Filozofická fakulta Univerzity Palackého

# The English of University Web Pages: A Corpus based study

(Bakalářská práce)

2021 Petr Uram

**Filozofická fakulta Univerzity Palackého**

**Katedra anglistiky a amerikanistiky**

# Anglický jazyk na webových stránkách univerzit: Korpusová studie

# The English of University Web Pages: A Corpus based study

**(bakalářská práce)**

Autor: Petr Uram

Studijní obor: Angličtina se zaměřením na komunitní tlumočení a překlad

Vedoucí práce: Mgr. Michaela Martinková, Ph.D.

Olomouc 2021

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a uvedl jsem úplný seznam citované a použité literatury.

V Olomouci dne 19. 8. 2021 ………………………

*Petr Uram*

Děkuji vedoucí, Mgr. Michaele Martinkové, Ph.D., za nadpozemskou trpělivost, obrovskou vstřícnost, veškerou pomoc a cenné rady při zpracování bakalářské práce. Také děkuji Danielu Agnewovi, Janu Velingerovi a Břetislavu Regnerovi za ochotu a poskytnutí informací o struktuře webových stránek českých univerzit.

# Abbreviations

UPOL – Palacký University

UK – Charles University

MUNI – Masaryk University

CAM – Cambridge University

OXF – Oxford University

UCL – University College of London

CUWC – Czech University Web Corpus

EUWC – English University Web Corpus

TTR – type token ratio

STTR – standardized type token ratio

# Table of Contents

# Introduction

Corpus based research has been over the time regarded as one of the most important analytical tools in linguistics. The corpus methodological analysis offers researchers a way of looking into language like no other device can. Structure, lexicon, colloquial expressions, stylistics, and more can be observed by using corpus in a proper manner. And not only that, it may also serve as a helpful tool for translators.

When it comes to translating, Tymoczko says: “Corpus translation studies is central to the way that Translation Studies as a discipline will remain vital and move forward” (1998, 1). Zanettin even claims that a majority of translators use computer corpora as a working aid and to create and access terminological databases and translation memories to make their jobs quicker and easier (2013, 20). Corpus studies allow the Translation Studies to move “from prescriptive approaches (…) to descriptive approaches” (Tymoczko 1998, 1), meaning translation scientists may focus on adaptability and changeability of source and target languages within translations. Though Corpus Based Translation Studies are usually more connected with the use of parallel corpora that feature the original texts and its translations back to back, monolingual comparable corpora are useful, too, as they enable the researchers to compare the translation and non-translation language (Bernardini 2011, 2).

Studies have shown that translated texts have a lower lexical diversity than non-translated texts written in the same language (Berman 1985, Laviosa 1998, Cvrček and Chlumská 2015), and that translation language tends to be “de-complexified” by means of simplification, explicitation, normalization/conservatism, or levelling out (Zanettin 2012, 13). All of these influence lexical richness in a way, but simplification, specifically, relates to “reduction of lexical variability in translated texts” (Cvrček and Chlumská 2015, 312). Recently, corpus linguists have also noticed that translated language shares certain features with what has been called “constrained language” as translation “involves bilingual language activation and is circumscribed by a previously produced text” which in turn cognitively constrains the language in a conspicuous manner (Kruger and van Rooy 2016, 27). Lower lexical diversity has, for example, been observed for the language produced by second-language (L2) learners (Mifková 2019, 39).

This thesis aims to investigate one candidate for such a constrained language: English versions of the web pages of Czech universities, namely selected three British university websites and the English versions of three Czech university websites. The prior investigation of the circumstances in which English versions of web pages of Czech universities are created suggests that they, at least to some extent, depend on the original Czech versions (are translations), and that the authors/translators are not always native speakers of English. First, it needs to be examined whether the English versions are less lexically richer than texts written originally in English. Then, the two corpora will be explored through keyword analysis. It is hypothesized that the original English websites will lexically richer and include more unique keywords (those appearing only in one corpus and not the other), and that there will be statistically significant differences between the amounts of lockwords (the most frequent words of both corpora).

Section 1 introduces the use of Corpus Linguistics in Translation Studies and in the studies of constrained language in general. This includes the use of corpora in Translation Studies and the issues and means of creating the web corpora. Sections 2 and 3 are concerned with a review of literature and theoretical framework concerning the two aspects of a lexical analysis: the former discusses the problematic of defining the lexical richness, various ways of calculating it, and their reliability for purposes of an exact statistical analysis; the latter expands on the notion of keyness, keywords and different types of them, and the issues of calculation and statistical methods for identifying and analysing them.

Section 4 describes the process of creating two comparable corpora of the Czech and English university website texts via Sketch Engine. It introduces the texts included in the corpora and the process of compiling them while dealing with the question whether the texts included in the English versions of the Czech university websites are translations or not. What comes next is the lexical analysis of the compiled corpora with a focus on differences in lexical richness and keywords. Finally, the thesis is concluded, and its findings are summarized.

# 1 Corpus Linguistics

One of the more computer-oriented branches of linguistic studies is Corpus Linguistics which analyses language using corpora. According to Johansson, corpora are “bodies of texts assembled in a principled way” (1995, 19) which allow linguists to study language samples based on empirical evidence in the form of concrete data. Corpus Linguistics, therefore, uses quantitative statistical methodology and focuses on frequencies of words and phrases in corpora (Březina 2018, 3).

There is a distinction to be made between corpus driven and corpus based research. While the former bottom-up approach “rejects the characterisation of Corpus Linguistics as a method” and sees corpora only as a “source of our hypotheses about language”, the latter top-down approach considers the corpus analysis to be a valid method and a tool for exploration of linguistic theories (Tognini-Bonelli, 2001).

The modern Corpus Linguistics date back to the early 1960s when the first one-million-word computer corpora were created, such as the Brown Corpus (Leech 1991, 10), but the specialized branch really started in later decades with the rise of multi-million-word corpora including the Longman-Lancaster English Language Corpus and Birmingham Collection of English Text (Kruger 2002, 71). Corpus Linguistics has been on a steady rise since 1980s thanks to accelerated technological progress and the raise of awareness of the multifunction of corpora studies (McEnery and Xiao 2008, 1). Corpora may be used for analysing both spoken and written texts by anyone from language students to translators.

## 1.1 The use of corpora in Translation Studies

There are three main approaches of Corpus Based Translation Studies: descriptive, theoretical, and applied. While the Descriptive Corpus Based Translation Studies focus on systemic, functional, and target-oriented approach to translation (Kruger 2002, 77), their theoretical counterpart relies purely on the corpus, which is seen as “the only repository of relevant denotative, connotative, pragmatic, and ideological meanings” (Laviosa 2004, 38). The applied approach represents the practical use of corpora for translation and interdisciplinary purposes (Ibid 2004, 40).

Although it may appear that the field of Corpus Based Translation Studies is primarily interested in multilingual parallel corpora, i.e. corpora containing both source texts and their translations aligned, monolingual comparable corpora consisting of translation (or at least non-native) and non-translation English are seen as "the major methodological advance" of corpus studies (Pym 2008, 321–322) and are often preferred to their multilingual counterparts which perpetuate a prescriptive, source-oriented view of translations as inferior to the originals almost by definition (Bernardini 2011, 2).

Representativeness and comparability of the texts in the corpora have to be also considered (Chesterman 2003, 214–215). Many criteria should but accounted for, such as size, text type, and genre (Chlumská 2014, 228). Kenny says that corpus design criteria depend on use of the corpora and representativeness of a specific type of language production (1998, 50), and Chlumská adds, that question of representability largely relies on whether parallel corpora or comparable corpora are analysed as the former depend on the choice of the originals and its translations, and the latter rely on the criteria for the text selection (2014, 228). The more criteria the texts meet, the more comparable they are, but also the more criteria are applied, the harder it is to create a suitable corpus.

## 1.2 Corpus based research of constrained language

Kruger and van Rooy say that Corpus Based Translation Studies often investigate "linguistic features that typify translated language as a distinct variety" (2016, 27). According to Baker, such features include simplification of the message, language, or both; explicitation, including additional information; normalization or conservatism to conform to the target language; and levelling out or homogenisation, meaning focusing more on the centre than the fringes of the continuum (1996, 176–177). These are referred as "translation universals", but as more and more corpus based studies are gradually disproving their universal status, they are also given more neutral names, such as properties or tendencies (Cvrček and Chlumská 2015, 311).

Translation language is hence seen as “constrained” because it is affected by such universals. Nevertheless, Lanstyák and Heltai claim that all communication is constrained to a degree, but some language variants are affected more because of certain communicative contexts where constraints play a conspicuous role (2012, 100). Kruger and van Rooy add that both L2 and translation language are constrained in a similar manner as they are both produced by means of bilingual language activation in situations of language contact (two languages colliding) (2016, 27) and therefore both succumb to similar psycholinguistic and contextual constraints (31). It is hypothesized that both translated and L2 texts are lexically poorer because of such constraints (Ibid 2016, 39).

## 1.3 The issues and advantages of creating the web corpora

Using website texts for corpora creation has many advantages as there is plenty of data to choose from, it is already in electronic form and easily accessible (Liu and Curran 2006, 233). However, there also some challenges, because one has no control over the composition and design of the web, the texts often mix different strains of language, genres, and text types (Zanettin 2012, 56–57) and include spam, duplicated and machine-generated content (“Sketch Engine” 2021), and lack of punctuation, which are all issues that can be solved by efficient corpus tools (Liu and Curran 2006, 234). Translators benefit from specialized web corpora as they may be used “as a source of corpus based descriptive translation studies as well as for translator training and practice,” because they can be easily assembled for particular assignments and later be disposed of (Zanettin 2012, 62, 64).

## 1.4 Sketch Engine

The online corpus analysis tool Sketch Engine was developed by Pavel Rychlý and Adam Kilgarriff. It is both the web service for exploring other user’s corpora and the software for creating, uploading, and managing your own (Kilgarrif et al. 2014, 8). Sketch Engine is widely used by language teachers and textbook creators, translators, terminologists, and language technology companies (Ibid 2014, 16).

The online corpus tool consists of many preloaded specialized types of corpora, including large general language corpora, parallel corpora, teaching corpora, historical corpora. etc. (Ibid 2014, 23). However, one of the biggest feats of Sketch Engine is the ease with which it allows its many users to create and manage their own corpus which is also the reason why it was chosen as a perfect tool for the thesis.The Sketch Engine also offers links to various tools or even some built-in ones which are useful for creation of web corpora, such *JusText* for removing unwanted content or *WebBootCaT* for deleting duplicated and other undesired content ("Sketch Engine" 2021).

# 2 Lexical richness

## 2.1 The issue of defining and calculating the lexical richness

The concepts of lexical richness and lexical diversity are considered interchangeable by many, nonetheless, some experts see them as something different (Wang 2014, 66). Other names for the concept include e.g. lexical variation (Laufer and Nation, 1995), vocabulary richness (Kubát and Milička, 2013), and lexical variability (Mazgutova and Kormos, 2015). According to Jarvis (2013, 15), lexical richness is calculated to measure a range of vocabulary in a text. It is affected by many factors. According to Laufer and Nation, it is influenced by vocabulary size and communicative purpose besides other things (1995, 308).

The authors also mention measures of lexical richness such as lexical originality, which concerns itself with words used by only one writer in a group, and lexical sophistication, which includes the "advanced" words in the text (Laufer and Nation 1995, 309). These measures are calculated by dividing the amount of a certain type of tokens (usually multiplied by hundred if a result expressed as a percentage is needed) by the total amount of tokens. Calculating lexical originality requires the amount of tokens unique to one writer, and lexical sophistication requires the amount of advanced tokens whose definition is influenced by both the levels of education and language proficiency and by the measures set by the researchers. However, lexical originality and lexical sophistication suffer from some reliability issues. The former is completely dependent on writer's performance relative to others who authored other pieces in the corpus, and the latter cannot be reliable when the definition of the advanced words is never set in stone.

