametry	Nepřekladové texty		Překladové texty	Odpovídá hypotéze	Statistická signifikance
		Korpus CRO		Korpus CET		
1	Absolutní frekvence 3-gramů (typů)	877	<	1097	ano	p < .001
2	Absolutní frekvence 3-gramů (tokenů)	7441	<	9715	ano	p < .001
3	Absolutní frekvence 100 nejčastějších 3-gramů (filtrované)	2176	<	2555	ano	p < .001
4	Signifikantně nadužívané 3-gramy (filtrované)	3	<	6	ano	viz Tabulka 6
5	Absolutní frekvence 4-gramů (typů)	50	<	74	ano	p < .05
6	Absolutní frekvence 4-gramů (tokenů)	338	<	504	ano	p < .001
7	Absolutní frekvence filtrovaných 4-gramů (typů)	40	<	58	ano	není signifikantní
8	Absolutní frekvence filtrovaných 4-gramů (tokenů)	278	<	420	ano	p < .001
9	Signifikantně nadužívané 4-gramy (filtrované)	1	>	0	ne	p < .01
10	Lexikální bohatost (TTR)	5,4091 %	>	5,0673 %	ano	p < .001
11	Průměrná délka vět	20,355	>	18,188	ano	p < .001

## 7 LIST OF FIGURES, GRAPHS AND TABLES

Figure 1: Creating the corpus.....	52
Figure 2: Uploading the corpus.....	52
Figure 3: Expansion of the archives.....	52
Figure 4: Compiling the corpus .....	53
Figure 5: Tagging the corpus .....	53
Figure 6: Frequency comparison of two samples .....	55
Figure 7: Testing for statistical significance (3-gram "na celém světě"): input.....	56
Figure 8: Testing for statistical significance (3-gram "na celém světě"): result.....	56
Figure 9: 3-gram search query .....	57
Figure 10: 4-gram search query .....	63
Figure 11: Statistical significance of the lexical richness .....	67
Figure 12: Statistical significance of the average sentence length.....	69
Graph 1: Number of texts in each corpus across the years .....	43
Graph 2: Number of words in each corpus across the years .....	44
Graph 3: Number of words and tokens in each corpus .....	44
Graph 4: Text length variability.....	45
Graph 5: Average/median text length .....	46
Graph 6: Number of authors per article (the CRO corpus).....	47
Graph 7: Number of articles per author (the CRO corpus).....	48
Graph 8: Authors by gender (the CRO corpus) .....	48
Graph 9: Number of articles according to gender (the CRO corpus) .....	49
Graph 10: Absolute frequency of 3-gram types.....	58
Graph 11: Absolute frequency of 3-gram tokens.....	59
Graph 12: Absolute frequency of 100 top 3-grams (filtered) .....	60
Graph 13: Relative frequency of 100 top 3-grams (filtered) in the CET corpus compared to the CRO corpus .....	60
Graph 14: Absolute frequency of 4-gram types.....	63
Graph 15: Absolute frequency of 4-gram tokens.....	64
Graph 16: Absolute frequency of 4-gram types (filtered).....	64
Graph 17: Absolute frequency of 4-gram tokens (filtered).....	65
Graph 18: Relative frequency of 3-grams (filtered) in the CET corpus compared to the CRO corpus .....	65



Graph 19: Comparison of lexical richness.....	67
Graph 20: Number of sentences in both corpora .....	68
Graph 21: Average sentence length .....	68
Table 1: Comparing the publication dates (original vs. translation).....	50
Table 2: Calculating the “sample size” (the total number of n-grams).....	55
Table 3: Comparison of the frequency distribution of the matching 3-grams .....	61
Table 4: Comparison of the distribution of the matching 4-grams .....	66
Table 5: Overview of the examined features .....	72
Table 6: Přehled výsledků.....	79

