

# **Design, compilation and applications of an English-Polish-Belarusian Parallel Literary Corpus**

Angelika Peljak-Łapińska

Submitted to Swansea University in fulfilment of the requirements for the Degree of Doctor of  
Philosophy

*Swansea University*  
2020

## Summary

The main goal of the project is to create the English-Polish-Belarusian Literary Parallel Corpus (EPB corpus) and present its applications in several linguistic disciplines, including translation studies and discourse analysis. The thesis provides an outline of corpus linguistics research and corpus linguistics as a methodology, then addresses the problem of the differences in the development of corpus linguistics in the three languages: English (as a *lingua franca*), Polish (a statutory national language) and Belarusian (a minority language). The analysis of available tools and resources for each of these languages proves the need for the EPB corpus in order to develop useful new resources for Belarusian in particular.

A substantial part of the thesis presents the documentation of the process of creating the corpus. Various aspects of corpus design, text collection and text encoding are discussed in the context of the availability and usability of numerous tools. Special attention is paid to the tools specifically designed for each language and to the solutions that enable the data processed by these tools to be merged.

Using corpus linguistics techniques (e.g. linguistic distribution, lexical density, vector-based semantic similarity measures) the thesis goes on to explore the application of the EPB corpus in investigating translation universals, in exploring the dependency between the author's and the translator's style, in supporting translation students and professionals, and in analysis of gender discourse. These case studies clearly show the practical value of the resource.

Finally, the thesis provides a detailed overview of the plans and possibilities for further development of the project in the broader context of the evolution of Polish and Belarusian corpus linguistics.

## DECLARATION

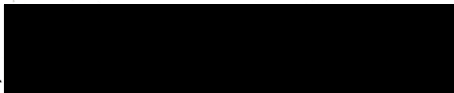
This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .....  (candidate)

Date ..... 20.05.2021 .....

## STATEMENT 1


This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .....  (candidate)

Date ..... 20.05.2021 .....

## STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .....  (candidate)

Date ..... 20.05.2021 .....

## Table of contents

Acknowledgements.....	5
Linguistics-related abbreviations and notes on transliteration.....	8
Chapter 1: Outlining the project.....	10
1.1. Specificity of the three languages: English, Polish & Belarusian.....	11
1.2. Research questions.....	17
1.3. General overview of corpus linguistics.....	18
1.4. Overview of corpus linguistics resources in English, Polish and Belarusian.....	25
1.5. Corpus linguistics as methodology.....	28
1.6. Outline of the next chapters.....	34
Chapter 2: Corpus compilation.....	36
2.1. Corpus design.....	36
2.2. Text collection.....	41
2.3. Text encoding.....	44
2.4. Alignment.....	51
2.5. Storing and sharing.....	55
Chapter 3: Applications of the EPB corpus (A): Translation studies.....	59
3.1. Theoretical translation studies: explication, simplification and levelling out in Polish and Belarusian translations of English literature.....	59
3.2. Descriptive translation studies: translator style vs. author style.....	80
3.3. Applied translation studies: how the corpus can improve the training of translators dealing with the Belarusian and Polish languages.....	96
Chapter 4: Applications of the EPB corpus (B): Discourse analysis.....	105
4.1. Gender and grammar.....	105
4.2. Gender discourse in English.....	108
4.3. Gender discourse in Polish translational data and in contrast with monolingual corpus.....	118
4.4. Gender discourse in Belarusian translational data and in contrast with monolingual corpus.....	133
Chapter 5: Looking ahead – the planned direction of the EPB corpus development.....	142
Conclusion.....	149
Glossary.....	151
Bibliography.....	154
Primary sources (A) – English-Polish-Belarusian Parallel Literary Corpus.....	154
Primary sources (B) – corpus of original Belarusian literary prose.....	173
Secondary sources.....	174
Appendix 1: Testing accuracy of Belarusian lemmatisers and POS taggers.....	202
Appendix 2: Testing relations between the contraction factor of the translations and the external factors included in the metadata.....	206
Appendix 3: Close reading of source text samples and their translations.....	207
Appendix 4: Dispersion of chosen words in the EPB corpus (results by AntConc).....	211
Appendix 5. Highest ranked collocates of chosen words from the EPB corpus.....	217

## Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor Professor Tom Cheesman. He believed in the initial project and supported my application for a Research Scholarship in Digital Humanities at Swansea University. Not only did he provide invaluable insights and inspirational guidance for my doctoral research, but also actively supported all of my extracurricular academic activities. Professor Cheesman helped me and also my family in settling into the new realities of life right after we moved to Wales and he will always remain our dear friend. I also thank my co-supervisors Dr Bob Laramée and especially Professor Andy Rothwell, who provided valuable advice on the penultimate draft.

I cannot thank enough everyone who helped turning this PhD project into a free resource available to anybody interested in Belarusian language and translation studies: Dr Volha Tratsiak and Maria Artowicz for help in processing the OCRed texts; Prof. Piotr Pęzik for introducing me to the Mantel application and thus helping me in the process of aligning the corpus; Dr Jan Wieczorek for initial help within CLARIN structures; and Wojciech Rauk for technical management of my data within the CLARIN infrastructure.

I would also like to thank colleagues I met during this journey: co-organisers of Swansea workshops on Computer-Assisted Literary Translation (CALT), Dr Roy Youdale and Dr Silke Lührmann as well as Professors Cheesman and Rothwell; CALT participants; fellow members of the Swansea Translation and Interpreting Group (STING); participants and organizers of the Lancaster Summer School in Corpus Linguistics; participants and the convenor of the Writing for Academic Publication workshop; students and lecturers of the Belarusian Studies Institute at the University of Warsaw.

Equal thanks are expressed to my friends and family, first and foremost to my husband, Marcin Łapiński. Without his support in everyday life and shared responsibility over our children this project would never have been conceived, let alone successfully finished.

## List of tables

Table 1.1. Examples of borrowings and calques from English in Polish and Belarusian.....	16
Table 2.2: UDPipe v2.5 tokenizer accuracy.....	46
Table 2.3: Number of tokens in the EPB corpus.....	47
Table 2.4: Tagging with UDPipe v2.5 accuracy.....	51
Table 2.5. Count of various types of alignment in 1-million-word EPB subcorpus.....	55
Table 3.6. Size of English-Belarusian translation data in the OPUS corpus (Tiedemann, 2012).....	61
Table 3.7. Size of Polish-English parallel data in the OPUS corpus (Tiedemann, 2012).....	62
Table 3.8. 'Would' and its Polish translations in the EPB corpus (first ten hits).....	64
Table 3.9. 'Am' and its Belarusian translation in the EPB corpus (first ten hits).....	65
Table 3.10. Size of English-Serbian parallel data in the OPUS corpus (Tiedemann, 2012).....	67
Table 3.11. Size of Czech-Polish parallel data in the OPUS corpus (Tiedemann, 2012).....	67
Table 3.12. Most frequent words in the English texts in EPB.....	68
Table 3.13. Standardised type/token ratios across corpora of English, Polish and Belarusian (calculated with WordSmith).....	72
Table 3.14. Frequency profiles of translational and non-translational corpora of English.....	73
Table 3.15. Frequency profiles of translational and non-translational corpora of Polish.....	74
Table 3.16. Frequency profiles of translational and non-translational corpora of Belarusian.....	75
Table 3.17. Average sentence length in EPB corpus components and corresponding corpora (calculated with WordSmith).....	77
Table 3.18. Highest scored verbs on the keyword lists of Christie's short stories in Belarusian.....	91
Table 3.19. Key n-grams in Belarusian translation of Christie's 'The Kidnapped Prime Minister'..	93
Table 3.20: Gender composition of texts authors (EPB corpus).....	95
Table 3.21. Occurrences of 'neck in the woods' phrase in the EPB corpus.....	97
Table 3.22. Occurrences of 'donkey's years' phrase in the EPB corpus.....	98
Table 3.23. Translations of 'bad luck' in the EPB corpus.....	100
Table 3.24. Examples of the 'set out' verb usage and their translations in the EPB corpus.....	102
Table 4.25. Instances of forms of address in the EPB corpus.....	107
Table 4.26. Feminine and masculine pronoun and nouns in the English subcorpus of the EPB corpus.....	110
Table 4.27. Collocates of the word 'man' (EPB corpus).....	113
Table 4.28. Collocates of the word 'men' (EPB corpus).....	114
Table 4.29. Collocates of the word 'boy' (EPB corpus).....	114
Table 4.30. Collocates of the word 'boys' (EPB corpus).....	115
Table 4.31. Collocates of the word 'woman' (EPB corpus).....	115
Table 4.32. Collocates of the word 'women' (EPB corpus).....	116
Table 4.33. Collocates of the word 'girl'.....	117
Table 4.34. Collocates of the word 'girls'.....	118
Table 4.35. Number of occurrences of female and male pronouns in the EPB corpus.....	118
Table 4.36. Lemmas 'kobieta', 'mężczyzna' and their synonyms (EPB corpus).....	119
Table 4.37: 'Man' translations into Polish and Belarusian (aligned section of the EPB corpus)....	120
Table 4.38: Examples of 'man' translations in the EPB corpus.....	121
Table 4.39. Collocates of the word 'kobieta' (EPB corpus).....	122
Table 4.40. Collocates of the word 'kobiety' (EPB corpus).....	122
Table 4.41. Collocates of the word 'dziewczyna' (EPB corpus).....	123
Table 4.42. Collocates of the word 'dziewczyny' (EPB corpus).....	124
Table 4.43. Collocates of the word 'mężczyzna' (EPB corpus).....	124