Therefore, the most common way of calculating the lexical richness is a method of type token ratio (TTR) which is calculated as the amount of types divided by the amount of tokens, that can be multiplied by a hundred if a result in percentage is desired. "A type is a unique word form in the corpus" and "a token (running word) is a single occurrence of a word form in the text" (Březina 2018, 39), which basically means that the amount of different words is divided by the complete amount of words in a text:

$$\text{type token ratio} = \frac{\text{amount of types } (\times\ 100)}{\text{amount of tokens}}$$

The higher the result (percentage), the richer a lexicon of a text is, but it is unrealistic to expect a number close to 1 (or 100 percent) as it would mean that (almost) every word form occurs only once. The reasons why TTR is used so often are the ease of calculation (basically every processing tool has information about the amount of tokens and types), its straightforward interpretation and low computational complexity (Cvrček and Chlumská 2015, 315).

However, it has to be noted again that TTR method may not always be reliable as its results are often affected by differing length of the text because the longer it is, the more words repeat (Laufer and Nation 1995, 310). Comparing texts of different lengths then can prove to be rather inconclusive. Another possible issue is reflected in some preliminary experiments which suggested that TTR is also affected by the type of the text, e. g. the average TTRs of fiction and journalistic texts of the same length may be significantly different (Cvrček and Chlumská 2015, 316), because journalistic texts are usually not as lexically diverse as their fictional counterparts (Knittlová 2000, 159). This should also not pose as a problem because both thesis corpora include similar text types.

## 2.2 Other methods of calculating lexical richness

To avoid the flaws of TTR, its standardized version (STTR or sometimes sTTR), which is also referred to as mean segmental type token ratio (MSTTR), can be used instead. According to Březina, STTR or MSTTR is calculated by dividing a text into same-size (usually 1,000 tokens or words) segments and then calculating TTR for each one to obtain its mean value (2018, 58). Such a method is better than TTR because it is not affected by text size, therefore it can be used for the lexical richness analysis of the differently sized parts of the corpora.

Another method is called moving average type token ration (MATTR), which is similar to STTR, but "instead of dividing the text into successive nonoverlapping segments, [it] uses an overlapping window smoothly moving through the text," which makes it “a more robust measure of lexical richness than STTR because it takes into account all possible segmentations of the text” (Březina 2018, 58). Other means of lexical diversity include moving window type token ratio (MWTTR) (first proposed by Köhler and Galle in 1993, but not called as such), moving window type token ratio distribution (MWTTRD) (Kubát and Milička, 2013), or TTR scaling (zTTR) (Cvrček and Chlumská, 2015). If the TTR and STTR results show a statistically significant difference in lexical richness between translation and non-translation English, it can be assumed that one is more lexically variable than the other.

## 2.3 Lexical richness and translation

According to Fang and Liu, lexical richness in translations should not be neglected because “vocabulary choice of translators has direct impact on the quality and readability of translation” (2015, 54). Nevertheless, research has shown that translated texts tend to be not as lexically rich as non-translated texts. The study by Laviosa (1998) on Translation English Corpus and British National Corpus proved the latter to be more lexically diverse than the former. Berman even describes the process of quantitative impoverishment which reflects a loss of lexical richness (1985, 291). Cvrček and Chlumská agree that lexicon of translated texts tends to be less lexically variable (2015, 312).

There are several reasons for this. Laviosa says that translations include relatively more content words than grammatical words than non-translations and that the most frequent words repeat more in translations than in non-translations (1998, 9). Berman mentions the simultaneous impoverishment and prolonging of the translated texts because signifier and signifying chains are proliferated while, on the other hand, other explicate and decorative signifiers or grammatical elements are added (1985, 291). Cvrček and Chlumská claim that this phenomenon relates to simplification of translations (2015, 312).

# 3 Keyness and keywords

The notion of "keyness" has always been a source of much discussion. Bondi says that while there is a disagreement whether it serves as a key to interpret a text or a culture, its study is central to the area of Corpus Linguistics (2010, 3). Scott even suggests that the "key" metaphor represents unlocking the layer which is otherwise inaccessible and enabling the possibility of seeing something which was hidden before (2010, 44) but he also notes it is more likely a simple pointer in the right analytical direction (Ibid 2010, 56).

According to Stubbs, there are three different approaches to "keywords" based on different notions of keyness (2010, 22–23). The first one is concerned with culturally, socially, and politically relevant words or phrases and is more of a qualitative nature and relevant rather to social sciences than linguistics. The history of this approach goes back to beginning of 20th century to German lexicographers, but this idea was elaborated on in 1930s with Firth's "pivotal words" (1935, 40) and then with *Keywords: A Vocabulary of Culture and Society* (1976) by Raymond Williams who created a list of about 120 words that he chose based on his intuition, education, and political analysis, the historical shift in their meaning, the spread of their use, and their ability to create and conceptualize categories (Stubbs 2010, 23–24). Nevertheless, even though such a keyword analysis tends to reveal a correlation between the social aspects of language and semantics, the meaning of keywords may not always relate to the social world (Ibid 2010, 39–40).

The second approach to keyness focuses more on phrases and collocations than on singular words and is more fit for corpus driven studies than for corpus based ones. It was first proposed by Gill Francis who saw meaning rather in various lexico-grammatical patterns which do not include any one essential word (1993, 155). While it is quite lexically challenging, the bottom-up empirical analysis may help to assess and evaluate meanings and contained cultural knowledge which is not intercepted by a fixed expression but only delineates a concept that captures a subjective perception of the social world conveyed by "extended lexical items which express much more complex and subtle evaluative acts" (Stubbs 2010, 29).

The third approach defines the keyness as “a textual manner” (Scott and Tribble 2006, 65) that is based on distribution and frequency of the content words, i.e. words carrying the semantic content of the text. Although this concept is concerned with keyness from the text perspective, it does not relate to text segmentation. Al-Rawi also adds: “Analysing text/texts, using keywords, does not depend on the meaning of the words themselves, but on their meaning within the context” (2017, 368), which needs to be borne in mind during the analysis.

## 3.1 Keywords and their types

As the notions of keyness vary, the relationship between qualitatively-defined and quantitatively-defined keywords needs to be defined. Although the latter can signify the cultural, social, or political concepts featured in the text, after the initial quantitative analysis, the qualitative analysis has to follow to establish and prove such a significance (Culpeper and Demmen 2015, 1).

Sketch Engine is endowed with a corpus analysis tool called “Keywords” which uses one corpus as a focus (also known as “a target”) and the other as a reference to compare contained terminology. Březina explains that “keywords are words that are considerably more frequent in one corpus than in another corpus; (..) that are typical of the corpus of interest when compared to another corpus” (2018, 79–80). Culpeper and Demmen stress that although the size of the reference corpus may play a role, it is typically its content which is more important (2015, 6).

Keywords can be vital for comparison and identification of typical vocabulary within a text (Groom 2010, 63) and, therefore, also for finding major lexical differences between two texts of the same language and genre. However, a minimum frequency cut-off parameter should be applied during analysis, because keywords may include, e. g. proper nouns and other rather localised phenomena (Culpeper and Demmen 2015, 7).

According to Scott, there are two types of keywords: positive keywords (used more often in the focus corpus than in the reference corpus) and negative keywords (used more often in the reference corpus than in the focus corpus) (2013, 96). Consequently, positive keywords of the focus corpus are negative keywords of the reference corpus and vice versa. Březina adds that there are also lockwords which are used in both corpora with a comparable frequency (2018, 80). Scott and Tribble also mention two more types: those related to the text content (open-class keywords), and those related to the text style (closed-class keywords) (2013, 196). Groom says that the latter trend to be omitted during analysis as keywords are usually used for analysing the content rather than the style (2010, 61). Al-Rawi argues that the matter is not so simple as some keywords may be both open and closed-classed (2017, 368). Groom agrees and avers that while closed-class keywords show little or no information about values and meanings expressed in a corpus as a whole, their contextual analysis may reveal indicated meaning of some texts and are, therefore, usually much more useful than even some open-class keywords such as proper nouns (although this depends on the type of the analysis and the corpus) (2010, 63–70).

## 3.2 The calculation method for identifying keywords

The wordlists of a focus and a reference corpora are compared using a frequency based statistical significance test (Culpeper and Demmen 2015, 7). Keywords are detected by counting occurrences in both corpora, dividing both numbers by the amount of words in those corpora, optionally multiplying it by a thousand or a million (depending whether the frequency per thousand or million is desired), and, finally, dividing one number by the other to see a ratio, which is, then, used to sort words and discover those which are more present in one corpus than the other.

Nevertheless, there are several issues with this method of calculation. Firstly, as it was already mentioned, the used text types and corpus construction matter hugely, because if two vastly different corpora are chosen to be compared, the resulting keywords will be probably either confusing or useless. Secondly, one word being a sole focus of one text in the corpora may also confuse the results. Thirdly, the division by zero is mathematically impossible which means that the words that are completely absent from the reference corpus cannot be the positive keywords of the focus corpus and vice versa. This is usually solved by “adding one” to every frequency which solves the issue conveniently (Manning and Schütze

1999, 203). Sketch Engine implements the “add one” method automatically (Kilgarrif 2009, 3).

Other statistical problems are posed by Scott (2010, 48–50). Firstly, statistics represents a game of chance, which means that every analysis runs a risk of spuriously getting odd results of statistical significance. Secondly, the keywords are represented by the notion of power law as some appear in enormous numbers but others only once or twice. It is, therefore, quite difficult to assess what the likely statistically significant results of the keyword analysis will be. Thirdly, the keyword likelihood is computed for every word-type individually, which means that the statistically based likelihood of getting a certain set of keywords cannot be calculated. This only confirms that keywords simply point in the direction of possible interpretations.

# 4 Corpus based study

## 4.1 Corpus design

For the aim of this study, two web corpora were compiled — one on the basis of texts from the English versions of three Czech university websites; Palacký University (UPOL), Charles University (UK), and Masaryk University (MUNI); and the other including texts from the websites of three British universities; Cambridge University (CAM), Oxford University (OXF), and University College London (UCL). The former mentioned will be henceforth referred to as the CUWC (i.e. "Czech Universities Web Corpus") and the latter as the EUWC (i.e. "English Universities Web Corpus").

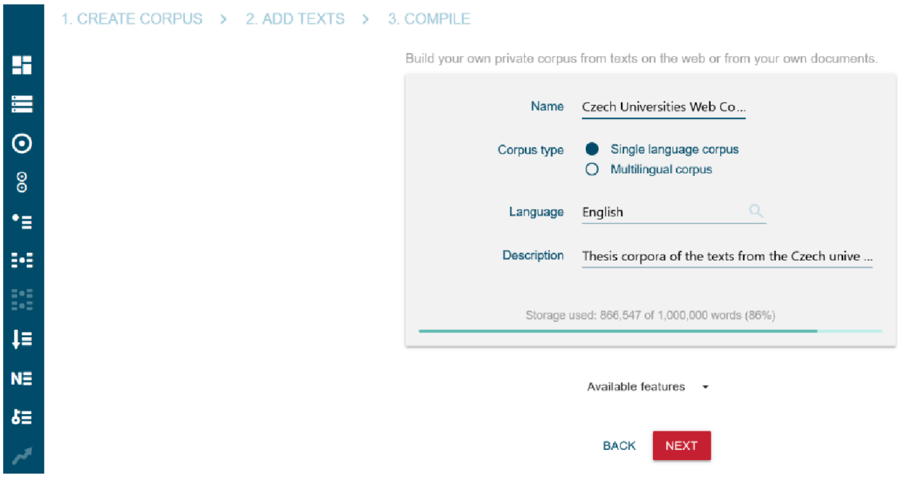
At first, both web corpora were supposed to be compiled to compare the non-translation English of the EUWC and the translation English of the CUWC focusing on translation universals, however, after discussing the webmasters of the Czech universities, it emerged that not all texts of the English versions of the websites are translations as some of them are written specifically for the English versions, predominantly not by L2 English speakers, although often at least edited by the native English speakers (Daniel Agnew of UPOL, Jan Velinger of UK, and Břetislav Regner of MUNI, personal communication, May 1, 2020). Such an approach also differs from faculty to faculty which only makes matters more complicated. For that reason it was decided to focus on lexical richness and keywords.