## 8 WORKS CITED

- Allen, David. 2009. 'Lexical Bundles in Learner Writing: An Analysis of Formulaic Language in the ALESS Learner Corpus'. *Komaba Journal of English Education*, no. 1/2009: 105–27.
- Apostol, Mihaela Simona, Adriana Anca Cristea, and Tatiana Corina Dosecu. 2015. 'The Peculiarities of Journalistic Discourse'. *Quality - Access To Success*, no. 16: 146–51.
- Baker, Mona. 1993. 'Corpus Linguistics and Translation Studies — Implications and Applications'. In *Text and Technology: In Honour of John Sinclair*, edited by Elena Tognini-Bonelli, Francis Gill, and Baker Mona, 233–50. Amsterdam and Philadelphia: John Benjamins.
- . 1996. 'Corpus-Based Translation Studies: The Challenges That Lie Ahead'. In *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager*, edited by Harold Somers, 175–86. Amsterdam: John Benjamins.
- . 2004. 'A Corpus-Based View of Similarity and Difference in Translation'. *International Journal of Corpus Linguistics* 9 (2): 167–93.
- Bani, Sara. 2006. 'An Analysis of Press Translation Process'. In *Translation in Global News*, edited by Kyle Conway and Susan Bassnett. United Kingdom.
- Bardoel, Thomas. 2012. 'Comparing N-Gram Frequency Distributions: Explorative Research on the Discriminative Power of N-Gram Frequencies in Newswire Corpora'. Master thesis, Tilburg: Tilburg University.
- Baroni, Marco, and Silvia Bernardini. 2003. 'A Preliminary Analysis of Collocational Differences in Monolingual Comparable Corpora'. In *Proceedings of the Corpus Linguistics 2003 Conference*, edited by Dawn Archer, Paul Rayson, and Tony McEnery, 16:82–91. Lancaster: Lancaster University.
- . 2005. 'Spotting Translationese A Corpus-Driven Approach Using Support Vector Machines'. In . Birmingham.
- . 2006. 'A New Approach to the Study of Translationese: Machine-Learning the Difference between Original and Translated Text.' *Literary and Linguistic Computing* 21 (3): 259–74.
- Baroni, Marco, and Stefan Evert. 2017. 'SIGIL: Corpus Frequency Test Wizard'. Online utility. *SIGIL*. April 6. <http://sigil.collocations.de/wizard.html>.

- Bednarek, Monika, and Helen Caple. 2012. *News Discourse*. London New York: Continuum.
- Berman, Antoine. 1985. 'Translation and the Trials of the Foreign'. In *The Translation Studies Reader*, edited by Lawrence Venuti, 285–97. London: Routledge.
- Bernardini, Silvia. 2011. 'Monolingual Comparable Corpora and Parallel Corpora in the Search for Features of Translated Language'. *SYNAPS – A Journal of Professional Communication* 26: 2–13.
- Biber, Douglas, Susan Conrad, and Viviana Cortes. 2004. 'If You Look At...: Lexical Bundles in University Teaching and Textbooks'. *Applied Linguistics* 25 (3): 371–405.
- Biber, Douglas, Stig Johansson, Edward Finegan, Geoffrey Leech, and Susan Conrad. 1999. *Longman Grammar of Spoken and Written English*. Harlow, England: Pearson Education Limited.
- Biel, Łucja. 2009. 'Corpus-Based Studies of Legal Language for Translation Purposes: Methodological and Practical Potential'. In *Reconceptualizing LSP. Online Proceedings of the XVII European LSP Symposium*.
- Bielsa, Esperança, and Susan Bassnett. 2009. *Translation in Global News*. London and New York: Routledge.
- Bisiada, Mario. 2015. 'Universals of Editing and Translation.' In *Empirically Modelling Translation and Interpreting*, edited by Silvia Hansen-Schirra, Sascha Hofmann, and Bernd Meyer, 1–33. Berlin: Language Science Press.
- Blum-Kulka, Shoshana. 1985. 'Shifts of Cohesion and Coherence in Translation'. In *The Translation Studies Reader*, edited by Lawrence Venuti, 298–313. London: Routledge.
- Chesterman, Andrew. 2003. 'Contrastive Textlinguistics and Translation Universals'. In *Contrastive Analysis in Language Identifying Linguistic Units of Comparison*, edited by Dominique Willems, Bart Defrancq, Timothy Coleman, and Dirk Noël, 213–29. Basingstoke: Palgrave Macmillan.
- . 2004. 'Hypotheses about Translation Universals'. In *Claims, Changes and Challenges in Translation Studies: Selected Contributions from the EST Congress, Copenhagen 2001*, edited by Gyde Hansen, Kirsten Malmkjær, and Daniel Gile, 1–13. Amsterdam/Philadelphia: John Benjamins.