Table 4.44. Collocates of the word 'mężczyźni' (EPB corpus).....	125
Table 4.45. Collocates of the word 'chłopiec' (EPB corpus).....	126
Table 4.46. Collocates of the word 'chłopcy' (EPB corpus).....	127
Table 4.47. Collocates of the word 'kobieta' (InterCorp).....	128
Table 4.48. Collocates of the word 'kobiety' (InterCorp).....	128
Table 4.49. Collocates of the word 'dziewczyna' (InterCorp).....	129
Table 4.50. Collocates of the word 'dziewczyny' (InterCorp).....	130
Table 4.51. Collocates of the word 'mężczyzna' (InterCorp).....	130
Table 4.52: Collocates of the word 'mężczyźni' (InterCorp).....	131
Table 4.53: Collocates of the word 'chłopiec' (InterCorp).....	132
Table 4.54: Collocates of the word 'chłopcy' (InterCorp).....	133
Table 4.55. Lemmas 'жанчына', 'мужчына' and their synonyms (EPB corpus).....	134
Table 4.56. Collocates of the word 'жанчына' (EPB corpus).....	135
Table 4.57. Collocates of the word 'дзяўчына' (EPB corpus).....	136
Table 4.58. Collocates of the word 'мужчына' (EPB corpus).....	136
Table 4.59. Collocates of the word 'хлопец' (EPB corpus).....	137
Table 4.60: Collocates of the word 'жанчына' (InterCorp).....	138
Table 4.61: Collocates of the word 'дзяўчына' (InterCorp).....	138
Table 4.62: Collocates of the word 'мужчына' (InterCorp).....	139
Table 4.63: Collocates of the word 'хлопец' (InterCorp).....	140

## List of illustrations

Illustration 1.1. Query from the EPB corpus in a KWIC format.....	21
Illustration 2.2. Workflow of building the corpus.....	42
Illustration 3.3. Differences in contraction factors of translated texts across the EPB corpus.....	60
Illustration 3.4. Contraction factor of Swedish language translations in the ASPAC corpus (University of Gothenburg, 2018). Language family codes according to ISO “Codes of representation of names of languages”, capitalised for better visibility.....	66
Illustration 3.5. Distribution of the number of works published in years 1920-2020 in the EPB database.....	70
Illustration 3.6. Frequency profiles in corpora of English.....	73
Illustration 3.7. Frequency profiles in corpora of Polish.....	75
Illustration 3.8. Frequency profiles in corpora of Belarusian.....	76
Illustration 3.8. PCA for translational and non-translational Belarusian texts.....	78
Illustration 3.9. Dendrogram of English texts in the EPB corpus (Classic Delta measure).....	82
Illustration 3.10. Dendrogram of English texts in the EPB corpus (Eder’s Delta measure).....	83
Illustration 3.11. Dendrogram of Polish texts in the EPB corpus (Classic Delta measure).....	85
Illustration 3.12. Dendrogram of Polish texts in the EPB corpus (Eder’s Delta measure).....	86
Illustration 3.13. Dendrogram of Belarusian texts in the EPB corpus (Classic Delta measure).....	88
Illustration 3.14. Dendrogram of Belarusian texts in the EPB corpus (Eder’s Delta measure).....	89
Illustration 4.15. Dispersion of the words ‘lady’ and ‘gentleman’ in the EPB corpus.....	111
Illustration 4.16: Word 'man' collocating with lemma 'surround' (EPB corpus).....	114
Illustration 4.17. Concordances of the verb 'love' collocating with the word 'woman' in EPB corpus.....	116
Illustration 4.18: Word 'chłopiec' collocating with 'oficer' (InterCorp).....	132
Illustration 5.19. KonText main window.....	145
Illustration 5.20. Concordance in the KonText application.....	146

## **Linguistics-related abbreviations and notes on transliteration**

ACE – Australian Corpus of English

AusNC – Australian National Corpus

ARF – average reduced frequency

CECL – Centre of English Corpus Linguistics

CLARIN – Common Language Resources and Technology Infrastructure

CLiC – Corpus Linguistics in Context

CORAL – CORpus ALigner

CST – Center for Sprogteknologi [Center for Language Technology]

DA – Discourse Analysis

DTS – descriptive translation studies

ELRA – European Language Resources Association

EPB corpus – English-Polish-Belarusian Parallel Literary Corpus

ERIC – European Research Infrastructure Consortium

IJP PAN – Instytut Języka Polskiego Polskiej Akademii Nauk [The Institute of the Polish Language at the Polish Academy of Sciences]

JSTOR – Journal Storage

KORBA – Korpus Barokowy [Baroque Corpus]

KWIC – keyword in context

MF/MD – multifeature/multidimensional

NLP – Natural Language Processing

NLTK – Natural Language Toolkit

OCLC – Online Computer Library Center

OPUS – Open Parallel Corpus

PARTHENOS – Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies

PELCRA – Polish and English Language Corpora for Research and Applications

PEN – Poets, Essayists, Novelists

POS – part of speech

SSR lab – Speech Synthesis and Recognition Laboratory

STTR – standardised type/token ratio  
SVO – subject-verb-object (word order)  
TEC – Translational English Corpus  
TEI – Text Encoding Initiative  
TTR – type/token ratio  
TU – translation universals  
UD – Universal Dependencies  
VLO – Virtual Language Observatory  
VVV – Version Variation Visualization  
XML – Extensible Markup Language

All translations from Polish and Belarusian sources are made by the author unless stated otherwise. Belarusian-language names and titles are transliterated with the use of Belarusian Latin Alphabet (Uladzimir Katkouski & Rrapo, 2005) unless the name given in the paper or the title is transliterated otherwise.

## Chapter 1: Outlining the project

This thesis presents the process of designing and compiling, as well as applications of an English-Polish-Belarusian Parallel Literary Corpus (EPB corpus). The technical solutions proposed in this project are discussed against a broader background of corpus linguistics, particularly in the three languages involved. In a nutshell, the EPB corpus is a collection of 20<sup>th</sup>- and 21<sup>st</sup>-century literary prose in English, Polish and Belarusian, where a text in one of those languages is a source text for both other languages. This thesis focuses on the initial version of this corpus, which contains only English-language literature translated into Polish and Belarusian (details in Chapter 2).

As discussed in Section 1.4, there is a multitude of currently available corpora, yet the corpus presented in this dissertation is unique. It is motivated by several factors. Firstly, the project supports the minority language: it strengthens the position of the Belarusian language in the digital world. Secondly, it creates an on-line resource for researchers, mainly in linguistics but also in connected disciplines. As explained in Section 1.3, the applications of parallel corpora include, apart from purely linguistic research, also sociolinguistic studies, discourse analysis and literary stylistics. Parallel corpora are crucial in theoretical, descriptive and applied translation studies. Even though some aspects of translations between Belarusian, Polish and English have been discussed in a number of papers, no study explores that topic comprehensively, partially due to the lack of resources. Yet another discipline that can possibly benefit from the use of the EPB corpus is language technology. With the corpus and the new data, methods of automatic processing of the Belarusian language can be enhanced. Thirdly, the EPB corpus helps to develop contrastive studies into languages other than English. “The majority of the existing parallel corpora involve English as one of the two languages. Consequently, prominent researchers like Rabadan (2005: 156), Tymoczko (1998:6; 2006) & Delabatista (2008:238) signal a need for research based on translations from other languages than English, and a need for comparable studies using data from different language pairs.” (Hareide & Hofland, 2012).

Parallel corpora have been used in research into English, Polish and Belarusian in a myriad of ways, as evident from dozens of books and hundreds of articles on this topic. Regarding English, researchers applied parallel corpora for investigating vocabulary (*Al-Ajmi, 2004; Chujo et al., 2006; Wu & Xia, 1994*) and morphology (*Šilić et al., 2007; Viberg, 2017*), for extracting paraphrases (*Barzilay & McKeown, 2001*), and, most commonly, in translation studies, eg. in statistical machine translation (*Goldwater & McClosky, 2005; Ramasamy et al., 2012; Tian et al., 2014*) or for translation training (*Alotaibi & M, 2017*).

Parallel corpora in research into Polish are used in a similar way. Various studies concern lexicography (*Ł. Grabowski, 2018; Perdek, 2012*), morphology (*Jurkiewicz-Rohrbacher, 2019*) and translation studies (*Biel, Łucja, 2010; Wołk, 2015; Zeldes, 2012*). A notable difference between the body of research concerning Polish and English parallel corpora is its size. There is significantly less literature about Polish language about English, as exemplified by Google Scholar search engine which returns over 900,000 results when asked about “English parallel corpus” and only 48,000 when for querying “Polish parallel corpus” or 2,000 in the case of the Polish analogue (“polski korpus równoległy”). Understandably, the level of parallel corpora use in research into Belarusian is

extremely low due to the lack of resources, there were however interesting attempts at breaching this gap with neural machine methods (*Karakanta et al., 2018*) and using a relatively small Belarusian language datasets (37,000 words on developmental stage and 22,000 for testing).

The EPB corpus, even in its initial version presented in this thesis, supplies high-quality parallel data for the Polish and Belarusian languages, thus helping in fulfilling the aforementioned need. In summary, the main objectives for building the EPB corpus are threefold: empowering the under-resourced language, supporting linguistic and interdisciplinary studies of the three languages involved in the project and enabling contrastive studies of languages other than English.