The text types vary slightly, but mostly belong to the administrative and technical discourses, as they are described by Knittlová (2000, 127, 137), with some exceptions (e.g. advertising purposes) (Ibid, 175). The texts compiled in the corpora are examples of coherent texts written in sentences and are consisting of information for website visitors, including students and employees.

## 4.2 Corpus compilation

Both corpora were compiled with the use of the web version of Sketch Engine (Figure 1). The texts, which were extracted from the university websites manually, divided into three text files for each corpus, and cleansed of unwanted and duplicated content using *WebBootCaT*,[1] were then uploaded using the option "I have my own texts" (Figure 2). As the upload was finalized, the corpora were set to be automatically compiled (Figure 3). The procedure was the same for both the CUWC and the EUWC corpora.

***Figure 1: Creating the corpus***



1. CREATE CORPUS > 2. ADD TEXTS > 3. COMPILE

Build your own private corpus from texts on the web or from your own documents.

Name: Czech Universities Web Co...

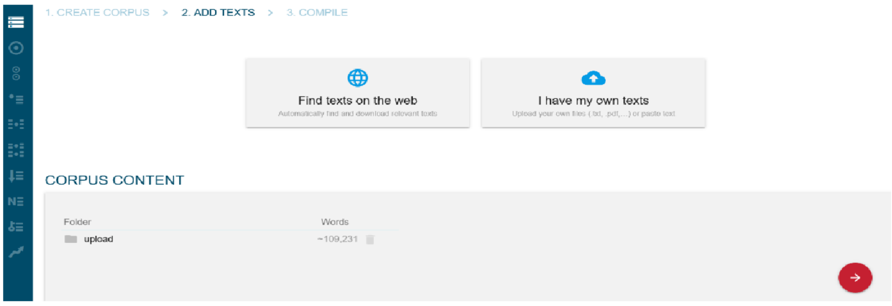
Corpus type: ● Single language corpus ○ Multilingual corpus

Language: English

Description: Thesis corpora of the texts from the Czech unive ...

Storage used: 866,547 of 1,000,000 words (86%)

***Figure 2: Uploading the text files***



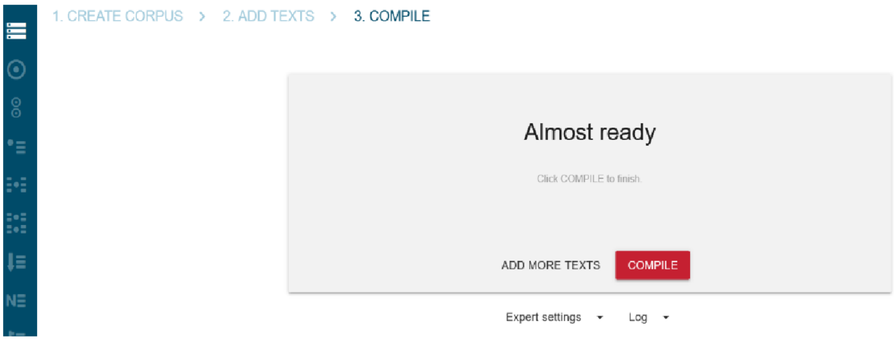
1. CREATE CORPUS > 2. ADD TEXTS > 3. COMPILE

Find texts on the web
Automatically find and download relevant texts
I have my own texts
Upload your own files (.txt, .pdf,...) or paste text

CORPUS CONTENT

| Folder | Words |
|---|---|
| upload | ~109,231 |

***Figure 3: Compiling the corpus***



1. CREATE CORPUS > 2. ADD TEXTS > 3. COMPILE

Almost ready

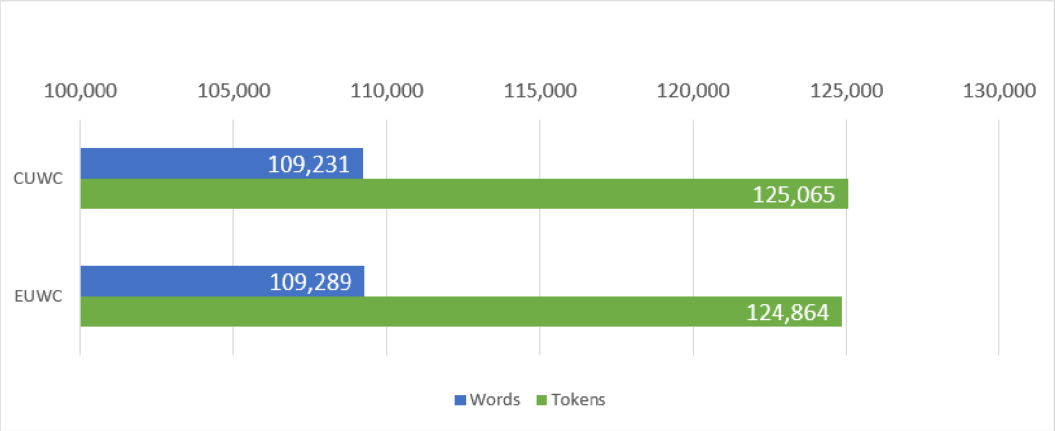
Click COMPILE to finish.

[1] *WebBootCaT* comes as a part of Sketch Engine corpus management tools.

## 4.3 Corpus size

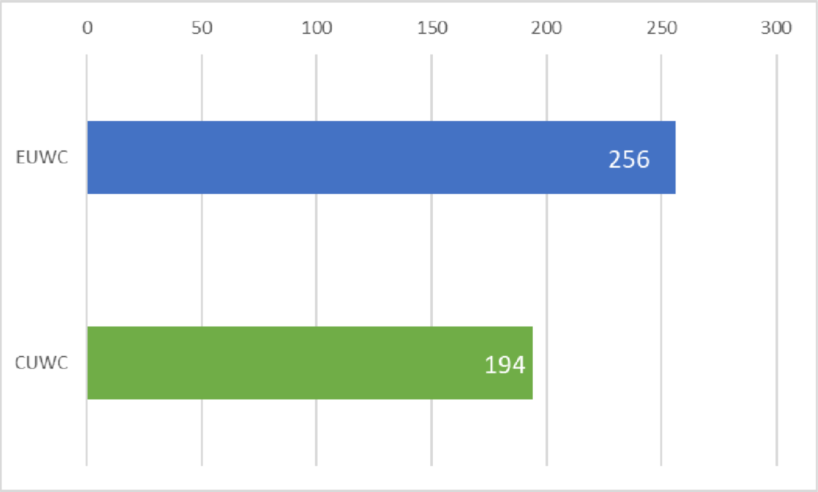
One of the objectives was to achieve a balanced number of words/tokens in both corpora to assure that the CUWC and the EUWC are comparable in size. The CUWC corpus includes 109,231 words and 125,065 tokens and the EUWC corpus includes 109,289 words and 124,864 tokens (see Graph 1).

***Graph 1: Number of tokens and words in each corpus***

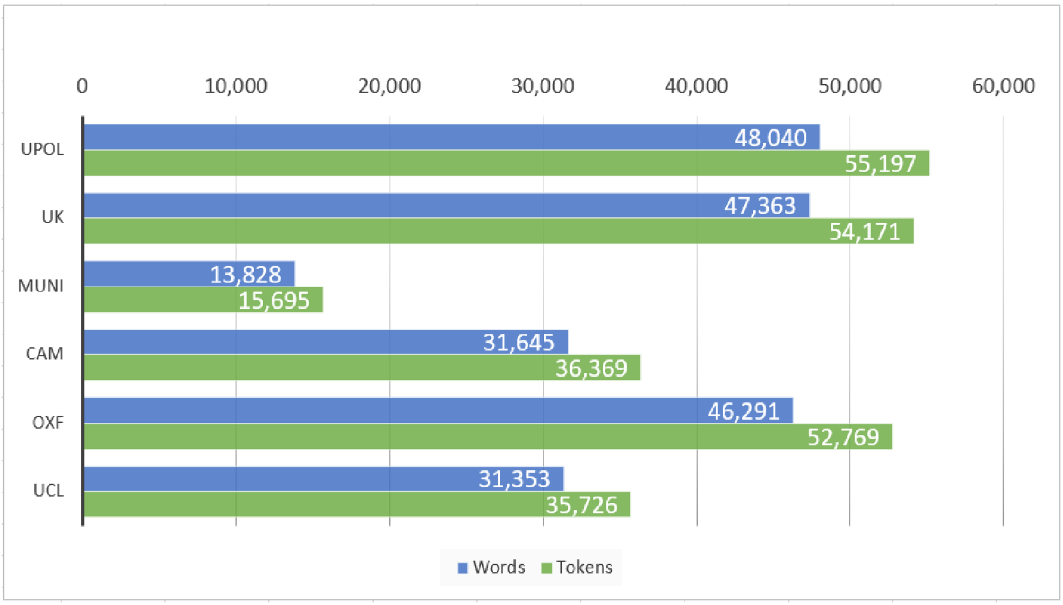


Another significant feature of both corpora is the size of the individual university website parts. Although both the CUWC and the EUWC are of similar size, not every university website featured a similar amount of English language texts. The CUWC includes 194 texts while the EUWC includes 256 texts (see Graph 2). Furthermore, the CUWC contains the biggest two parts of both corpora coming from the English versions of the UPOL website with 48,040 words (55,197 tokens) and of the UK website with 47,363 words (54,171 tokens). The English version of the MUNI website was able to provide only 13,828 words (15,695 tokens). The British universities websites are a bit more size-comparable, even though OXF website part with 46,291 words (52,769 tokens) is quite bigger that the other two. The sizes of the CAM website part with 31,645 words (36,369 tokens) and of the UCL website with 31,353 words (35,726 tokens) are, however, almost identical (see Graph 3).

*Graph 2: The number of texts in each corpus*



*Graph 3: Number of words and tokens by university*



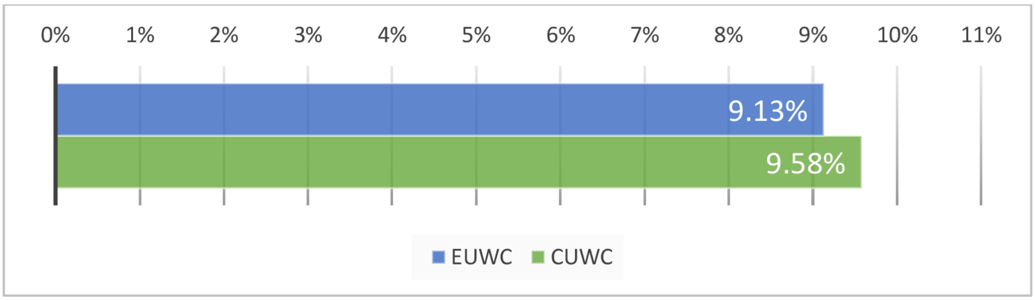
Both corpora are comprised of various texts that were featured on the university websites in Spring 2021. These include general information about universities and their faculties, departments, and colleges, their history, research and other tips and requirements for undergraduate and postgraduate students, alumni, employees, etc. The aim was to feature similar types of information for every university and faculty, yet, not every faculty and college section included the same amount of texts about their research and history, and some sections did not even include such information at all. The pictures and their descriptions, the web links, inserted tables, and majority of texts excerpts written in points were excluded due to facilitating the subsequent automated corpora analysis and keeping the flow of the texts.

## 4.4 Lexical richness analysis

One of the thesis aims is to compare lexical richness of the EUWC and the CUWC corpora. As at it was already mentioned, while the EUWC includes exclusively non-translation texts, the CUWC is compiled of indistinguishable combination of both translated texts and non-translated texts written by Czech native speakers. It is hypothesised that the texts from English university websites may be lexically richer than their Czech counterparts, given the inclusion of probably less lexically variable translations from the Czech language into English and L2 English texts in the CUWC.