- . 2010. 'Why Study Translation Universals?' *Acta Translatologica Helsingiensia* 1: 38–48.
- . 2011. 'Translation Universals'. In *Handbook of Translation Studies, Vol. 2*, edited by Yves Gambier and Luc van Doorslaer, 195–179. Amsterdam/Philadelphia: John Benjamins.
- Chlumská, Lucie. 2014. 'Není korpus jako korpus: Korpusy v kontrastivní lingvistice a translologii'. *Časopis pro Moderní Filologii* 96 (2): 221–232.
- . 2015. 'Překladová čeština a její charakteristiky'. PhD dissertation, Praha: Charles University.
- . 2016. '(Ne)typické slovní kombinace v českých překladech a možnosti jejich zkoumání'. Edited by Anna Čermáková, Lucie Chlumská, and Markéta Malá, *Studie z korpusové lingvistiky*, no. 23: 235–66.
- Chlumská, Lucie, and Olga Richterová. 2014a. 'Jak zkoumat překladovou češtinu: Výzkum simplifikace na korpusu Jerome'. *Korpus – Gramatika – Axiologie*, no. 09/2014: 16–29.
- . 2014b. 'Překladová Čeština v Korpusech.' *Naše řeč*, no. 4–5: 259–69.
- Cvrček, Václav, and Lucie Chlumská. 2015. 'Simplification in Translated Czech: A New Approach to Type-Token Ratio'. *Russian Linguistics*, no. 39(3): 309–25.
- Detrani, Jason R. 2011. *Journalism: Theory and Practice*. Oakville, Ont.: Apple Academic Press. <http://dx.doi.org/10.1201/b13161>.
- Doorslaer, Luc van. 2010. 'Journalism and Translation'. In *Handbook of Translation Studies*, edited by Luc van Doorslaer and Yves Gambier, 1:180–84. John Benjamins.
- Ellis, Nick C., Rita Simpson-Vlach, and Carson Maynard. 2008. 'Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL'. *TESOL Quarterly* 42 (3): 375–296.
- Evans, Nicholas, and Stephen C Levison. 2009. 'The Myth of Language Universals: Language Diversity and Its Importance for Cognitive Science'. *Behavioral and Brain Sciences* 32 (05): 429–48.
- Gambier, Yves. 2006. 'Transformations in International News'. In *Translation in Global News*, edited by Kyle Conway and Susan Bassnett. United Kingdom.

- Giannossa, Leonardo. 2016. 'Corpus-Based Studies'. In *Researching Translation and Interpreting*, edited by Claudia V. Angelelli and Brian James Baer, 195–202. New York: Routledge.
- Gries, Stefan Th., John Newman, and Cyrus Shaoul. 2011. 'N-Grams and the Clustering of Registers'. *Empirical Language Research Journal* 5 (1): 1–13.
- Halliday, M.A.K., Anna Cermáková, Wolfgang Teubert, and Collin Yallop. 2004. *Lexicology and Corpus Linguistics*. Continuum.
- Hoffmann, Sebastian, Stefan Evert, David Lee, and Ylva Berglund Prytz. 2008. *Corpus Linguistics with BNCweb: A Practical Guide*. English Corpus Linguistics, v. 6. Frankfurt am Main: Peter Lang.
- House, Juliane. 2008. 'Beyond Intervention: Universals in Translation?' *Trans-Kom* 1 (1): 6–19.
- Hyland, Ken. 2008. 'As Can Be Seen: Lexical Bundles and Disciplinary Variation'. *English for Specific Purposes* 27 (2008): 4–21.
- Ilisei, Iustina, Ruslan Mitkov, D Inkpen, and Gloria Corpas Pastor. 2010. 'Identification of Translationese: A Machine Learning Approach'. In *Computational Linguistics and Intelligent Text Processing*, 503–11. Springer Berlin Heidelberg.
- Jiménez-Crespo, Miguel A. 2010. 'The Future of "universal" Tendencies: A Review of Papers Using Localized Websites'. In *UCCTS Conference*, 1–34.
- Kamenická, Renata. 2007. 'Defining Explicitation in Translation.' *Brno Studies in English* 33 (1): 45–57.
- Kenny, Dorothy. 1998. 'Corpora in Translation Studies'. In *Routledge Encyclopedia of Translation Studies*, edited by Mona Baker, 50–53. London: Routledge.
- Koizumi, Rie, and Yo In'nami. 2012. 'Effects of Text Length on Lexical Diversity Measures: Using Short Texts with Less than 200 Tokens'. *System* 40 (4): 522–32. doi:10.1016/j.system.2012.10.017.
- Konšalová, Petra. 2007. 'Explicitation as a Universal in Syntactic De/Condensation.' *Across Languages and Cultures* 8 (1): 17–32.
- Koppel, Moshe, and Noam Ordan. 2011. 'Translationese and Its Dialects'. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1318--1326. 1: Association for Computational Linguistics.