The next sections describe the features of English, Polish and Belarusian, followed by the research questions which need to be answered in the process of compiling a corpus, with special attention to the issues that are specific to a minority language, namely Belarusian. What follows is an overview of corpus linguistics, the available corpus resources in the context of the three languages in question, and corpus linguistics as methodology. This chapter concludes with the outline of the rest of the thesis.

### **1.1. Specificity of the three languages: English, Polish & Belarusian**

The corpus compiled for the purposes of this thesis contains texts written in three languages. Two of them are Slavic languages and, to some degree, are similar in grammatical aspect, nevertheless they cannot be compared in terms of their status.

#### **1.1.1. Belarusian**

*Ethnologue* (Simons & Fenning, 2017) classifies Belarusian as a status 1 language (statutory national language, but with “largely symbolic use”) with approximately 2.5 million users; the symbolic use is due to the fact that Belarus has two official languages, Russian and Belarusian, however the disproportion in the usage of the two is significant. According to the latest census<sup>1</sup>, 53.2% of Belarusians indicated the Belarusian language as their mother tongue, but only 23% use it at home. 70% declare Russian is their main language of everyday communication (National Statistical Committee of the Republic of Belarus, 2009).

Apart from Belarus, Belarusian is used as an official language in a few Eastern communes in Poland. Three main dialects of Belarusian are: Northeast Belarusian (regions of Połock, Vicebsk and Mohiloŭ), Southwest Belarusian (regions of Harodnia, Baranovičy, Šłuck and Mazyr) and Central Belarusian (Simons & Fenning, 2017). The differences between them are fairly small.

Census data demonstrate that Belarusian is a minority language in Belarus however the actual numbers might be even lower due to the “so-called *trasianka* factor” (Vasilevich, 2013). *Trasianka* is the term referring to the mixture of Russian and Belarusian, it differs from speaker to speaker and the number of features of each language varies in many respects, however in many cases *traskanika* is primarily Russian language with only a minor shift towards Belarusian. Still, some people might

---

<sup>1</sup> The first findings of the 2019 census were announced in February 2020, however processing the data will take until July 2020 and the full results – including information regarding language – will be available in the Autumn 2020.

consider that variety to be the Belarusian language and thus the numbers given in the census could be, in fact, very misleading and over-estimate the real use of Belarusian.

Another peculiarity of the Belarusian language is the parallel existence of two orthographic variants: *Taraškievica*, referred to as Belarusian Classical Orthography (created by Branisłaŭ Taraškievič in 1918), and *narkamaŭka* which is a version that was reformed in 1933, during the Soviet times (Mayo, 1978). *Taraškievica* is not officially acknowledged in Belarus, it is, however, broadly used by emigrants who are critical about *narkamaŭka* because many of its elements draw Belarusian orthography towards Russian.

Another important aspect to be considered is purely the linguistic characteristics of the Belarusian language. Regarding the typology based on the language affinity, Belarusian comes from the Indo-European family, Balto-Slavic sub-family, Slavic group and Eastern sub-group. By virtue of being an East-Slavic language, Belarusian shares a number of features with Russian and Ukrainian but on the lexical level a strong influence of Polish is visible. Studies reveal that Belarusian is lexically more similar to Polish than Russian (Katkouski, 2005). With respect to the basic word order Belarusian represents the most typical SVO (subject-verb-object) kind. Belarusian uses stress to distinguish words and is a synthetic language with a high degree of inflection (six cases and three genders). Recently English language influence has been apparent, especially in vocabulary from the domain of new technologies.

Considering the circumstances mentioned in previous sections, UNESCO's decision to classify the Belarusian language as vulnerable (Moseley, 2010) is understandable. However the language situation in Belarus is not simply an outcome of a one-time decision of giving Russian and Belarusian equal status as official languages. It is a result of complicated socio-historical processes that can be traced back to the 13<sup>th</sup> century and the Grand Duchy of Lithuania that encompassed all Belarusian duchies. Citizens of the Grand Duchy used the Ruthenian language (also called Old Belarusian), which almost entirely replaced the Old Slavonic language. Old Belarusian developed without obstacles, until in the 16<sup>th</sup> century the country became a part of the Polish-Lithuanian Commonwealth and Polish culture and language started to dominate in the territory, specially among the higher social strata, whereas the most numerous part of society, that is peasants or villagers, still used Belarusian, and it was the village that a few centuries later became a starting point for recovering the language.

Meanwhile, the 19<sup>th</sup> century brought the rule of the Russian Empire and replacement of the Polish language with Russian, predominantly through the use of terror. In the 20<sup>th</sup> century turbulent political and social changes took place. After World War I, in 1918 Belarus became an independent state and fast Belarusization<sup>2</sup> started. Belarusization "was part of the authorities' agenda to bring the government policy closer to the people and to diminish the cultural and linguistic differences between the cities and the countryside" (Rudling, 2014, p. 126). Owing to the fact that Belarusian

---

2 Some scholars refer to this process as *Belarusification* (Ioffe G., 2003, *Understanding Belarus: Questions of Language*), which resembles the term *Russification*, only with prefix *Bela-* that conveys no meaning for English language speakers. In a historical sense, Russification refers to Imperial Russia and the Soviet Union enforcing the Russian culture and language by means of violence and fear, therefore it has strong negative connotations and for this reason in this work the term *Belarusization* (created as an analogy to *беларусізацыя* [bielarusizacyja]) will be used.

survived in villages it could now have been restored and introduced in all spheres of public life. That concerned, however, only the central and eastern parts of the Belarusian lands, as the west was incorporated by Poland and was being Polishised.

Belarusization survived only until the 1930's – that is when Stalin's policy changed and the unification of the “Soviet nation” started. Because of the *Great Terror* (Yekelchik, 2008) nearly all Belarusian elites were either killed or exiled to Siberia. World War II ruined the country economically, but was celebrated as a Soviet victory and called the *Great Patriotic War* (rus. *Великая Отечественная война*; Portnov, 2011). It all added to the strong Russification of Belarus that was now one of the Soviet Union's republics. Belarus became independent again in 1991, but the situation did not change significantly – Belarusian was restored as a state language and an attempt at Belarusization was made again, however three years later Aliaksandr Lukashenka took office as president and acting in accordance with the results of a national referendum introduced Russian as a second state language. What followed the decision was the supremacy of Russian and what is called by researchers *replacive bilingualism* (Zaprudski, 2007).

Replacive, or subtractive bilingualism is “experienced by many ethnic minority groups who because of national educational policies and social pressures of various sorts are forced to put aside their ethnic language for a national language” (Lambert, 1975). Such a situation is most common among immigrants and its result is very often lower proficiency in both languages (Cummins, 1984). It seems unsettling in the Belarusian context, as Belarusians are neither immigrants or ethnic minority in Belarus, rather the opposite.

Nevertheless, until very recently the government policy fought directly and indirectly against the Belarusian language. Actions undertaken by the officials concern many spheres of life and influence people from the early stages of their life. The prevalence of the Russian language in the education system is enormous even though the proportions of Belarusian and Russian in everyday use were more balanced in the 1990's (36.7%-62.8%), only after the 1994 referendum Russian started displacing Belarusian more decidedly (Smolicz & Radzik, 2004; Spasiuk, 2015; Viačorka, 2017). The numbers of Belarusian language lessons was systematically decreased and the language of instruction of nearly all subjects is Russian. Requests for forming Belarusian-language groups are denied and the only fully Belarusian secondary school, Belarusian Humanities Lyceum, continues working underground after officially being closed in 2003 (Spasiuk, 2017). Meanwhile the Belarusian minority living in Poland managed to organise pre-school and first grade Belarusian-language classes in Warsaw (‘Mniejszość Białoruska – Szkoła Podstawowa nr 395 [Belarusian Minority - Ground school no 395]’, n.d.) with the institutional support of Polish officials that is guaranteed by European laws.

Officials in various situations refuse to use Belarusian, such as judges in courts (Human Rights Center ‘Viasna’, 2015), or the Central Electoral Commission during elections (Smolicz & Radzik, 2004). Also everyday activities are impeded, for instance sending a telegram or filling out a registration form in Belarusian (Smolicz & Radzik, 2004). Mass media are dominated by Russian, as is clear from the figures provided by the Belarusian Ministry of Information: out of 1395 “editions” (which include paper titles and radio and TV channels) 526 (37,7%) are delivered in both

Russian and Belarusian, only 32 (2,3%) in Belarusian and consequently more than a half is exclusively in Russian. Also most books are published in this language and publishing houses that promote books in Belarusian usually face administrative hindrances (Shearlaw, 2015). An example of such an obstruction is the requirement of registering with the ministry of information – a publisher has to prove they do not act against the interest of the Republic of Belarus. An instance of an activity contradicting the country's welfare and deemed extremist was, as it was evident in the case of Lohvinaŭ publisher, selling the book "Belarus Press Photo 2011" which presented images of the policy brutality towards the citizens of Belarus (Shearlaw, 2015).

Only recently, since 2014, has the government's attitude started to change and two main factors contributed to this shift. Firstly, the steadily deteriorating economic situation in Belarus which is evident in redenominations (in 2000 and 2016) as well as in devaluations (by 20% in 2009 and 56% in 2011). The constant need for money forced Belarus to seek the help of Russia but since 2014 the direction of the requests for help changed due to "the war in Ukraine, Russia's increasingly problematic relations with all of its neighbours and Russia's own economic troubles" (Wilson, 2016, p. 78, 2016). Russia stopped being a reliable source of money and additionally it started posing a security risk. That is why the government introduced the policy of *soft Belarusization* (Mojeiko, 2015). *Soft* means in this context actions that are seemingly insignificant, in order not to alert Russia, and, in contrast to previous attempts of Belarusization, not rapid and non-institutional.