To calculate lexical richness of the whole corpora, the analysis will start with a simple TTR measure, using the number of unique word forms (types) in the corpus and the number of all word tokens in the corpus as calculated by Sketch Engine. The calculation focuses on the number of word tokens (number of words in the corpus) because the number of all tokens include punctuation which is rather useless for the measure of lexical richness. According to the Sketch Engine corpus info, the EUWC consists of 9,985 types and 109,289 word tokens and the CUWC of 10,470 types and 109,231 word tokens. If the amount of types (multiplied by a hundred to get a percentage) is divided by the amount of word tokens, the TTR index for EUWC is 0.0913 (9.13%) and 0.0958 (9.58%) for the CUWC (see Graph 4).

***Graph 4: Comparison of lexical richness by measure of TTR***



In the other words, the lexical richness index of the EUWC is by 0.0045 (0.45%) lower than the CUWC one, and the result seems to be significant at p<.001.[2] This goes against the hypothesis stated in section XX: translated texts and texts written by non/native speakers of English were expected to have a lower lexical richness than original texts written in English by English native speakers. I believe the reason behind this rather counter-intuitive finding is the measure used for calculating lexical richness, TTR.

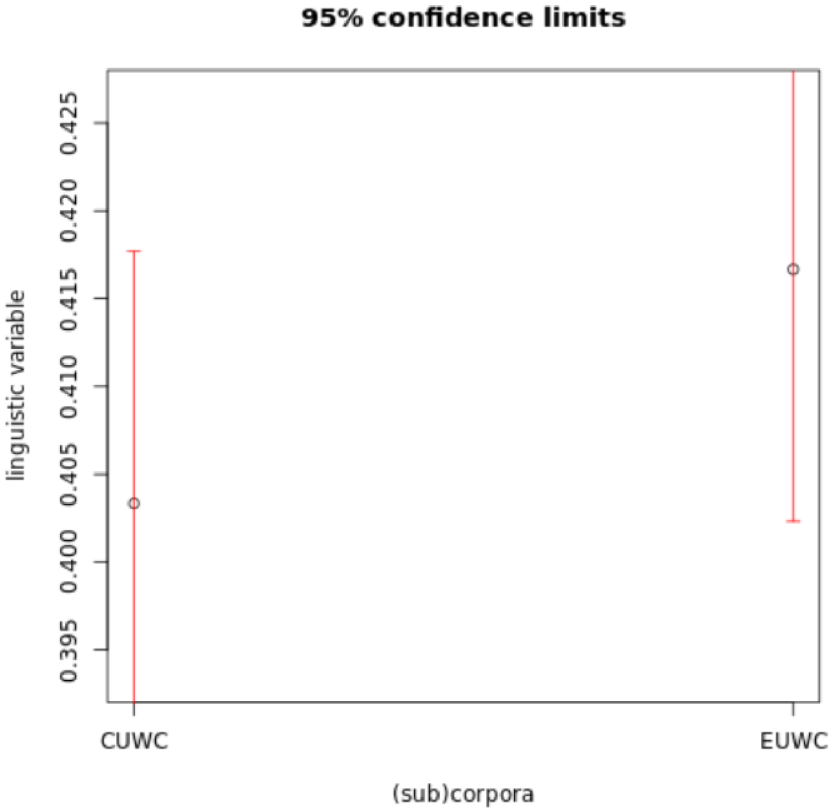
As has been stated earlier, the TTR measure as calculated above relies on numbers of words and tokens supplied by Sketch Engine for the whole corpora. Internal variation is thus ignored, yet internal variation may play a major role: it has been stated that longer texts have a lower TTR. One possible way to get round this problem is using the STTR measure, described in Section CC, and easily calculated by tools available via Lancaster Stats Tools online (Březina, 2018),[3] more specifically by the Vocabulary: Frequency and Dispersion Tool.

However, the STTR cannot be calculated for the whole corpora since their sizes exceed the 500,000-character limit for STTR calculation set by Lancaster Stats Tools. For that reason, the corpora were divided into parts consisting of texts downloaded from individual universities websites and the STTR was calculated manually for each university website separately (see Appendix). The Graph Tool was then used to compare the STTRs of the two corpora. As it can be seen in the Graph 5, the results suggest a different picture than the one based on TTR.

[2] According to *Corpus Frequency Wizard* tool (Baroni and Evert, 2017), which is available online at http://sigil.collocations.de/wizard.html.

[3] http://corpora.lancs.ac.uk/stats/toolbox.php

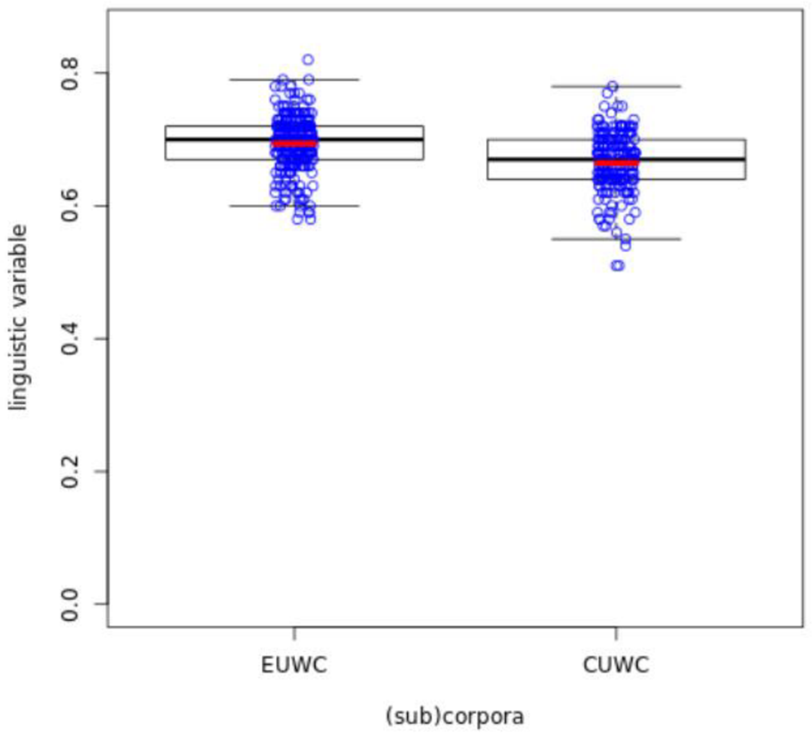
*Graph 5: The STTR values of the CUWC and EUWC*



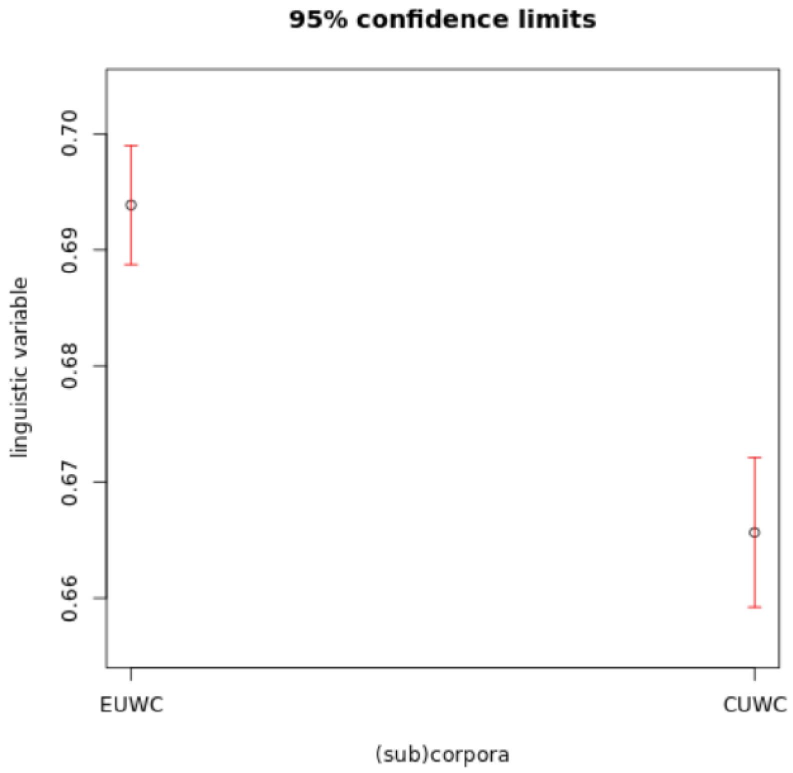
Using the STTR results, lexical richness of the UK and MUNI websites is at 0.40 (40%), of the UPOL and UCL websites at 0.41 (41%), and of the CAM and OXF websites at 0.42 (42%). It can be asserted that the mean STTR value for the CUWC is 0.403 (40.33%) and 0.416 (41.66%) for the EUWC, which means that the latter is by 0.013 (1.33%) lexically richer than the former. This, unlike the TTR results, confirms the hypothesis, although the difference is not statistically significant as the two vertical lines on the Graph 5 overlap.

This is because the more data is taken into equation, the more is the difference apparent as both corpora are compiled of hundreds of texts of various sizes linked from the main university pages. Although the EUWC includes 256 texts and the CUWC only 194, their sizes vary with some consisting of only a hundred tokens and others featuring even several thousand. Therefore, the most accurate results can be obtained by calculating STTR of every text with the normalisation basis set on 100 and comparing them (see Graph 6 and Graph 7).

***Graph 6: STTR of the individual texts of both corpora***



***Graph 7: Inference graph of STTR values of the individual texts***



Graph 6 shows individual values of SSTR for the two corpora. It pictures mean values that demonstrate a bigger lexical richness for the EUWC. Nevertheless, it also reflects noteworthy differences between individual texts. STTR values of the CUWC range from 0.51 (51%) to 0.78 (78%) (the difference of 0.21) with the median value between around 0.66 (66%) and 0.67 (67%). The EUWC STTR values range from 0.58 (58%) to 0.82 (82%) (the difference of 0.24) with the median value between around 0.69 (69%) and 0.7 (70%). Both corpora feature more texts that are of above median value (90 in the CUWC and 107 in the EUWC) than those of below median value (79 in the CUWC and 100 in the EUWC) or of median value (25 in the CUWC and 49 in the EUWC). Graph 7 shows the difference between lexical richness levels even more clearly. The difference shows to be significant as the two vertical lines do not overlap.

## 4.5 Keyword analysis

For the purpose of the thesis, the corpora are used both as the focus corpus and the reference corpus for each other. All tokens in the Sketch Engine corpora are automatically converted to lower-case to make the search case insensitive ("Sketch Engine" 2021). As is, the sample list of 20 most common positive keywords in the CUWC and EUWC includes predominantly proper names, acronyms, and regional vocabulary, that are mostly well-reflective of the selected universities and their regions (see Table 1).