- Kubáčková, Jana. 2009. 'Keeping Czech in Check: A Corpus-Based Study of Generalization in Translation.' *SKASE Journal of Translation and Interpretation* 4 (1): 33–51.
- Laviosa, Sara. 1998a. 'Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose'. *Meta* 43 (4): 1–15.
- . 1998b. 'The Corpus-Based Approach: A New Paradigm in Translation Studies'. *Translators' Journal* 43 (4): 474–79. doi:10.7202/003424ar.
- . 2010. 'Corpora'. In *Handbook of Translation Studies. Volume 1*, edited by Yves Gambier and Luc van Doorslaer, 80–86. Amsterdam/Philadelphia: John Benjamins.
- Lee, Changsoo. 2013. 'Using Lexical Bundle Analysis as Discovery Tool for Corpus-Based Translation Research'. *Perspectives* 21 (3): 378–95. doi:10.1080/0907676X.2012.657655.
- Lind, Sarah. 2007. 'Translation Universals (or Laws, or Tendencies, or Probabilities, Or...?)'. *TIC Talk* 63: 1–10.
- Malmkjær, Kirsten. 2008. 'Norms and Nature in Translation Studies'. *Incorporating Corpora-Corpora and the Translator. Clevedon: Multilingual Matters*, 49–59.
- McEnery, Tony, and Andrew Hardie. 2012a. *Corpus Linguistics*. Cambridge: Cambridge University Press.
- . 2012b. 'Concordancing Tools'. *Corpus Linguistics: Method, Theory and Practice*. October 31. <http://corpora.lancs.ac.uk/clmtp/2-conc.php>.
- McEnery, Tony, and Richard Xiao. 2008. 'Parallel and Comparable Corpora: What Are They up To?' In *Incorporating Corpora. The Linguist and the Translator*, edited by Gunilla Anderman and Margaret Rogers, 18–31. Clevedon: Multilingual Matters.
- Meyer, Charles F. 2004. *English Corpus Linguistics An Introduction*. Cambridge University Press.
- Munday, Jeremy. 2009. *The Routledge Companion to Translation Studies*. London: Routledge.
- Pawley, Andrew. 2009. 'What Is Formulaic Language'. In *Formulaic Language*, edited by Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali, and Kathleen M. Wheatley. Vol. 1. Typological Studies in Language 82. Amsterdam: John Benjamins.

- Pinna, Antonio, and David Brett. 2012. 'Fixedness and Variability: Using PoS-Grams to Study Phraseology in Newspaper Articles'. *Collected Abstracts from the 10th Teaching and Language Corpora Conference, Warsaw*.
- Pym, Anthony. 2008. 'On Toury's Laws of How Translators Translate'. *Benjamins Translation Library* 75: 1-23.
- Rafiee, Marzieh, and Mahbube Keihaniyan. 2012. 'A Comparative Analysis of Lexical Bundles in Journalistic Writing in English and Persian: A Contrastive Linguistic Perspective.' *International Journal of Foreign Language Teaching and Research* 1.2: 37-44.
- Rafiee, Marzieh, Mansoor Tavakoli, and Zahra Amirian. 2011. 'Structural Analysis of Lexical Bundles across Two Types of English Newspapers Edited by Native and Non-Native Speakers.' *Modern Journal of Applied Linguistics* 3, no. 3.2: 218-36.
- 'Redakce, Dopisy.' 2015. *Respekt*. September 20.  
<https://www.respekt.cz/tydenik/2015/39/dopisy>.
- 'Respekt'. 2017. *Respekt*. Accessed May 3. <https://www.respekt.cz/>.
- 'Respekt – Inzerce'. 2017. *Economia*. Accessed October 3.  
<http://economia.ihned.cz/inzerce/respekt/>.
- Salazar, Danica. 2014. *Lexical Bundles in Native and Non-Native Scientific Writing: Applying a Corpus-Based Study to Language Teaching*. Studies in Corpus Linguistics, volume 65. Amsterdam ; Philadelphia: John Benjamins Publishing Company.
- Santos, Diana. 1995. 'On Grammatical Translationese.' In , 59-66.
- 'Sketch Engine'. 2017. Accessed January 3. <https://www.sketchengine.co.uk/>.
- Stubbs, Michael. 2004. 'Language Corpora'. In *The Handbook of Applied Linguistics*, edited by Alan Davies and C. Elder, 106-32. Blackwell Handbooks in Linguistics 17. Malden, MA: Blackwell Pub.
- . 2007. 'Notes on the History of Corpus Linguistics and Empirical Semantics'. *Collocations and Idioms*, 317-29.
- 'SPJ Code of Ethics'. 2014. *Society of Professional Journalists*. June 9.  
<https://www.spj.org/ethicscode.asp>.
- 'The Economist: About Us'. 2017. *The Economist*. Accessed November 3.  
<http://www.economist.com/help/about-us>.