This new language policy of soft Belarusization focuses primarily on promoting the Belarusian language in the public space. One of the most well-known cases involved the Belarusian president who on Independence Day in 2014 gave a speech in Belarusian (Dynko & Bigg, 2014). It was a real sensation for Belarusians who heard the president using the country's titular language for the first time in his career. Politicians expressed the wish for Belarusian to be used more frequently in schools, specially in Belarusian history and geography lessons (Astapenia, 2016). Other instances of encouraging people to use Belarusian include social campaigns featuring world-class sports figures: Alaksandra Hierasimienia, Olympic champion in swimming (marketing.by, 2015) and Vital Hurkoŭ, Muay Thai champion (MovaNanova, 2016). These two cases are just a sample of a growing phenomenon of Belarusian language use among sports personalities (Kul, 2015).

Apart from the increase of Belarusian language presence in the public space in Belarus certain actions have been undertaken abroad, especially in Poland. The Cultural Centre of Belarus, a unit created in 2008 by the Belarusian Embassy in Poland, in fact started being active in late 2014 by announcing recruitment for Belarusian language courses (Belarusian Embassy in Poland, 2017). A few months later the Centre, in co-operation with the Belarusian Studies Institute at the University of Warsaw, organised an event entitled "Belarusian Culture in Panorama" which has since been repeated annually. Its highlights included a theatre play based on a book by Aleksijevič – Literary Noble Prize winner who is not acknowledged in Belarus due to her negative comments about the Belarusian president. Another interesting element of the "Belarusian Culture in Panorama" was a lecture about Belarusian cuisine that was given by a journalist of Belsat, the only fully Belarusian television channel based in Warsaw, regarded as oppositional.

These actions might lead to a change in the status of the Belarusian language. However the government's change in policy coincided with an already ongoing grass-roots initiative and it is this initiative that seems more likely to ultimately influence people. Belarusian language courses – “Mova ci kava” (Borowska, 2014) running from 2013 and “Mova Nanova” (MovaNanova, 2017) running from 2014 – turned out to be a huge success and attracted thousands of people across the country and abroad. The most important aspect of these courses was the involvement of citizens who attended meetings simply by virtue of their interest in the language, rather than to demonstrate oppositional political views.

The development of the Belarusian language is possible also by virtue of an active web community. Even though the number of web pages created in the Belarusian language is not high, volunteers create Wikipedia (*Wikipedia*, n.d.) and the number of Belarusian entries – 154,113 (‘Hałoŭnaja staronka [Main page]’, 2018) – is comparable with small yet not minoritised European languages, such as Estonian (175,545; (‘Esileht [Main page]’, 2017), Croatian (191,512; (‘Glavna stranica [Main page]’, 2017), Greek (146,820; (‘Πύλη’, 2018) or Lithuanian (187,913; (‘Pagrindinis puslapis [Main page]’, 2018).

Belarusian Wikipedia is a collaborative work of dedicated volunteers. In recent years events promoting Belarusian Wikipedia were organised in Belarus (Falanster, 2013), and in countries where big Belarusian communities live – Poland (Bladyniec, 2013c, 2015), Lithuania (Bladyniec, 2013a), Czech Republic (Bladyniec, 2013b) and in Sweden (Eastern Partnership Civil Society Forum, 2013). The actions of the small community proved to be effective as the number of articles in Belarusian Wikipedia tripled in the last five years.

### 1.1.2. Polish

Having discussed the status and linguistic characteristics of the Belarusian language, the next section of this chapter addresses the issue of the features of the Polish language. According to *Ethnologue* (Simons & Fenning, 2017) the Polish language is a status 1 language (statutory national language in Poland) and it is used by almost 41 million users. *Ethnologue* names a few countries as places where Polish is spoken but it does not take into account a numerous population of Polish emigrants. According to the latest census, Polish has become the most common non-native language used in England and Wales (Rawlinson, 2013) and Poles are reported to be the biggest non-UK born nationality (Office for National Statistics, 2017). Polish, similarly to Belarusian, is quite homogeneous and specialists report no significant differences between regional variants.

Polish belongs to the same language group as Belarusian and to the West-Slavic sub-group. As mentioned above, due to the historical influence of Polish on Belarusian, these languages share a number of lexical features, however they differ in other respects. One characteristic of Polish is word order that is less restricted than in Belarusian and mainly used for stressing particular information in the sentence. *Pani miała psa* [Lady had a dog], *Psa pani miała* [A dog had lady] and *Miała pani psa* [Had lady a dog] are equally grammatically correct but convey different meaning – the first example is declarative sentence with neutral meaning (unless indicated otherwise by intonation) while the second and third examples underline the element positioned at the beginning

of the sentence, additionally they convey emotional information (the second example could be easily understood as irony). Another feature of Polish that poses a possible difficulty in the process of computer analysis is its morphological richness.

For roughly 180 thousand base forms of words, almost 4 million inflected word forms exist. The inflection paradigms are complex, and even their exact number is a matter of a dispute (single exceptions might be thought to create a new paradigm). Even native speakers have problems with properly inflecting many words [...] (Miłkowski, 2012)

The English language has been frequently reported to influence Polish not only on the lexical and structural levels (Sztencel, 2009) but also in semantics (Zabawa, 2008). Similarly to Belarusian, borrowings and calques appear mostly in the domain of science and technology.

English	Polish	Belarusian
laser, smart phone, video, scanner, photography, (computer) mouse	laser, smartfon, wideo, skaner, fotografia, mysz/-ka (komputerowa)	лазер [łazier], смартфон [smartfon], відэа [videa], сканер [skanier], фатаграфія [fatahrafija], (кампутарная) мыш [(kamputarnaja] myś]

Table 1.1. Examples of borrowings and calques from English in Polish and Belarusian.

Among recent developments of Polish, experts report increased usage of feminine forms for professions and simplification in terms of morphology (namely inflection) as well as in lexis (Miłkowski, 2012).

Polish is a statutory language of Poland with all its rights. All official business, commercial activities and education are conducted in Polish. According to the latest census (Statistics Poland, 2015, p. 69) Polish is the only language of everyday contacts for 96.19% of the Polish population and 2% use both Polish and another language. What is worth mentioning is that among non-Polish languages the two most popular – Kashubian and Silesian – are regional variants of Polish.<sup>3</sup> This indicates a very strong position of Polish in its titular country. Another aspect of the Polish language's position among other languages is its presence on the Internet. The strong Polish-speaking web community is easily measured by the number of Wikipedia entries (*Wikipedia*, n.d.) – with over 1.27 million headwords it is comparable with Italian and Russian.

### 1.1.3. English

English is a widely used language, with 378 million native users, but it is also a *lingua franca* for an estimated more than a billion users (Simons & Fenning, 2017). English is an official language in 53 countries on the globe. Each country may be associated with its own dialect of English and a number of sub-dialects, additionally pidgins and creole English-based languages developed all over the world (Ananiadou et al., 2012). English is not only cultivated as a national language in many countries but it also dominates international communication in various fields, such as science, information technology, business, diplomacy, education, employment, travel and tourism, media or entertainment (*Rao, 2019*). It all adds up to the strong global position of English.

3 Both languages have debatable status, being referred to as *languages* or as *dialects* by various linguists. Although only Kashubian has an official status of regional language and has been introduced as an auxiliary language in commune offices, the authors of the Census report treat both Kashubian and Silesian as languages.

From a purely linguistic point of view English is, like Polish and Belarusian, an Indo-European language but it belongs to the Germanic group and West sub-group. English has minimal inflection compared to Polish and Belarusian. Among the peculiarities of English one can name its complicated spelling system, which is the reflection of complex historical and political changes, as well as its large number of phrasal verbs (Ananiadou et al., 2012). Other properties typical of English are the flexibility of function and openness of vocabulary (Potter, 2020). Flexibility of function is actually caused by the lack of inflection and is manifested in the freedom of using many words as both nouns and verbs. An example of this phenomenon can be sentences, such as *I was booking a place while placing a book at the shelf* or *I can can a can*. This rule also works with some adverbs, adjectives and pronouns. The openness of vocabulary in English is nothing else than the readiness of adopting and adapting words from other languages, as well as the ease of creating compounds and derivatives.

The impact of English is clearly visible on the Internet. “In 2010, there were an estimated 536 million users of the English language Internet, constituting 27.3% of all Internet users. This makes the English Internet the most used in the world” (Ananiadou et al., 2012, pp. 13–14). Because of the number of English-speaking users of the Internet it has become a standard to create web sites with at least one language variant being English. The dominance of English in the digital world is thus undeniable.

## **1.2. Research questions**

Compiling a corpus inevitably poses a number of problems and it is a complex task on its own. It becomes even more complicated when minority languages are involved. Therefore a number of research questions emerge when one is embarking on creating such a resource. The general concern is what is the most efficient way of building a corpus having the desired features and fit for the range of tasks described in the next section. This issue can be dealt with if divided into components, each of which is assigned research questions that lead to solving the problem.

Firstly, how and where to find data for a corpus of a minority language, such as Belarusian? It is an important question in the phase of designing the corpus and dealing with issues of availability of texts; identifying potential constraints and planning ways of managing them is crucial at this stage of building a corpus.