*Table 1: The sample table of positive keywords in both corpora*

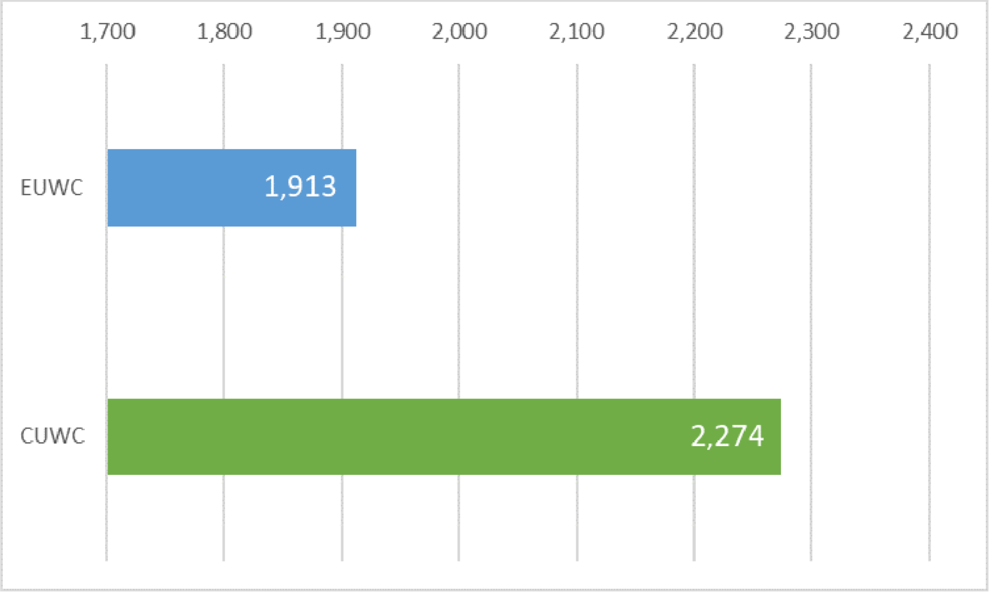
|  | CUWC | EUWC |
|---|---|---|
| 1 | olomouc | ucl |
| 2 | palacký | ucas |
| 3 | masaryk | world-leading |
| 4 | rector | bloomsbury |
| 5 | prague | barlett |
| 6 | czech | tripos |
| 7 | cu | bodleian |
| 8 | mu | england |
| 9 | czechoslovak | henry |
| 10 | hussite | lots |
| 11 | brno | christ |
| 12 | slovak | green |
| 13 | ga | whilst |
| 14 | czk | divisions |
| 15 | youth | professorship |
| 16 | pedagogical | postgraduates |
| 17 | optic | tutorials |
| 18 | bilateral | slade |
| 19 | hap | wrangler |
| 20 | moravia | parents |

Proper names include names connected with universities (Palacký, Masaryk, Bodleian), cities (Olomouc, Prague, Brno), regions (Moravia), countries (Czech, Czechoslovak, Slovak, England), or even religion (Hussite, Christ). Interestingly, the EUWC features Bloomsbury, Henry, and Slade as keywords. Concordances show that Bloomsbury is featured in reference to the London district, Henry in reference to many famous men mentioned (especially kings), and Slade in reference to the Slade School Fine Art of UCL. The acronyms are connected with universities (UCL, CU, MU), academic bodies (GA being Grant Agency and UCAS The Universities and Colleges Admissions Service), currency (CZK) or, interestingly enough, the academic employee evaluation system (HAP).

All of this says little about the lexical differences between the corpora, though, as the thesis is more interested in finding overall lexical dissimilarities which are not based just on regional distinctions. The more proper keyword list should omit such regional keywords (which are specific to the countries of origin and to the selected universities themselves, therefore it is expected that such words do not appear in the reference corpus) and concern itself with only non-site-specific vocabulary. The simplest way of doing this is to omit upper case items and setting minimum length for keywords to at least three signs. The table is adjusted to fit those demands.

By setting the keywords focus on the rarest of words, the list of words that are featured in one but not the other corpus is created. Such words are called "unique keywords" for purposes of the thesis. Altogether, the CUWC includes 2,274 of unique keywords that are not featured in the reference corpus while the EUWC includes 1,913 of those (see Graph 8). Although such a measure is not a measure of lexical richness, it is notable that the CUWC consists of significantly more unique keywords than the EUWC at $p < .001$.

***Graph 8: The difference between the amounts of unique keywords***



When analysing the lockwords, the list of predominantly closed-class items is obtained. If the wordlists of the most frequent words in both corpora is analysed, the CUWC features more open-class items than the EUWC. Both wordlists also include some keywords that are not included in the other or only scarcely. For example, the CUWC wordlist features Czech or Charles, while the EUWC features UCL or Oxford. Therefore only lower-case items should be considered. The Table 2 includes the sample wordlist of top 30 most frequent lower-case words in both corpora with their absolute frequencies, the frequencies in the other corpora, the difference $x^2$, and the statistical significance p.

***Table 2: The table of most frequent lower-case words in both corpora***

| | CUWC | Frequency in the CUWC | Frequency in the EUWC | $x^2$ | p | EUWC | Frequency in the EUWC | Frequency in the CUWC | $x^2$ | p |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | the | 8,964 | 6,353 | 469.315 | < .001 | the | 6,353 | 8,964 | 469.315 | < .001 |
| 2 | of | 5,589 | 3,993 | 273.424 | < .001 | and | 5,012 | 4,492 | 30.336 | < .001 |
| 3 | and | 4,492 | 5,012 | 30.336 | < .001 | of | 3,993 | 5,589 | 273.424 | < .001 |
| 4 | be | 3,234 | 3,426 | 5.947 | < .05 | be | 3,426 | 3,234 | 5.947 | < .05 |
| 5 | in | 2,958 | 2,423 | 53.285 | < .001 | to | 3,100 | 2,050 | 219.892 | < .001 |
| 6 | a | 2,052 | 2,629 | 73.177 | < .001 | a | 2,629 | 2,052 | 73.177 | < .001 |
| 7 | to | 2,050 | 3,100 | 219.892 | < .001 | in | 2,423 | 2,958 | 53.285 | < .001 |
| 8 | university | 1,733 | 900 | 264.351 | < .001 | for | 1,569 | 1,392 | 10.875 | < .001 |
| 9 | for | 1,392 | 1,569 | 10.875 | < .001 | you | 1,213 | 377 | 442.666 | < .001 |
| 10 | faculty | 1,002 | 366 | 295.351 | < .001 | student | 1,204 | 837 | 66.767 | < .001 |
| 11 | student | 837 | 1,204 | 66.767 | < .001 | university | 900 | 1,733 | 264.351 | < .001 |
| 12 | research | 810 | 825 | 0.144 | insignificant | our | 890 | 164 | 501.979 | < .001 |
| 13 | with | 808 | 692 | 8.684 | < .01 | as | 867 | 690 | 20.305 | < .001 |
| 14 | on | 784 | 716 | 2.903 | insignificant | research | 825 | 810 | 0.144 | insignificant |
| 15 | by | 773 | 483 | 66.367 | < .001 | have | 814 | 545 | 53.574 | < .001 |
| 16 | at | 748 | 642 | 7.807 | < .01 | college | 732 | 39 | 624.134 | < .001 |
| 17 | study | 730 | 517 | 35.88 | < .001 | on | 716 | 784 | 2.903 | insignificant |
| 18 | as | 690 | 867 | 20.305 | < .001 | your | 693 | 263 | 193.944 | < .001 |
| 19 | programme | 620 | 260 | 146.394 | < .001 | with | 692 | 808 | 8.684 | < .01 |
| 20 | or | 561 | 640 | 5.217 | < .05 | at | 642 | 748 | 7.807 | < .01 |
| 21 | have | 545 | 814 | 53.574 | < .001 | or | 640 | 561 | 5.217 | < .05 |
| 22 | education | 522 | 173 | 174.175 | < .001 | we | 637 | 198 | 231.231 | < .001 |
| 23 | from | 470 | 625 | 22.003 | < .001 | from | 625 | 470 | 22.003 | < .001 |
| 24 | it | 422 | 360 | 4.675 | < .05 | that | 591 | 304 | 92.183 | < .001 |
| 25 | also | 390 | 315 | 7.670 | < .01 | study | 517 | 730 | 35.88 | < .001 |
| 26 | international | 385 | 221 | 43.687 | < .001 | will | 491 | 250 | 78.351 | < .001 |
| 27 | their | 383 | 355 | 0.947 | insignificant | by | 483 | 773 | 66.367 | < .001 |
| 28 | its | 378 | 299 | 8.885 | < .01 | academic | 386 | 378 | 0.076 | insignificant |
| 29 | academic | 378 | 386 | 0.076 | insignificant | faculty | 366 | 1,002 | 295.351 | < .001 |
| 30 | you | 377 | 1,213 | 442.666 | < .001 | it | 360 | 422 | 4.675 | < .05 |

Although the difference between the frequencies of both the definite article (the) and indefinite article (a) is statistically significant at p < .001, the CUWC contains way more definite articles than the EUWC. There are also statistically significant differences between the frequencies of some prepositions (*of*, *in*, *with*, *at*, *as*, and *by* are more frequent in the CUWC, only *from* is more frequent in the EUWC), conjunctions (*and* and *or* are both more frequent in the EUWC), and (auxiliary) verbs (*be* and *have* both more frequent in the EUWC). All in all, majority of differences are statistically significant except for the words *research*, *on*, *their*, and *academic*.

# Conclusion

Corpus based studies have been, are, and more than likely will be one of the most useful tools for the purposes of Translation Studies. Although it may seem that translators benefit from multilingual parallel corpora the most, monolingual comparable corpora have proven to be an equally great, if not better, contribution to the field as it helps to understand the difference between translation and non-translation language or, as in case of the thesis, between language strains coming from two geographically and culturally different places. While the translation universals analysis cannot be directly applied in this case, some of its methods may be utilized and, given the specific nature of such language strains, show rather interesting and unpredictable results.

Lexical richness and keyword analysis represent the two methods that found their use across the area of Corpus Linguistics and even beyond. How diverse the lexicon of a corpus is may indicate many attributes of included texts from their type to the questions whether they are translations or not. Keywords, on the other hand, even though they also may suggest an inclusion of constrained language, may have much deeper implications about the text as they can be analysed not only from a lexical, but also from the semantical, social, political, or cultural point of view. While both methods are seen as measures of Corpus Based Statistical Linguistics, their potential does not end there.

The aim of the thesis was to create and analyse two monolingual comparable corpora of the selected six Czech and English university (Palacký University – UPOL, Charles University – UK, Masaryk University – MUNI, Cambridge University – CAM, Oxford University – OXF, and University College London – UCL). As there was the issue that it was impossible to recognize which texts featured in the CUWC were translations and which non-translations, it was decided not to focus on translation universals analysis as whole but only focus on lexical richness and keywords to assess the difference between the two corpora. Given the inclusion of constrained language in the CUWC, it was hypothesized that the CUWC will be less lexically rich than the EUWC, that it will include less unique keywords, and that there will be statistically significant differences between the amounts of lockwords.

The type token ratio (TTR) method disproved the hypothesis as it showed lexical richness of the CUWC to be at 0.0837 (8.37%) and of the EUWC at 0,0799 (7,99%) with the statistically significant difference of 0,0034 (0.34%) at p < .001. Nevertheless, the TTR method is often found to be unreliable, therefore the standardized type token ratio (STTR) method was used, too, to compare the singular same-sized parts of both corpora. This showed interesting results such as the UPOL website being as lexically rich as the UCL website at 0.41 (41%) while UK and MUNI, and CAM and OXF being at 0.40 (40%) and 0.42 (42%), respectively. The overall STTR values proved to be at 0.403 (40.33%) for the CUWC and 0.416 (41.66%) for the EUWC with the statistically insignificant difference of 0.013 (1.33%). What followed was STTR calculation of every single text in the CUWC and EUWC. STTR values of the CUWC ranged from 0.51 (51%) to 0.78 (78%) (the difference of 0.21) with the median value between around 0.66 (66%) and 0.67 (67%). The EUWC STTR values ranged from 0.58 (58%) to 0.82 (82%) (the difference of 0.24) with the median value between around 0.69 (69%) and 0.7 (70%), and the overall difference proved to be statistically significant, ultimately proving the hypothesis.

The keyword analysis focused on unique keywords of both corpora, which are those that appear in only one corpus and not the other, and lockwords, which are included in both corpora the most frequently. After excluding proper names and acronyms from the equation, it was found that the CUWC contains significantly more unique keyword than the EUWC. When analysing the lockwords as the most frequent words in the wordlists of both corpora, it showed that there are statistically significant differences in frequencies of articles (more definite articles in the CUWC), prepositions, conjunctions, and (auxiliary) verbs.

Ultimately, the Corpus Based Translation Studies remain one of the most important and interesting fields of current Linguistics and it is hoped that this thesis, its findings, and the created corpora will be used for further research, e. g. focusing on the hapax legomema frequency, the open and closed class items ratio, and the phraseology by means of n-gram analysis.