- Torruella, Joan, and Ramon Capsada. 2013. 'Lexical Statistics and Tipological Structures: A Measure of Lexical Richness'. *Procedia – Social and Behavioral Sciences* 95 (October): 447–54. doi:10.1016/j.sbspro.2013.10.668.
- Toury, Gideon. 1995. *Descriptive Translation Studies and beyond*. Amsterdam and Philadelphia: John Benjamins.
- . 2004. 'Probabilistic Explanations in Translation Studies: Welcome as They Are, Would They Qualify as Universals?' In *Translation Universals: Do They Exist?*, edited by Anna Mauranen and Pekka Kujamäki, 15–32. Amsterdam: John Benjamins.
- Tymoczko, Maria. 1998. 'Computerized Corpora and the Future of Translation Studies'. *Meta* 43 (4): 652–59.
- Valdeón, Roberto A. 2012. 'From the Dutch Corantos to Convergence Journalism: The Role of Translation in News Production'. *Meta* 57 (4): 850–65.
- . 2014. 'From Adaptation to Appropriation: Framing the World Through News Translation'. *Linguaculture*, no. 1: 51–62.
- Wollin, Lars. 2005. 'The Language of 19th and 20th Century Swedish'. In *The Nordic Languages*, edited by Wiegand Herbert Ernst, 2:1506–12.
- Xiao, Richard. 2011. 'Word Clusters and Reformulation Markers in Chinese and English: Implications for Translation Universal Hypotheses'. *Languages in Contrast* 11 (2): 145–71.
- Xiao, Richard, Lianzhen He, and Yue Ming. 2008. 'In Pursuit of the Third Code: Using the ZJU Corpus of Translational Chinese in Translation Studies: Using the ZJU Corpus of Translational Chinese in Translation Studies'. *Zhejiang University*.
- Zanettin, Federico. 2013. 'Corpus Methods for Descriptive Translation Studies'. *Procedia – Social and Behavioral Sciences* 95 (2013): 20–32.
- Zelizer, Barbie, and Stuart Allan. 2010. *Keywords in News and Journalism Studies*. Maidenhead: Open University Press.



## 9 ABSTRACT

This paper presents a corpus-based contrastive study based on a monolingual comparable corpus of journalistic texts. It comprises a subcorpus of texts originally written in Czech (non-translations) and a subcorpus of Czech translations from English. It investigates possible differences between the original and translated language and tries to establish whether such differences can provide a basis for distinguishing between the two. Based on the theory of T-universals (language specific translation universals), it examines features which researchers consider the most helpful for distinguishing between translated and non-translated language, namely distribution of lexical bundles (3-grams and 4-grams), lexical richness and average sentence length.

### **Key words**

Corpus, Corpus research, Czech, translationese, translated language, non-translated language, translation universals, T-universals, lexical bundles, n-gram, journalism, explicitation, simplification, normalization, lexical richness, average sentence length, patterns

## 10 ANOTACE

Tato práce se zabývá překladovým jazykem, konkrétně překladovou češtinou v kontrastu s češtinou nepřekladovou (originální). Na základě kontrastivní analýzy jednojazyčného srovnatelného korpusu žurnalistických textů si klade za cíl identifikovat možné rysy překladové češtiny ve srovnání s češtinou originální (nepřekladovou). Pro tyto účely byly sestaveny dva subkorpusy, z nichž první obsahuje originální česky psané texty a druhý české překlady z angličtiny. Tato práce vychází z hypotézy takzvaných T-univerzálií, která předpokládá, že překladový jazyk vykazuje jisté společné rysy, které jej odlišují od jazyka textů nepřekladových. Na základě rešerše odborné literatury zabývající se typickými rysy překladového jazyka bylo identifikováno několik základních rysů, které by dle výzkumníků mohly pomoci rozlišit jazyk překladu a nepřekladového originálu. Konkrétně práce zkoumá distribuci n-gramů (3-gramů a 4-gramů), lexikální bohatost a průměrnou délku vět.

### **Klíčová slova**

Korpus, korpusový výzkum, čeština, překladová čeština, nepřekladový jazyk, překladové univerzálie, T-univerzálie, lexikální svazky, n-gram, žurnalistika, explicitace, simplifikace, normalizace, lexikální bohatost, průměrná délka vět