Secondly, how to efficiently collect texts for the corpus? It seems a rather naive question but the problem is in fact complex. The answer is (almost) obvious for texts in digital form, but when it comes to the data that is accessible only in analogue form, further questions are raised, such as what is the best tool for automated text processing, especially in the Belarusian language? After the data is obtained the problem of its storage emerges. How to deposit the data so it would be safe, easily accessible and also available in the far future?

Thirdly, how to make the data more useful for researchers? Obviously, by encoding it, enriching it with additional linguistic data. It is a straightforward task but it implies a bunch of additional questions. How to encode the corpus most efficiently, if the corpus comprises data in three languages, retaining high accuracy at the same time? How far should the encoding go, or in other

words – on how many levels should one annotate the data? How to bring together the encoding with alignment of the corresponding parts of datasets in three languages?

Fourthly, how to make the data accessible for the widest possible range of scholars and respect copyrights at the same time? And finally, to what degree is it possible to use the collected data in the areas indicated in Section 1.3?

Each question reflects separate stage of building the corpus: design, collection and encoding, as well as making it publicly available. These research questions are answered in Chapter 2 and the answers are accompanied by a thorough analysis of possible solutions to the challenges posed by the main task of this project. The last question concerning the possibility of a wide range of EPB corpus applications is dealt with in Chapter 3 and 4 of this thesis, and it demonstrates the steps followed to ensure the quality of the work done while compiling the corpus. With various narrow research problems, this chapter exemplifies how the data gathered in the project can contribute to the development of a number of research domains.

To conclude this section, the research questions posed above might and should be treated as directions for conducting the research. The next section describes the procedures and methods used in corpus linguistics for investigating the data. This methodology is used later on in the exemplification of the EPB corpus applications.

### **1.3. General overview of corpus linguistics**

Corpus linguistics emerged from the need to provide quantitative evidence confirming the phenomena scholars describe in their works. What we have been observing in the past few decades is a shift towards empiricism in linguistics and linguists who can no longer support their theories with intuition alone. “Because we are looking for typical patterns, analyses cannot rely on intuitions or anecdotal evidence. In many cases humans tend to notice unusual occurrences more than typical occurrences” (Biber et al., 1998, p. 3).

Even though nowadays we associate corpora with computers, researchers underline that it is not a modern invention as a huge body of texts were gathered from the 13<sup>th</sup> century when biblical studies were one of the main linguists’ concerns (Kennedy, 1998; O’Keeffe & McCarthy, 2010). These corpora were thoroughly analysed, sometimes at a great scale, such as in the project led by a German stenographer, Friedrich Wilhelm Kaeding, who, with the help of hundreds of volunteers, performed the count of frequencies of over 250,000 words within 11 million words of material (Kuebler & Zinsmeister, 2015) in the late 19<sup>th</sup> century.

Understandably, such an endeavour was extremely time-consuming and prone to mistakes, therefore it was the use of computers that revealed the real potential of corpora. Even though the definition of corpus continues to be a topic of discussion, scholars agree that a modern corpus is a body of texts in electronic form, meaning that it is machine-readable (McEnery & Wilson, 2001), and it “can serve as a basis for linguistic analysis and description” (Kennedy, 1998, p. 1). Some researchers find it necessary that a corpus must also be finite-size and sampled (*McEnery & Wilson, 2001*) to represent a language or a variety of it (Biber et al., 1998; Kuebler & Zinsmeister, 2015).

Independently from various definitions of a corpus, researchers use multiple taxonomies of corpora, however a few categories are most commonly considered. Firstly, the mode of the primary data is a criterion for distinguishing between spoken, sometimes referred to as speech corpus (Kuebler & Zinsmeister, 2015; O’Keeffe & McCarthy, 2010), written and multi-modal corpora (Kuebler & Zinsmeister, 2015). Secondly, depending on the purpose behind the corpus design, scholars name general corpus (Kennedy, 1998; Weisser, 2016) in contradistinction with specialized (Kennedy, 1998), specific (Weisser, 2016) or special (O’Keeffe & McCarthy, 2010) corpus – the three last terms may refer to e.g. a learners’ corpus or a corpus of a particular variety of a language. With regard to time, researchers unanimously distinguish between synchronic and diachronic corpora. Persistency (Kuebler & Zinsmeister, 2015), on the other hand, is a criterion specifying the regularity of adding new texts to the corpus. On this basis corpora can be divided into static, in which the texts are gathered just once and then the corpus is regarded finished, and monitor (Kennedy, 1998; Kuebler & Zinsmeister, 2015) or snapshot (Weisser, 2016) corpus, which is constantly supplemented with new texts. Lastly, corpora can be divided into sample and full-text. The first type, as evident from its name, contains only samples of certain texts, while the other one comprises full texts.

Yet another distinction needs to be introduced for the purposes of this thesis and it is the one based on the number of languages involved in the corpus. Scholars agree that we can distinguish between mono-, bi- and multilingual corpora, however a terminology problem arises around the terms *parallel* and *comparable corpora* (McEnery et al., 2005) or *corpora for comparison* (O’Keeffe & McCarthy, 2010) – each of these terms can be used in relation to a bi- and multilingual corpus.

A parallel corpus is mostly understood as a set of texts in language A together with a set of their translations into one or more other languages, whereas a comparable corpus consists of texts written primarily in language A and language B (or more) and having a feature in common, such as the topic (O’Keeffe & McCarthy, 2010), and the same sampling frame (McEnery et al., 2005). “The *frame* or the *sampling frame* is any material or device used to obtain observational access to the finite population of interest” (Särndal, Swensson, & Wretman, 2003, p. 9). Some researchers (Aijmer et al., 1996) use the two terms inversely, that is parallel in regard to comparable and comparable in regard to parallel data, nevertheless in this thesis the definition by McEnery is followed. Parallel corpora, as McEnery asserts, might be further divided into uni-, bi- and multidirectional, depending on the direction of the translations – from language A into language B, both ways or, in the case of a multilingual corpus, translations from multiple languages into multiple languages.

The distinctions presented above entail important questions that need to be answered while designing a corpus. Those questions are addressed in Section 2.1 which contains the details of the design of the corpus compiled for this particular research.

The history and development of corpora provoke some scholars to distinguish between corpus-based studies and corpus linguistics (Kennedy, 1998; McEnery & Wilson, 2001), corpus linguistics being a methodology, as opposed to corpus-based studies which simply use corpora without a consistent methodological approach. Scholars making that distinction point to the fact that

following Chomskyan criticism of corpus data in the 1950s, the structuralists developed more sophisticated criteria for collecting real language data and put them at the centre of linguistic studies, and indicate that the term *corpus linguistics* emerged only in the 1980s (O’Keeffe & McCarthy, 2010). On the other hand, to some researchers corpus linguistics is indistinguishable from the corpus-based approach, and they point to the extensive use of computers as an important feature of it (Biber et al., 1998).

Similarly to *corpus*, the term *corpus linguistics* is defined in various ways, however most scholars underline that it is a tool, a methodology (McEnery & Wilson, 2001) and a source of evidence (Kennedy, 1998) for answering the questions about various linguistic disciplines. What is generally acknowledged is the importance of manual qualitative analysis (Biber et al., 1998) and the most significant task of a researcher being to ask meaningful questions (Kennedy, 1998; McEnery & Wilson, 2001), as corpus linguistics is not about quantitative findings but about interpreting and assessing the importance of those findings.

Using corpora changed the perception of many well-established theories and a new approach, corpus-driven (Tognini-Bonelli, 2001), as opposed to corpus-based, emerged. Put in simple words, “the corpus-driven linguists come to a corpus with no preconceived theory, with the aim of postulating categories entirely on the basis of corpus data” (McEnery et al., 2005, p. 10). That kind of approach seems to be extreme because, firstly, it rejects the centuries of linguistics history and development as well as the intuitions that, in fact, are the product of linguistic experience, and, secondly, as McEnery (2005) claims, one would need to have no linguistic education to truly practice a corpus-driven approach.

However, for these very reasons the corpus-driven approach provides new insights into languages. For example, it enables researchers to analyse the lexical grammar of English and to confirm that patterns and meaning are connected (Hunston & Francis, 2000). In recent years the corpus-driven approach has also been used for researching English in the international context and for identifying interesting phenomena, such as recognising the 3<sup>rd</sup> person singular zero (verbs in the 3<sup>rd</sup> person singular with the -s morpheme omitted) as a legitimate variant rather than a mistake or error (Dewey & Cogo, 2012). The corpus-driven approach is also successfully adapted in researching endangered languages – even though the amount of data is significantly smaller than in the case of the English language, it has enabled researchers, for example, to define a scale of language mixing (Adamou, 2016) which is a broad term for “the combination of elements drawn from two or more languages which are actively spoken in a given community as observed in the spontaneous speech of individual speakers” (Adamou, 2016, p. 212). The scale<sup>4</sup> is meant to serve as a universal tool for comparisons between and classification of various languages in contact with each other.

Corpus linguistics offers possibilities which are explored in every branch of linguistics, as well as in interdisciplinary approaches. Initially studies with the use of corpora focused on the biblical texts and belonged to the domain of lexicography and phraseology – the aim was, in most cases, to count the frequency of a given word or expression. Next, syntax and phonetics started being explored by means of corpus linguistics. Even though corpora were initially applied mainly to the study of

---

4 Even though there is a considerable amount of research concerning Russian and Belarusian mixing, the scale has never been used to discuss this particular phenomenon.

grammar, and a significant number of works are devoted to that topic (Jones & Waller, 2015), nowadays the scope of corpus linguistics is much broader and includes, but is not limited to, sociolinguistics (P. Baker, 2014), language teaching and learning (Harris & Moreno Jaen, 2011), pragmatics (Romero-Trillo, 2008), literary stylistics (Fischer-Starcke, 2010), discourse analysis (Groom et al., 2015) and translation studies (Laviosa, 2002). Some of these areas should be discussed in more detail.