# Shrnutí

Korpusové studie byly, jsou a s největší pravděpodobností i budou jednou z nejužitečnějších pomůcek pro translatologii. Ačkoliv se může na první pohled zdát, že překladatelé nejvíce těží z vícejazyčných paralelních korpusů, jednojazyčné srovnávací korpusy znamenají stejný, ne-li větší přínos oboru, protože napomáhají porozumět rozdílu mezi překladovým a nepřekladovým jazykem nebo jako v případě této práce mezi dvěma vzorky jazyka, které pochází ze dvou geograficky a kulturně odlišných míst. Přestože se v tomto případě nedalo využít přímé analýzy překladových univerzálií, na některé její metody vzhledem ke specifické podstatě těchto jazykových vzorků také došlo, přičemž odhalily zajímavé a nepředvídatelné výsledky.

Jedná se o analýzu lexikální bohatosti a klíčových slov, která se využívá napříč korpusovou lingvistikou i jinde. Rozmanitost slovní zásoby korpusu může naznačit různé vlastnosti zahrnutých textů od jejich typu po otázku, zdali se jedná o překlady či ne. Klíčová slova mohou zase nejen svědčit o přítomnosti fenoménu „constrained language“, ale i mnohem hlubších textových atributech, protože jejich analýza nemusí být jen lexikální, ale i sémantická, sociální, politická nebo kulturní. I když se obě metody vnímají především jako nástroje korpusové statistické lingvistiky, jejich potenciál zde zdaleka nekončí.

Cílem práce bylo vytvořit a analyzovat dvojici jednojazyčných srovnávacích korpusů textů vybraných šesti webových stránek českých (CUWC) a anglických (EUWC) univerzit se (Univerzity Palackého – UPOL, Univerzity Karlovy – UK, Masarykovy univerzity – MUNI, Cambridgeské univerzity – CAM, Oxfordské univerzity – OXF a University College London – UCL) zaměřením na lexikální bohatost a klíčová slova. Jelikož nebylo možné rozpoznat, které texty v CUWC jsou překlady a které nikoliv, práce se při hledání rozdílů mezi korpusy nezaměřila na překladové univerzálie jako celek, ale pouze na lexikální bohatost a klíčová slova. Kvůli výskytu „constrained language“ se předpokládalo, že CUWC nebude natolik lexikálně bohatý jako EUWC, že bude zahrnovat méně jedinečných klíčových slov a že analýza odhalí statisticky významné rozdíly mezi počty lockwords.

Metoda poměru typů a tokenů (TTR) hypotézu vyvrátila, protože u CUWC odhalila lexikální bohatost o výši 0,0837 (8,37 %) a u EUWC 0,0799 (7,99 %), což představovalo statisticky významný rozdíl o velikosti 0,0034 (0,34 %) s p < ,001. Metoda TTR se ovšem mnohdy považuje za nespolehlivou, a proto se k porovnání jednotlivých stejně velkých částí korpusů svolilo k metodě standardizovaného poměru typů a tokenů (STTR). Výsledky ukázaly, že hodnoty lexikální bohatosti webových stránek UPOL a UCL jsou 0,41 (41 %), UK a MUNI 0,40 (40 %) a OFX a CAM 0,42 (42 %). Průměrná hodnota pro CUWC pak činila 0,403 (40,33 %) for a 0,416 (41,66 %) pro EUWC se statisticky významným rozdílem o velikosti 0,013 (1,33 %). Následovalo vypočítání STTR každého textu v CUWC a EUWC. Hodnoty STTR v CUWC se pohybovaly na škále od 0,51 (51 %) do 0,78 (78 %) (s rozdílem 0,21) s mediánem mezi 0,66 (66 %) a 0,67 (67 %). Hodnoty STTR v EUWC se pohybovaly na škále 0,58 (58 %) do 0,82 (82 %) (s rozdílem 0,24) s mediánem mezi 0,69 (69 %) a 0,7 (70 %). Rozdíl mezi oběma škálami se ukázal být statisticky významný a hypotéza se potvrdila.

Analýza klíčových slov se zaměřila na jedinečná klíčová slova v obou korpusech, což jsou ty, které se vyskytují pouze v jednom korpusu a v druhém už ne, a na lockwords, které se v obou korpusech vyskytují nejčastěji. Po vypuštění vlastních jmen a akronymů se ukázalo, že CUWC obsahuje více jedinečných klíčových slov než EUWC. Při analýze lockwords jako nejčastěji se vyskytujících slovech na wordlistech obou korpusů vyšlo najevo, že mezi jejich výskyty byly statisticky významné rozdíly, např. mezi četnostmi členů (mnohem více určitých členů se nacházelo v CUWC), předložek, spojek a (pomocných) sloves.

Korpusová translatologie i nadále zůstává důležitým a zajímavým odvětvím současné lingvistiky a předpokládá se, že tato práce, výsledky a korpus budou využity při dalším výzkumu zaměřeném kupříkladu na frekvenci hapax legomema, poměru open a closed class items a frazeologii s využitím analýzy n-gramů.

## List of figures, graphs, and tables

# References

Al-Rawi, Mustafa Khalid Saleh. 2017. "Using Antconc: A Corpus-Based Tool, To Investigate and Analyse the Keywords in Dickens' Novel 'a Tale of Two Cities'.*" International Journal of Advanced Research* 5, no. 2: 366–72.

Baker, Mona. 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, 175-186. Amsterdam: John Benjamins Publishing Company.

Baroni, Marco, and Stefan Evert. 2017. "SIGIL: Corpus Frequency Test Wizard". Online utility. SIGIL. June 13, 2021. http://sigil.collocations.de/wizard.html.

Berman, Antoine. 1985. "Translation and the Trials of the Foreign". In *The Translation Studies Reader*, edited by Lawrence Venuti, 284–97. 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.

Bondi, Marina. 2010. "Perspectives on keywords and keyness". In *Keyness in texts*, edited by Marina Bondi, and Mike Scott, 1–18. Amsterdam: John Benjamins.

Březina, Václav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Chesterman, Andrew. 2003. 'Contrastive Text Linguistics and Translation Universals'. In *Contrastive Analysis in Language Identifying Linguistic Units of Comparison*, edited by Dominique Willems, Bart Defrancq, Timothy Colleman, and Dirk Noël, 213–29. Basingstoke: Palgrave Macmillan.

Chlumská, Lucie. 2014. 'Není korpus jako korpus: Korpusy v kontrastivní lingvistice a translatologii'. *Časopis pro Moderní Filologii* 96 (2): 221–32.

Culpeper, Jonathan, and Jane Demmen. 2015. "Keywords". In *The Cambridge Handbook of English Corpus Linguistics*, edited Douglas Biber, and Randi Reppen, 90–105. Cambridge: Cambridge. University Press.

Cvrček, Václav, and Lucie Chlumská. 2015. "Simplification in Translated Czech: A New Approach to Type-Token Ratio". *Russian Linguistics, no. 39(3).* 309–25.

Fang, Yu, and Haitao Liu. 2015. "Comparison of vocabulary richness in two translated Hongloumeng". *Glottometrics* 31: 54–75.

Firth, John. 1935. "Technique of semantics". In *Transactions of the Philological Society*, 36–72.

Francis, Gill. 1993. "A corpus-driven approach to grammar: Principles, methods and examples". In *Text and Technology*, editen by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 137–56. Amsterdam: John Benjamins.

Groom, Nicholas. 2010. "Closed-class keywords and corpus-driven discourse analysis". In *Keyness in texts*, edited by Marina Bondi, and Mike Scott, 59–78. Amsterdam: John Benjamins.

Jarvis, Scott. 2013. "Defining and measuring lexical diversity." In *Vocabulary Knowledge: Human ratings and automated measures*, edited by Scott Jarvis and Michael Daller, 13–44. Amsterdam/Philadelphia: John Benjamins.

Johansson, Stig. 1995. "Mens sana in corpore sano: On the role of corpora in linguistic research". *The European English Messenger* 4(2), 19–25.

Kenny, Dorothy. 1998. "Corpora in Translation Studies". In *Routledge Encyclopedia of Translation Studies*, edited by Mona Baker, 50–3. London: Routledge.

Kilgarriff, Adam. 2009. "Simple maths for keywords". In *Proceedings of Corpus Linguistics*. Liverpool, edited by Michaela Mahlberg, Victorina González-Diaz, and Catherine Smith.

———, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. "The Sketch Engine: ten years on". In *Lexicography* 1, 7–36.

Knittlová, Dagmar. 2000. *K teorii i praxi překladu*. 2$^{nd}$ edition. Olomouc: Palacký University.

Köhler, Reinhard, and Matthias Galle. 1993. "Dynamic aspects of text characteristics". In *Quantitative Text Analysis*, edited by Luděk Hřebíček and Gabriel Altmann, 46–53. Trier: WVT Wissenschaftlicher Verlag.

Kruger, Alet. 2002. "Corpus-based translation research: Its development and implications for general, literary and Bible translation". *Acta Theologica* 22 (1), 70–106.

Kruger, Haidee, and Bertus van Rooy. 2016. “Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English." *English World-Wide* 37:1, 26-57.

Kubát, Miroslav, and Jiří Milička. 2013. “Vocabulary richness measure in genres.” *Journal of Quantitative Linguistics* 20 (4), 339–49.

Lanstyák, István, and Pál Heltai. 2012. “Universals in Language Contact and Translation”. *Across Languages and Cultures* 13: 99–121.

Laufer, Batia, and Paul Nation. 1995. “Vocabulary Size and Use: Lexical Richness in L2 Written Production”. *Applied Linguistics Vol. 16*, No. 3, 307–22.

Laviosa, Sara. 1998. “Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose”. *Meta* 43 (4): 1–15.

———. 2004. “Corpus-based translation studies: Where does it come from? Where is it going?”. *Language Matters* 35 (1): 6–27.

Leech, Geoffrey. 1991. “The state of the art in corpus linguistics”. In *English Corpus Linguistics*, edited by Karin Aijmer, and Bengi Altenberg, 8–29. London: Routledge.

Liu, Vinci, and James R. Curran. 2006. “Web Text Corpus for Natural Language Processing”. In *Proceedings of EACL*, 233–240.

Manning, Chris, and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.

Mazgutova, Diana, and Judit Kormos. 2015. “Syntactic and lexical development in an intensive English for Academic Purposes programme.” *Journal of Second Language Writing* 29, 3–15.

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.

Mifková, Hana. 2019. “Lexical Issues in L2 English”. Bachelor thesis, Olomouc: Palacký University.

Pym, Anthony. 2008. “On Toury’s Laws of how Translators Translate”. In *Beyond Descriptive Translation Studies. Investigations in Homage to Gideon Toury,* edited by Anthony Pym et al., 311–28. Amsterdam: John Benjamins.

Scott, Mike. 2010. “Problems in investigating keyness, or cleaning the undergrowth and marking out trails”. In *Keyness in texts*, edited by Marina Bondi, and Mike Scott, 43–58. Amsterdam: John Benjamins.

———. 2013. *WordSmith Tools Manual*. Version 6.0. Liverpool: Lexical Analysis Software Ltd. See http://www.lexically.net/downloads/version6/wordsmith6.pdf (accessed 4 March 2021).

———, and Christopher Tribble. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

‘Sketch Engine’. 2021. Accessed March 20. https://www.sketchengine.co.uk/.

Stubbs, Michael. 2010. “Three concepts of keywords”. In *Keyness in texts*, edited by Marina Bondi, and Mike Scott, 21–42. Amsterdam: John Benjamins.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing Company.

Tymoczko, Maria. 1998. “Computerized Corpora and the Future of Translation Studies”. *Meta: Journal Des Traducteurs* 43 (4): 652–59.

Wang, Xuan. 2014. “The Relationship between Lexical Diversity and EFL Writing Proficiency.” In *University of Sydney Papers in TESOL* 9, 65–88.

Zanettin, Federico. 2012. *Translation-Driven Corpora*. Manchester: St Jerome.

———. 2013. “Corpus Methods for Descriptive Translation Studies”. Procedia – Social and Behavioral Sciences 95: 20–32.

## Abstract

The bachelor thesis presents a corpus based study which is concerned with constrained language of translation and non-translation language produced by native and second-language speakers with focus on their lexical richness and keywords. The purpose of the thesis is to compile and analyse two monolingual comparable corpora consisting of Czech and English university website texts. It is hypothesized that the English versions of Czech university websites will be not as lexically rich as their native English counterparts, that they will feature less unique keywords (those appearing only in one corpus and not the other), and that there will be statistically significant differences between the frequencies of lockwords (the most frequent words in both corpora).

## Key words

monolingual comparable corpus, web corpus, constrained language, lexical richness, keywords, university websites

## Anotace

Tato bakalářská práce představuje korpusovou studii, která se zabývá fenoménem „constrained language“ překladového a nepřekladového jazyka rodilých a nerodilých mluvčí se zaměřením na lexikální bohatost a klíčová slova. Účelem práce je vytvořit a analyzovat dva jednojazyčné srovnávací korpusy, které obsahují texty webových stránek českých a anglických univerzit. Předpokládá se, že anglické verze webů českých univerzit nebudou tak lexikálně bohaté jako jejich rodilé anglické protějšky, že budou obsahovat méně jedinečných klíčových slov (těch, které se nachází pouze v jednom z korpusů, a ne v druhém) a že mezi četnostmi lockwords (slov, které se v obou korpusech vyskytují nejčastěji) budou statisticky významné rozdíly.

## Klíčová slova

jednojazyčný srovnávací korpus, webový korpus, constrained language, lexikální bohatost, klíčová slova, webové stránky univerzit

# Appendix

| CUWC | STTR | Type | Token | EUWC | STTR | Type | Token |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 0.67 | 366 | 974 | 1 | 0.65 | 66 | 145 |
| 2 | 0.65 | 593 | 1,555 | 2 | 0.72 | 230 | 473 |
| 3 | 0.64 | 945 | 4,419 | 3 | 0.68 | 256 | 641 |
| 4 | 0.59 | 103 | 278 | 4 | 0.69 | 262 | 607 |
| 5 | 0.66 | 91 | 201 | 5 | 0.69 | 190 | 406 |
| 6 | 0.70 | 404 | 888 | 6 | 0.72 | 154 | 325 |
| 7 | 0.70 | 603 | 1,481 | 7 | 0.69 | 193 | 479 |
| 8 | 0.68 | 234 | 509 | 8 | 0.72 | 157 | 328 |
| 9 | 0.65 | 202 | 589 | 9 | 0.66 | 159 | 378 |
| 10 | 0.72 | 973 | 2,970 | 10 | 0.66 | 255 | 722 |
| 11 | 0.68 | 71 | 168 | 11 | 0.75 | 320 | 758 |
| 12 | 0.69 | 211 | 459 | 12 | 0.7 | 103 | 217 |
| 13 | 0.69 | 695 | 1,896 | 13 | 0.69 | 247 | 593 |
| 14 | 0.64 | 74 | 184 | 14 | 0.64 | 141 | 335 |
| 15 | 0.67 | 334 | 959 | 15 | 0.73 | 122 | 261 |
| 16 | 0.62 | 132 | 328 | 16 | 0.73 | 459 | 1,203 |
| 17 | 0.68 | 291 | 801 | 17 | 0.72 | 195 | 439 |
| 18 | 0.61 | 123 | 297 | 18 | 0.66 | 104 | 222 |
| 19 | 0.67 | 162 | 371 | 19 | 0.72 | 234 | 535 |
| 20 | 0.62 | 138 | 340 | 20 | 0.66 | 334 | 898 |
| 21 | 0.64 | 172 | 437 | 21 | 0.69 | 88 | 174 |
| 22 | 0.69 | 131 | 296 | 22 | 0.69 | 116 | 250 |
| 23 | 0.66 | 274 | 622 | 23 | 0.6 | 144 | 392 |
| 24 | 0.71 | 362 | 807 | 24 | 0.69 | 417 | 984 |
| 25 | 0.73 | 143 | 303 | 25 | 0.71 | 168 | 320 |
| 26 | 0.66 | 125 | 289 | 26 | 0.72 | 244 | 497 |
| 27 | 0.72 | 102 | 222 | 27 | 0.68 | 280 | 765 |
| 28 | 0.68 | 68 | 167 | 28 | 0.63 | 110 | 247 |
| 29 | 0.71 | 166 | 347 | 29 | 0.68 | 163 | 347 |
| 30 | 0.70 | 344 | 846 | 30 | 0.72 | 247 | 634 |
| 31 | 0.71 | 348 | 755 | 31 | 0.63 | 240 | 699 |
| 32 | 0.70 | 269 | 635 | 32 | 0.62 | 350 | 1,283 |
| 33 | 0.64 | 137 | 362 | 33 | 0.7 | 242 | 514 |
| 34 | 0.57 | 303 | 1,409 | 34 | 0.58 | 142 | 385 |
| 35 | 0.68 | 351 | 891 | 35 | 0.68 | 206 | 415 |
| 36 | 0.63 | 107 | 266 | 36 | 0.79 | 125 | 214 |
| 37 | 0.71 | 103 | 222 | 37 | 0.62 | 170 | 407 |
| 38 | 0.69 | 344 | 1,036 | 38 | 0.68 | 94 | 201 |
| 39 | 0.64 | 268 | 751 | 39 | 0.75 | 125 | 235 |
| 40 | 0.72 | 231 | 549 | 40 | 0.65 | 289 | 714 |
| 41 | 0.68 | 337 | 871 | 41 | 0.68 | 129 | 273 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 42 | 0.68 | 251 | 589 | 42 | 0.66 | 232 | 537 |
| 43 | 0.67 | 268 | 651 | 43 | 0.66 | 178 | 385 |
| 44 | 0.69 | 217 | 481 | 44 | 0.65 | 124 | 268 |
| 45 | 0.70 | 298 | 830 | 45 | 0.72 | 437 | 933 |
| 46 | 0.66 | 273 | 663 | 46 | 0.7 | 357 | 813 |
| 47 | 0.65 | 425 | 1,439 | 47 | 0.61 | 61 | 146 |
| 48 | 0.69 | 1509 | 5,473 | 48 | 0.6 | 130 | 280 |
| 49 | 0.72 | 237 | 496 | 49 | 0.69 | 357 | 797 |
| 50 | 0.70 | 329 | 757 | 50 | 0.64 | 423 | 996 |
| 51 | 0.62 | 200 | 512 | 51 | 0.68 | 201 | 464 |
| 52 | 0.66 | 296 | 783 | 52 | 0.7 | 869 | 2,522 |
| 53 | 0.66 | 322 | 911 | 53 | 0.68 | 441 | 1,120 |
| 54 | 0.64 | 383 | 1,118 | 54 | 0.67 | 88 | 186 |
| 55 | 0.67 | 289 | 714 | 55 | 0.74 | 102 | 190 |
| 56 | 0.72 | 197 | 394 | 56 | 0.68 | 82 | 164 |
| 57 | 0.63 | 114 | 258 | 57 | 0.71 | 172 | 320 |
| 58 | 0.64 | 162 | 374 | 58 | 0.74 | 237 | 512 |
| 59 | 0.68 | 494 | 1,227 | 59 | 0.68 | 254 | 552 |
| 60 | 0.61 | 129 | 301 | 60 | 0.72 | 95 | 195 |
| 61 | 0.68 | 183 | 398 | 61 | 0.76 | 153 | 295 |
| 62 | 0.65 | 203 | 482 | 62 | 0.7 | 75 | 150 |
| 63 | 0.70 | 352 | 852 | 63 | 0.71 | 190 | 386 |
| 64 | 0.64 | 243 | 586 | 64 | 0.72 | 284 | 553 |
| 65 | 0.64 | 86 | 207 | 65 | 0.63 | 129 | 294 |
| 66 | 0.67 | 449 | 1,336 | 66 | 0.74 | 74 | 144 |
| 67 | 0.63 | 153 | 390 | 67 | 0.71 | 239 | 463 |
| 68 | 0.69 | 123 | 286 | 68 | 0.67 | 111 | 219 |
| 69 | 0.57 | 241 | 896 | 69 | 0.7 | 191 | 401 |
| 70 | 0.56 | 201 | 603 | 70 | 0.75 | 267 | 523 |
| 71 | 0.67 | 623 | 1,878 | 71 | 0.66 | 114 | 245 |
| 72 | 0.71 | 381 | 913 | 72 | 0.74 | 164 | 310 |
| 73 | 0.67 | 136 | 295 | 73 | 0.72 | 83 | 167 |
| 74 | 0.72 | 322 | 750 | 74 | 0.71 | 233 | 470 |
| 75 | 0.62 | 271 | 687 | 75 | 0.7 | 215 | 434 |
| 76 | 0.64 | 137 | 353 | 76 | 0.68 | 110 | 240 |
| 77 | 0.65 | 345 | 1,029 | 77 | 0.71 | 352 | 996 |
| 78 | 0.69 | 465 | 1,372 | 78 | 0.7 | 392 | 955 |
| 79 | 0.67 | 210 | 546 | 79 | 0.71 | 238 | 515 |
| 80 | 0.70 | 183 | 398 | 80 | 0.73 | 183 | 408 |
| 81 | 0.70 | 238 | 589 | 81 | 0.72 | 256 | 575 |
| 82 | 0.51 | 109 | 385 | 82 | 0.67 | 155 | 356 |
| 83 | 0.72 | 310 | 683 | 83 | 0.68 | 168 | 362 |
| 84 | 0.67 | 270 | 643 | 84 | 0.74 | 347 | 774 |
| 85 | 0.65 | 332 | 956 | 85 | 0.74 | 341 | 873 |
| 86 | 0.64 | 116 | 291 | 86 | 0.78 | 90 | 167 |