One of the newest fields of research is corpus pragmatics which involves methodologies at the intersection of corpus linguistics and pragmatics. In corpus linguistics the data is typically analysed vertically – this is best illustrated in the KWIC (keyword in context) format, in which the unit that was searched for is displayed in the centre of the line and all lines containing the unit are displayed one under another. Therefore the researcher scans lines in a vertical manner.

tinkled sweetly as they moved. The hats of the <b>men</b> were blue; the little woman's hat was white
The procession moved towards the Great House. The Head <b>Man</b> watched it leave. "He likes to hear your
". What the strag will think is that any <b>man</b> who can hitch the length and breadth of the galaxy
seem. For instance, on the planet Earth, <b>man</b> had always assumed that he was more intelligent than dolphins
masturbating while watching a television set on which a young <b>man</b> is lying on his side and masturbating while watching a
snail in a shell, where the soul of the <b>man</b> cowered in fear of the natural blaze of life.
on them!" Arthur didn't notice that the <b>men</b> were running from the bulldozers; he didn't notice
—why do you avoid life?" But the young <b>man</b> was not listening. He was peering past the old
criminals are on the track; but even the rough <b>men</b> about in the streets and hotels could hardly have kept
He crouched among them, glared back at the Head <b>Man</b> in fury and shouted the words at the top of
. The screen in front of him shows a young <b>man</b> lying on his side and masturbating while watching a television
Of course, Great House! What else does any <b>man</b> need?" "His potter," said the
blinked furiously. "There, " said the Head <b>Man</b> . "You were doing so well but you spoil
shouted Arthur. He pointed at Prosser. "That <b>man</b> wants to knock my house down!" Ford glanced
husband, who has just ejaculated in her. The <b>man</b> watches her unconcernedly. As she starts to come lhe
by him. The crowd was swirling round the blind <b>man</b> who became a toy, a shouting doll. Pretty
. He did not know why that gross, obese <b>man</b> excited in him so violent a repulsion. But his
Prince shrank down behind his boulder and listened as the <b>man</b> went back. He was trembling, and went on
with the possibility that he might fall in with white <b>men</b> who would reveal his hiding-place; perhaps the
the rest. On the green lawn before it many <b>men</b> and women were dancing. Five little fiddlers played as
an American sailor, and he had deserted from a <b>man</b> -of-war in Apia. He had induced
and the God smiled in his sleep. The Head <b>Man</b> took his arms and folded them across each other so
the Wailing Well field consists of three women, a <b>man</b> , and a boy. The shock experienced by Algernon
just because the skipper was so gross and dull a <b>man</b> the whim seized him to talk further. "You
them." "'Anyone'," said the Head <b>Man</b> , "would blink or rub them. That's
unnoticed. The French police and our own Scotland Yard <b>men</b> , and the military are straining every nerve. It
located in one of Seoul's nicer neighborhoods, a <b>man</b> thirty-seven years old has surrounded himself with sumptuous
masturbating while watching a television set on which a young <b>man</b> is lying on his side and masturbating while watching a
side of it was a small pack of rather ugly <b>men</b> who they could only assume were the heavy mob of
forced inactivity. Finally the door reopened, and the <b>men</b> emerged, escorting three prisoners - a woman and two
if his life was held in it. The Head <b>Man</b> spoke quietly, like a physician explaining a disease.
And then he approached me one late afternoon, a <b>man</b> of nearly seventy who was still broad-shouldered and
companion masturbating. The latter is a gray-haired <b>man</b> of fifty-four, who looks at the young
yet I seem to see him more clearly than many <b>men</b> , my brothers, for instance, with whom I
"Guess I gotta fill in for the old <b>man</b> this year." "Right." Mr. Summers
told you of." "Is he a good <b>man</b> ?" inquired the girl anxiously. "He is
. He wanted to stay on friendly terms with these <b>men</b> . Even he, the Supreme Commander, might need
"What d'you mean, three women and a <b>man</b> ?" said Stanley, turning over for the first
the main road? But if he was a dishonest <b>man</b> , why did he start the car again when only
trundled away. In the sudden silence, the Blind <b>Man</b> was heard at last. "The Prince is going

*Illustration 1.1. Query from the EPB corpus in a KWIC format.*

In pragmatics, on the other hand, data is usually read horizontally, because what is surrounding the searched unit – the context – is most important. The idea of combining the vertical and the horizontal analysis is tempting but difficult in realisation because of a mismatch between form and function in most pragmatic phenomena. A good example of that is negation. A short phrase “Yes, sure” usually means agreement, however if said in ironical tone it is quite the opposite, and the only way of determining the function of these two words is by looking at the context. The problem with

form and function mismatch shifted the focus of pragmatic research by means of corpus linguistics to conventionalised speech acts (e.g. Aijmer, 1996).

The solution for that problem, however far from being perfect, is pragmatic annotation. While this kind of mark-up allows examination of form and function dependencies, it is difficult to implement because it cannot be done automatically, it always demands human intervention and thus high resources usage. Nevertheless, currently corpora are exploited in nearly all pragmatics sub-fields: speech acts (Garcia McAllister, 2015), pragmatics principles (Diani, 2015), pragmatic markers (Aijmer, 2015), evaluation (Partington, 2015), reference (Rühlemann & O'Donnell, 2015) and turn-taking (Clancy & McCarthy, 2015).

Regarding discourse analysis, one has to remember that *discourse* is understood in multiple ways. Two basic meanings of the term refer to different phenomena: either to the interaction between interlocutors, or to the analysis of people's beliefs and principles, otherwise referred to as ideology. It is the second meaning that is exploited in critical discourse analysis, a field that changed considerably as a result of introducing corpus linguistics methodology.

Even though the variety of tools accessible in corpus linguistics is impressive and provides undeniable quantitative evidence for critical discourse theories, it is interpretation, as noted above, that continues to be an essential element of corpus-based research. It is especially meaningful in discourse analysis, as is exemplified in Paul Baker's (2015) experiment, for which he gathered five scholars to work independently and use exactly the same dataset. The outcomes of these autonomous analyses differed considerably and thus the experiment proved that the researcher's bias cannot be removed completely and even though the data is objective, the way of looking at it remains subjective.

What is particularly important for the project discussed in this thesis is corpus in the context of translation studies and contrastive studies. Corpora are a point of interest for applied, descriptive and theoretical translation studies with a considerable number of works dedicated to each of the three domains. Mona Baker (1993) was the first scholar to use corpus linguistics as a methodology in translation studies, and according to Laviosa (2002) it was Baker's work that actually caused corpus-based translation studies to become a new paradigm in translation studies rather than just a methodology.

Theoretical translation studies, the aim of which is to identify general principles that can be used to explain and to anticipate certain phenomena in the process of translation, apply corpus data in the search for translation universals (TU) that itself started over three decades ago. Frawley (1984) put forward the notion of translation as *third code*, indicating that the product of a translation process differs substantially from both the source and the target language. That notion, in turn, became an inspiration for searching for the TU in what is nowadays known as *translationese* (Gellerstam, 1986). Mona Baker (1996) defined TU as certain linguistic characteristics found in translations simply by the virtue of the text being translated from another language, and investigated them systematically. Three universals are most frequently a point of interest for other scholars: simplification, explication and textual conventionality (normalisation).

Simplification, defined as a “tendency to simplify the language used in translation” (P. Baker, 2012, p. 250), was already investigated in the pre-corpus era (Blum-Kulka & Levenston, 1978), however according to Laviosa (2002), the evidence of different strategies used for simplification which has been presented in studies (such as Klaudy, 1996; Toury, 1995), is not coherent as it is derived from various analyses that cannot be compared due to the differences in research questions, investigated data etc. Because of that it cannot be used for generalisation. However, more systematic research could confirm the existence of simplification in translation, It could also determine whether the discrepancy of lexical organisation and representation of cultural concepts in any language pair or the translation itself causes simplification.

Mona Baker (1996) conducted such a systematic investigation using corpora and identified some indicators of simplification: lower average sentence length, lower lexical density (ratio of the number of content words to the number of grammatical words) and lower lexical variety (type-token ratio). Similarly, Laviosa (2002), investigating the English Comparable Corpus (Laviosa, 1996), tested three hypotheses: lexical variety, information load and sentence length, and found them to be consistent with the notion of simplification being a TU. Many other scholars, e.g. Scarpa (2006) who investigated specialised English-Italian translations, also confirmed these discoveries.

The second TU frequently referred to is explicitation. It was first recognised in the 1950s as a “translation technique involving the introduction in the target language of information which is only implicit in the source language, but which is retrievable from the context or situation” (Laviosa, 2002, p. 32), and later Blum-Kulka (Blum-Kulka, 1986) suggested the process of translation rather than linguistic and cultural differences causes explicitation. Øverås (1998) conducted one of the first systematic studies of this phenomenon; going beyond intuition he investigated a bi-directional English-Norwegian parallel corpus. Later also Olohan and Mona Baker (2000), as well as Olohan (2002) raised the issue of explicitation in their research on translational and original English.

The third TU explored in translation studies is normalisation, that is the translators’ tendency to render distinctive features of a given text in such a way that they are in accordance with the distinguishing features of the target language, sometimes to the point that these typical characteristics are exaggerated. Normalisation has been an object of theoretical and empirical studies, however little data supports the notion. Corpus-based studies provided scholars with new facts about normalisation and put forward new methodologies that can be further used in the investigation of the phenomenon in different languages pairs. Examples of such studies are Scott’s examination of Portuguese-English translation (1998) and Kenny’s (1999) research into German-English translations.

In terms of descriptive translation studies (DTS), which are primarily concerned with the question of why a particular text was translated in one way and not another, the following typology is adopted: product-oriented DTS, process-oriented DTS, function-oriented DTS (J. Holmes, 1988). Data and tools provided by corpus linguistics have been used to explore each of these branches of the descriptive approach to translation.

Regarding DTS oriented to the product of translation, most of the research conducted in that domain is focused on delivering evidence supporting the notion of TU – even though, according to some researchers (Tymoczko, 1998), scholars cannot identify TU because in order to do that they would have to collect translations from all times and all languages. However, another interesting and broadly explored area of product-oriented DTS is the translator's style.

Stylometry – the quantitative study of writing style – has been used in many studies concerning various pairs and directions of translation, e.g. from English into Turkish (Patton & Can, 2012), from Spanish and Portuguese into English (Saldanha, 2011a) and from Chinese into English (Ji & Oakes, 2012). Apart from typical corpus linguistics tools, dedicated software packages for stylistic analysis, such as Stylo, are available (Eder et al., 2016) and successfully used in translation studies (Cheesman et al., 2017; Rybicki, 2012).

With respect to process-oriented DTS, it attempts to explore what is happening in the mind of a translator, sometimes applying Think-Aloud Protocols (Bernardini, 2002), however it is difficult to examine the mental processes in isolation from its product. An example of such an approach is Munday's (1998) examination of Spanish-English translations of a short story. The analysis he conducted provides an insight into the decision-making process and enables him to deduce from it the translational norms followed by the translator. What is interesting is that, instead of simply describing the material, Munday uses general corpora for comparison and explaining the shifts.

In the case of function-oriented DTS the sociocultural context is most important. This is exemplified in the work undertaken by Laviosa (2000), who analysed semantically related words that are significant in the socio-cultural context and investigated the impact that the translated texts might have on the set of beliefs of the target audience. Similar studies are conducted also for minority languages, such as Zulu (Masubelele, 2004). The purpose of the study was to investigate what role the translation of the Bible played in the expansion of written Zulu in South Africa.

As far as applied translation studies are concerned, special attention is being paid to one aspect, namely teaching and learning. Corpora – comparable, parallel and monolingual in the target language – are used in multiple ways. One of the most obvious is analysing them by means of KWIC concordances in order to find the expressions that sound most natural while conveying the same meaning and having the same function in the target language. Studies (Bowker, 1998) indicate that specialised translations performed with the help of specialised corpora are of higher quality in parallel terms of understanding the field that the source text is involved in and in terms of choosing the appropriate terms and idiomatic expressions. Apart from teaching translators, corpora are often used as a resource for teaching foreign languages – data retrieved from parallel corpora has been frequently used, for example to create exercises enabling the students to infer the grammatical rules from the examples (Laviosa, 2002).

In summary, it has been demonstrated in this review that nearly all linguistic disciplines have gained from the introduction of corpus linguistics methodology. Corpora have enabled researchers dealing with a wide variety of tasks, to answer questions about theoretical issues, improve existing knowledge and enhance performance.

Having defined what is meant by corpus linguistics and what its applications are, we will now move on to discuss the specificity of the field in the three languages involved in this research: English, Polish and Belarusian. This section will start with reviewing the English language as it is the language that corpus linguistics started with and by now has the biggest advances in the field, therefore it constitutes a point of reference for all other languages.

#### **1.4. Overview of corpus linguistics resources in English, Polish and Belarusian**

The first element to consider is the wide variety of corpora in English. A noticeable fact is that the first computerised corpus – the Brown corpus (Kucera & Francis, 1964) – was a compilation of texts written in English. The same applies to most of the early written corpora, such as:

- Lancaster-Oslo/Bergen (Johansson, 1978), another corpus of British English,
- Kolhapur (Shastri, 1986), the corpus of Indian English,
- ACE – the Australian Corpus of English (Peters, 1989).

Several varieties of English are present also in modern corpora:

- British in the British National Corpus (University of Oxford, 2015),
- American in the Open American National Corpus (American National Corpus Project, 2015) and in the Corpus of Contemporary American English (Davies, 2008),
- Canadian in the Strathy Corpus ('Corpus of Canadian English (Strathy)', n.d.),
- Australian in the AusNC (Australian National Corpus Inc, n.d.),
- multiple varieties of English in the corpus of Global Web-based English (Davies, 2013) and in the International Corpus of English (The ICE Project, 2016).

All of the corpora listed above are general written corpora (in some cases with small parts of spoken materials) and they are sometimes referred to as *mega corpora* (Weisser, 2016) by virtue of their size, however they constitute only a fraction of the total number of corpora containing materials in English. Other types of corpora include:

- spoken, such as the London-Lund Corpus of Spoken English (Greenbaum & Svartvik, 1990), the Machine-Readable Corpus of Spoken English (L. J. Taylor & Knowles, 1988), and the Hong Kong Corpus of Spoken English (Cheng et al., 2005);
- academic, such as the British Academic Spoken English Corpus (Nesi & Thompson, 2006) or the Michigan Corpus of Academic Spoken English (The University of Michigan, 2007);
- learner, including the International Corpus of Learner English (Granger et al., 2009) and the Louvain Corpus of Native English Essays (Centre for English Corpus Linguistics (CECL), 2018);
- pragmatically annotated, for example the Speech Act Annotated Corpus (Leech & McEnery, n.d.) or the Switchboard Dialog Act Corpus (Potts, 2011);

- diachronic, such as a Representative Corpus of Historical English Registers (The University of Manchester, n.d.) and the Corpus of Historical American English (Davies, 2010).

The examples listed above are a small sample of the multitude of English language corpora. Another significant aspect in the context of development of corpus linguistics is the existence of research institutions dealing exclusively with questions of that field. An example of such is the Centre of Corpus Research at Birmingham (University of Birmingham, 2017) that provides corpus resources and tools, such as the CLiC web application (University of Birmingham & University of Nottingham, 2017) that has been primarily developed to analyse Dickens's novels but can be also applied to other literary texts. Another research centre offering an impressive choice of resources and tools is University Centre for Computer Corpus Research on Language (Lancaster University, 2014). These centres and other units of the kind are occupied not only with research but also with education – as demonstrated by lectures, workshops and summer schools they organise. Corpus linguistics is offered as well as a course in pursuit of the Master of Arts degree (University of Wolverhampton, 2018).

What is significant is the broad co-operation between many institutes and universities that is exemplified not only in joint projects of the centres mentioned and outside partners but also in the existence and continuous dynamic development of consortia. An important example is CLARIN (Common Language Resources and Technology Infrastructure) with 21 European countries as members and 4 countries (including South Africa and USA, which makes the project intercontinental) as observers or third party members, and within each country a number of units from research institutes and universities. The United Kingdom has currently a status of observer but it offers a wide range of resources for the English language (CLARIN-UK, 2018).

Turning now to corpus linguistics in the domain of the Polish language, it is worth mentioning that even though it has been developing for a much shorter time than corpus linguistics in English, it already has significant achievements. Researchers on the Polish language can use a variety of corpus resources. These resources include:

- the National Corpus of Polish (National Corpus of Polish, 2012), which is an example of a mega corpus with primarily written language data;
- Spokes (CLARIN PL, 2014), the corpus of spoken language;
- Monco ('Wyszukiwarka korpusowa Monco [Corpus browser Monco]', 2016), a monitor corpus;
- Paralela (Pęzik, 2016), a Polish-English parallel corpus;
- a historical corpus of 15<sup>th</sup>/16<sup>th</sup> century texts by Ioannes Dantiscus (Skolimowska & Turska, n.d.) and KORBA (The Institute of the Polish Language at the Polish Academy of Sciences (IJP PAN), n.d.), a corpus of 17<sup>th</sup> and 18<sup>th</sup> century texts; apart from that few other projects, involving even older texts, are being developed (Ogrodniczuk, 2013);
- PELCRA Learner English Corpus (Uniwersytet Łódzki, 2013);
- a corpus of 21<sup>st</sup> century youth language (Laboratorium językowe, 2014).

Apart from the existing corpora, researchers are developing a number of new projects, such as the Polish Sign Language Corpus (Pracownia Lingwistyki Migowej, 2019) which only recently (in December 2019) has been translated into the written Polish language and is currently undergoing the process of annotation.

In terms of tools dedicated to the analysis of the Polish language, CLARIN-PL (CLARIN PL, 2015) should be noted as the most significant contributor. Services offered by CLARIN include tools for corpus annotation, such as the morphosyntactic tagger (Walkowiak & Kaliński, 2013) or the system for editing annotated corpora, Inforex (Marcinčuk et al., 2017), as well as tools for syntactic, semantic and stylistic analysis of text. An important part of CLARIN's activity are its educational efforts. Workshops, repeatedly organised in major Polish cities, introduce text analysis tools to researchers from various branches of linguistics.

CLARIN, as mentioned above, is a consortium gathering various institutes, but corpus linguistics is also an object of interest of Polish universities. Development of dedicated units (University of Warsaw, 2020), introducing corpus linguistics as a subject in some linguistics courses and researchers' engagement in international projects clearly demonstrates an interest in the well-being of this field in Poland.