|  |  |  |  |  |  |  |  |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 87 | 0.61 | 485 | 1,888 | 87 | 0.72 | 107 | 224 |
| 88 | 0.59 | 120 | 326 | 88 | 0.82 | 99 | 174 |
| 89 | 0.63 | 314 | 1,056 | 89 | 0.77 | 123 | 232 |
| 90 | 0.69 | 145 | 303 | 90 | 0.74 | 279 | 606 |
| 91 | 0.72 | 133 | 279 | 91 | 0.79 | 135 | 234 |
| 92 | 0.69 | 197 | 442 | 92 | 0.67 | 283 | 660 |
| 93 | 0.74 | 85 | 190 | 93 | 0.73 | 226 | 457 |
| 94 | 0.71 | 280 | 747 | 94 | 0.73 | 101 | 187 |
| 95 | 0.54 | 67 | 191 | 95 | 0.73 | 275 | 594 |
| 96 | 0.71 | 259 | 593 | 96 | 0.72 | 119 | 244 |
| 97 | 0.62 | 175 | 602 | 97 | 0.76 | 142 | 260 |
| 98 | 0.58 | 77 | 251 | 98 | 0.78 | 159 | 441 |
| 99 | 0.70 | 126 | 310 | 99 | 0.73 | 262 | 580 |
| 100 | 0.68 | 237 | 660 | 100 | 0.68 | 80 | 169 |
| 101 | 0.66 | 223 | 623 | 101 | 0.7 | 384 | 1,146 |
| 102 | 0.68 | 231 | 610 | 102 | 0.68 | 188 | 475 |
| 103 | 0.69 | 213 | 539 | 103 | 0.72 | 109 | 215 |
| 104 | 0.55 | 136 | 420 | 104 | 0.71 | 358 | 1,111 |
| 105 | 0.68 | 387 | 1,068 | 105 | 0.72 | 428 | 1,084 |
| 106 | 0.69 | 322 | 882 | 106 | 0.66 | 470 | 1,868 |
| 107 | 0.62 | 235 | 648 | 107 | 0.59 | 99 | 259 |
| 108 | 0.59 | 98 | 243 | 108 | 0.66 | 195 | 529 |
| 109 | 0.66 | 292 | 758 | 109 | 0.66 | 172 | 492 |
| 110 | 0.75 | 97 | 207 | 110 | 0.68 | 254 | 708 |
| 111 | 0.70 | 125 | 280 | 111 | 0.65 | 492 | 2,026 |
| 112 | 0.62 | 276 | 840 | 112 | 0.65 | 189 | 532 |
| 113 | 0.65 | 192 | 513 | 113 | 0.72 | 162 | 352 |
| 114 | 0.72 | 292 | 732 | 114 | 0.68 | 248 | 577 |
| 115 | 0.72 | 92 | 214 | 115 | 0.66 | 299 | 763 |
| 116 | 0.65 | 76 | 187 | 116 | 0.68 | 178 | 396 |
| 117 | 0.67 | 132 | 314 | 117 | 0.7 | 135 | 262 |
| 118 | 0.71 | 550 | 1,569 | 118 | 0.7 | 335 | 665 |
| 119 | 0.72 | 369 | 1,113 | 119 | 0.65 | 225 | 517 |
| 120 | 0.60 | 94 | 264 | 120 | 0.74 | 145 | 282 |
| 121 | 0.71 | 288 | 719 | 121 | 0.61 | 130 | 321 |
| 122 | 0.70 | 93 | 209 | 122 | 0.68 | 171 | 359 |
| 123 | 0.65 | 275 | 708 | 123 | 0.69 | 255 | 551 |
| 124 | 0.66 | 94 | 234 | 124 | 0.69 | 309 | 716 |
| 125 | 0.66 | 606 | 1,866 | 125 | 0.74 | 128 | 248 |
| 126 | 0.64 | 225 | 528 | 126 | 0.7 | 405 | 1,056 |
| 127 | 0.64 | 420 | 1,155 | 127 | 0.7 | 199 | 420 |
| 128 | 0.71 | 124 | 265 | 128 | 0.71 | 242 | 524 |
| 129 | 0.60 | 172 | 462 | 129 | 0.72 | 215 | 460 |
| 130 | 0.63 | 188 | 471 | 130 | 0.67 | 315 | 807 |
| 131 | 0.70 | 274 | 771 | 131 | 0.59 | 152 | 391 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 132 | 0.64 | 96 | 258 | 132 | 0.61 | 122 | 298 |
| 133 | 0.59 | 121 | 371 | 133 | 0.72 | 141 | 271 |
| 134 | 0.65 | 228 | 493 | 134 | 0.65 | 125 | 286 |
| 135 | 0.63 | 159 | 448 | 135 | 0.74 | 151 | 292 |
| 136 | 0.65 | 188 | 478 | 136 | 0.69 | 163 | 342 |
| 137 | 0.63 | 89 | 224 | 137 | 0.68 | 160 | 328 |
| 138 | 0.65 | 499 | 1,419 | 138 | 0.68 | 185 | 341 |
| 139 | 0.65 | 263 | 700 | 139 | 0.72 | 96 | 201 |
| 140 | 0.68 | 226 | 517 | 140 | 0.7 | 98 | 203 |
| 141 | 0.71 | 263 | 534 | 141 | 0.72 | 93 | 178 |
| 142 | 0.70 | 133 | 323 | 142 | 0.69 | 71 | 149 |
| 143 | 0.71 | 97 | 221 | 143 | 0.71 | 132 | 260 |
| 144 | 0.68 | 540 | 1,354 | 144 | 0.7 | 87 | 174 |
| 145 | 0.65 | 262 | 665 | 145 | 0.63 | 117 | 254 |
| 146 | 0.64 | 147 | 366 | 146 | 0.77 | 191 | 334 |
| 147 | 0.71 | 396 | 862 | 147 | 0.69 | 133 | 258 |
| 148 | 0.68 | 209 | 445 | 148 | 0.71 | 133 | 249 |
| 149 | 0.68 | 149 | 347 | 149 | 0.73 | 123 | 244 |
| 150 | 0.64 | 123 | 299 | 150 | 0.7 | 102 | 205 |
| 151 | 0.68 | 448 | 996 | 151 | 0.76 | 171 | 307 |
| 152 | 0.65 | 114 | 248 | 152 | 0.61 | 98 | 213 |
| 153 | 0.69 | 276 | 616 | 153 | 0.67 | 67 | 144 |
| 154 | 0.69 | 292 | 716 | 154 | 0.71 | 144 | 292 |
| 155 | 0.58 | 119 | 304 | 155 | 0.67 | 95 | 184 |
| 156 | 0.73 | 172 | 360 | 156 | 0.64 | 111 | 245 |
| 157 | 0.64 | 113 | 254 | 157 | 0.6 | 107 | 240 |
| 158 | 0.69 | 73 | 170 | 158 | 0.74 | 406 | 894 |
| 159 | 0.65 | 73 | 186 | 159 | 0.72 | 173 | 344 |
| 160 | 0.68 | 102 | 221 | 160 | 0.73 | 136 | 260 |
| 161 | 0.66 | 74 | 174 | 161 | 0.78 | 196 | 354 |
| 162 | 0.75 | 85 | 183 | 162 | 0.68 | 90 | 177 |
| 163 | 0.73 | 200 | 392 | 163 | 0.74 | 130 | 230 |
| 164 | 0.67 | 150 | 379 | 164 | 0.75 | 89 | 168 |
| 165 | 0.65 | 245 | 651 | 165 | 0.73 | 84 | 166 |
| 166 | 0.61 | 210 | 623 | 166 | 0.67 | 262 | 525 |
| 167 | 0.58 | 404 | 2,276 | 167 | 0.71 | 157 | 324 |
| 168 | 0.65 | 146 | 382 | 168 | 0.63 | 82 | 191 |
| 169 | 0.72 | 170 | 344 | 169 | 0.74 | 147 | 275 |
| 170 | 0.75 | 107 | 226 | 170 | 0.72 | 235 | 448 |
| 171 | 0.77 | 350 | 721 | 171 | 0.69 | 119 | 238 |
| 172 | 0.72 | 134 | 277 | 172 | 0.71 | 127 | 254 |
| 173 | 0.66 | 69 | 169 | 173 | 0.72 | 132 | 241 |
| 174 | 0.69 | 132 | 306 | 174 | 0.62 | 102 | 212 |
| 175 | 0.68 | 317 | 800 | 175 | 0.74 | 185 | 357 |
| 176 | 0.65 | 400 | 1,352 | 176 | 0.58 | 102 | 227 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 177 | 0.64 | 104 | 242 | 177 | 0.7 | 118 | 243 |
| 178 | 0.62 | 73 | 182 | 178 | 0.7 | 111 | 207 |
| 179 | 0.64 | 86 | 202 | 179 | 0.73 | 131 | 248 |
| 180 | 0.64 | 87 | 210 | 180 | 0.66 | 209 | 453 |
| 181 | 0.71 | 159 | 337 | 181 | 0.71 | 276 | 586 |
| 182 | 0.72 | 140 | 293 | 182 | 0.71 | 239 | 511 |
| 183 | 0.58 | 110 | 278 | 183 | 0.72 | 195 | 385 |
| 184 | 0.78 | 158 | 305 | 184 | 0.72 | 264 | 533 |
| 185 | 0.68 | 112 | 236 | 185 | 0.63 | 156 | 341 |
| 186 | 0.64 | 249 | 615 | 186 | 0.66 | 788 | 2,293 |
| 187 | 0.72 | 144 | 294 | 187 | 0.71 | 107 | 220 |
| 188 | 0.51 | 89 | 234 | 188 | 0.59 | 83 | 195 |
| 189 | 0.62 | 94 | 232 | 189 | 0.68 | 187 | 448 |
| 190 | 0.72 | 219 | 513 | 190 | 0.73 | 910 | 2,598 |
| 191 | 0.65 | 267 | 641 | 191 | 0.72 | 184 | 374 |
| 192 | 0.63 | 206 | 541 | 192 | 0.7 | 176 | 353 |
| 193 | 0.67 | 269 | 607 | 193 | 0.72 | 90 | 175 |
| 194 | 0.69 | 297 | 678 | 194 | 0.69 | 109 | 205 |
| | | | | 195 | 0.67 | 100 | 208 |
| | | | | 196 | 0.67 | 118 | 244 |
| | | | | 197 | 0.7 | 228 | 493 |
| | | | | 198 | 0.7 | 223 | 449 |
| | | | | 199 | 0.7 | 118 | 240 |
| | | | | 200 | 0.62 | 109 | 216 |
| | | | | 201 | 0.69 | 180 | 357 |
| | | | | 202 | 0.7 | 220 | 432 |
| | | | | 203 | 0.73 | 95 | 182 |
| | | | | 204 | 0.66 | 94 | 187 |
| | | | | 205 | 0.71 | 268 | 707 |
| | | | | 206 | 0.67 | 126 | 578 |
| | | | | 207 | 0.66 | 134 | 306 |
| | | | | 208 | 0.71 | 184 | 419 |
| | | | | 209 | 0.67 | 170 | 373 |
| | | | | 210 | 0.7 | 367 | 1,069 |
| | | | | 211 | 0.6 | 137 | 332 |
| | | | | 212 | 0.61 | 142 | 369 |
| | | | | 213 | 0.75 | 112 | 221 |
| | | | | 214 | 0.67 | 343 | 962 |
| | | | | 215 | 0.71 | 186 | 383 |
| | | | | 216 | 0.74 | 185 | 360 |
| | | | | 217 | 0.74 | 218 | 408 |
| | | | | 218 | 0.77 | 83 | 155 |
| | | | | 219 | 0.73 | 261 | 558 |
| | | | | 220 | 0.72 | 226 | 515 |
| | | | | 221 | 0.69 | 151 | 347 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | 222 | 0.71 | 167 | 345 |
| | | | | 223 | 0.73 | 260 | 508 |
| | | | | 224 | 0.7 | 313 | 740 |
| | | | | 225 | 0.68 | 141 | 309 |
| | | | | 226 | 0.74 | 294 | 644 |
| | | | | 227 | 0.69 | 212 | 504 |
| | | | | 228 | 0.66 | 134 | 297 |
| | | | | 229 | 0.71 | 269 | 624 |
| | | | | 230 | 0.69 | 265 | 559 |
| | | | | 231 | 0.7 | 132 | 288 |
| | | | | 232 | 0.66 | 123 | 290 |
| | | | | 233 | 0.69 | 232 | 615 |
| | | | | 234 | 0.67 | 295 | 766 |
| | | | | 235 | 0.72 | 152 | 307 |
| | | | | 236 | 0.68 | 108 | 224 |
| | | | | 237 | 0.71 | 307 | 638 |
| | | | | 238 | 0.72 | 215 | 464 |
| | | | | 239 | 0.62 | 106 | 240 |
| | | | | 240 | 0.7 | 220 | 438 |
| | | | | 241 | 0.75 | 177 | 332 |
| | | | | 242 | 0.68 | 471 | 1,487 |
| | | | | 243 | 0.67 | 124 | 323 |
| | | | | 244 | 0.67 | 446 | 1,246 |
| | | | | 245 | 0.7 | 612 | 1,593 |
| | | | | 246 | 0.71 | 212 | 484 |
| | | | | 247 | 0.76 | 287 | 624 |
| | | | | 248 | 0.66 | 208 | 476 |
| | | | | 249 | 0.71 | 399 | 943 |
| | | | | 250 | 0.63 | 459 | 1,375 |
| | | | | 251 | 0.68 | 460 | 1,191 |
| | | | | 252 | 0.67 | 192 | 464 |
| | | | | 253 | 0.7 | 368 | 842 |
| | | | | 254 | 0.68 | 346 | 883 |
| | | | | 255 | 0.67 | 165 | 337 |
| | | | | 256 | 0.68 | 679 | 1,905 |