In contrast to corpus linguistics in English and Polish, the field is under-resourced in the Belarusian language. The biggest corpus of Belarusian, Belarusian N-corpus (BNkorpuz, 2019), offers access to about 163 million tokens (which is half of the balanced part of the National Corpus of Polish and 10% of the total size of the Polish corpus) and few tools for linguistic analysis. Apart from the BN-corpus only a handful of resources containing Belarusian language are accessible to the public. These resources are:

- beTenTen Corpus (Lexical Computing CZ s.r.o., n.d.) containing 63 million tokens crawled from the Internet in October 2016;
- a Biblical Corpus (BNkorpuz, 2017) containing 16 Belarusian and 6 non-Belarusian translations of the Bible;
- Corpus Albaruthenicum (Belarusian National Technical University, n.d.) containing scientific texts;
- a parallel Russian-Belarusian corpus (Nacionalnyj korpuz russkogo jazyka, 2017).

Apart from the sources that are primarily concerned with the Belarusian language there are multilingual corpora that contain materials in the Belarusian language, such as the ParaSol corpus (von Waldenfels, 2011) with nearly 0.5 million Belarusian words, the Amsterdam Slavic Parallel Aligned Corpus with 200,000 Belarusian tokens (University of Gothenburg, 2016), and the Belarusian newspaper corpus based on material crawled in 2011 (Deutscher Wortschatz, 2018) with almost 7.5 million tokens.

Despite the lack of corpus resources and tools, computational linguistics in the Belarusian language is developing really well due to the efforts of two communities. First, NLP group Belarus ('Natural Language Processing group Belarus', 2017) gathers people concerned with computational

linguistics and willing to share their experience. Members of the group share their projects in the GitHub repository (nlproc.by, 2018), while on their website they publish interviews with NLP engineers and linguists, reports about their meetings, and information about institutions in Belarus and abroad where it is possible to study computational linguistics. The second community contributing to the development of the field is the Speech Synthesis and Recognition Laboratory (Speech Synthesis and Recognition Laboratory, 2018b), which is a part of the United Institute of Informatics Problems of the National Academy of Sciences of Belarus. Researchers working in that institution have created multiple tools for language processing (Speech Synthesis and Recognition Laboratory, 2019), and some of them, such as Part-of-Speech Tagger, are certainly useful in corpus linguistics.

This section has reviewed three key aspects of the field of this research: corpus linguistics basic definitions, its applications and its situation in the English, Polish and Belarusian languages. It has been revealed that the Belarusian language is under-resourced in terms of available corpora and tools for corpus analysis. This state of affairs is due to the differences in the status of the three languages under consideration, as discussed in Section 1.1.

### 1.5. Corpus linguistics as methodology

According to McEnery (2011), corpus linguistics “is an area which focuses upon a set of procedures, or methods, for studying language”. This section explores some of the available methods, and tools that incorporate them. The most popular methods for analysing the data in corpus linguistics can be listed as follows:

- (a) wordlists (Kennedy, 1998), frequency counting (McEnery et al., 2005) or simply frequency lists (Crawford & Csomay, 2016); frequency – the number of occurrences of a word within a dataset – can be an indicator of markedness and “can be used (with supporting contextual or additional information) in order to uncover evidence for bias” (P. Baker, 2010, p. 125). What Baker understands as being “marked” is one concept being less frequent than the another and thus, in Cixous’ (1975 in Baker, 2010) terms, being the ‘outsider’ (such as *abnormal* and *unnatural* compared to *normal* and *natural*) or one concept being more frequent, such as *homosexual* versus *heterosexual*- “in this case, homosexuality is marked *because* society views it as unusual” (P. Baker, 2010, p. 126; emphasis original).

One needs to remember a few important aspects when analysing wordlists. Firstly, the frequency should be regarded in relation to the corpus size (McEnery et al., 2005) – frequency count 100 is less important in a 100-million than in a 1-million word corpus. Secondly, Kennedy (1998) points out the importance of interpretation of results and “quite radical differences” (1998, p. 247) between different domains. Thirdly, Weisser (2016) indicates another two issues: definition of a *word* (this could be misleading taking into consideration spelling, e.g. *icecream/ice-cream/ice cream*) and the differentiation between types and tokens. What is usually represented in a frequency list is a *type* or *lemma* (e.g. *say*) and the count of each individual occurrence of that type, that is the number of *tokens* (e.g. *says, said, say*). Since tokens may occur in all forms, not only basic, a corpus should be *lemmatised* in order to avoid misleading results. In a lemmatised corpus each word is

assigned a *lemma*, or basic form, and the search is conducted in fact within these basic forms.

Yet another aspect influencing the objective analysis of the frequency list is the word's dispersion or "the rate of occurrence of a word or phrase across a particular file or corpus" (P. Baker et al., 2006, p. 59); usually, even dispersion in the corpus is desired and researchers use a minimum threshold for the number of documents containing a specific word. If one text from the group is highly focused on one topic and contains a lot of occurrences of one word, it may skew the picture of the whole group, as this particular word is characteristic of one text, not the collection. However, if one is interested in identifying a single text's features, such as a topic, a density of a searched word in specific texts can be an indicator of it.

The method of analysing wordlists is used in Section 3.1.

- (b) linguistic distribution (P. Baker et al., 2006); this is the analysis of frequency across various genres of text, as well as across groups of texts divided according to other factors, such as the sex or age of the authors (or speaker in the case of spoken texts); this method allows one to differentiate the usage of synonyms across various registers or otherwise classified texts. This method is used in Sections 3.2 and 3.4.
- (c) keyword lists (Weisser, 2016); they are based on "comparing two word lists, one from the corpus under investigation – referred to as the *source* corpus – and the other from a (general) reference corpus – the *target* corpus. The output then shows the *significant differences* as a filtered word frequency list of the corpus that's being analysed" (Weisser, 2016, p. 169). This method is used in Section 3.2.
- (d) concordances (Kennedy, 1998) and KWIC (Crawford & Csomay, 2016); concordances are certain strings, words or phrases, searched in the corpus and displayed in KWIC (keyword in context) form. These outcomes may be sorted according to the context on the left and/or the right side which shows the relations between the query and its usual position in the sentence. This list illustrates "how the word [...] spreads within each of your texts" (Crawford & Csomay, 2016, p. 88). Concordances become really useful if they are analysed systematically (Tribble, 2010): the evidence not only has to be observed but also sorted by a feature; general patterns that might be observed on the right or the left side of the keyword have to be interpreted and linked by a common hypothesis and then further inspected in terms of any variation in the pattern or possible additional patterns. This method is used in Sections 3.1, 3.2 and 3.4.
- (e) collocations (Crawford & Csomay, 2016; McEnery et al., 2005) and colligations (P. Baker et al., 2006); collocates are frequently, that is more often than one would expect by chance, co-occurring strings, words or phrases. Although examples presented in the literature usually present collocates next to each other (e.g. sour milk, spend money) it is important that they could be separated by other words (e.g. milk was sour, spend that money) and this is very often the case in the Polish language that has a free word order. Colligation, on the

other hand is a frequent co-occurrence at grammatical level, e.g. adjectives usually colligate with nouns; colligates are recognised also in relation to particular words or phrases, e.g. “a word like *window* tends to colligate with prepositions” (P. Baker et al., 2006, p. 36). This method is used in Chapter 4.

- (f) n-grams (Crawford & Csomay, 2016; Weisser, 2016); n-grams are “sequences of words explored as a unit, where the value of *n* denotes how many words there are in that unit” (Crawford & Csomay, 2016, p. 41). When looking for a one-word unit one might talk about uni-grams (which were discussed in the first two methods above). Bi-grams are different from collocates because they occur in sequence, one after another, whereas collocates do not have to keep to the order (as demonstrated in the previous point). Therefore not all bi-grams are collocates, and not all collocates are bi-grams. A specific term is connected to tri- and four-grams. “When these [...] combinations occur at least 10 or 20 or more times in a million words [...] and appear in at least five different texts (to avoid idiosyncratic [...] use) in a register, they are referred to as *lexical bundles*” (Crawford & Csomay, 2016, p. 49). Lexical bundles were first identified and then thoroughly described by Biber (1999). This method is used in Section 3.2.
- (g) type/token ratio (P. Baker et al., 2006); this is the proportion of the number of types – unique words or lemmas – to the number of tokens – all word occurrences in the text; TTR is regarded as a measure of the lexical richness of a text and while its value is strongly related to the size of the corpus – the bigger the corpus the lower the ratio – it is useful when comparing two texts, especially in a parallel corpus (Merkel, 2001). This method is used in Section 3.1.
- (h) lexical density or lexical richness (P. Baker et al., 2006); this measure is interpreted in different ways by various researchers – some identify it with TTR, mentioned above, others calculate it with the percentage of the content words (or unique content words), that is words having lexical meaning in the text (or in the text without function words, that is without words fulfilling a grammatical function only).
- (i) POS tags (Crawford & Csomay, 2016) and regular expressions (Weisser, 2016); POS (part-of-speech tagging) is an additional layer of information encoded within the text and it enables, for example, content words to be sorted from function words. Regexp are chains of characters that allow for describing patterns and are an integral part of system tools as well as text editing programs and programming languages. Used together with POS tags they enable a more flexible search for structural patterns in the data, for instance nouns with descriptors of certain type and separated by other words, such as in *my little brother*, where the possessive pronoun is separated from the noun by an adjective.
- (j) multifeature/multidimensional (MF/MD) analytical framework (Biber, 1992); this is an approach to “the linguistic analysis of texts, genres, text types, styles or registers” (Biber, 1992, p. 332), which enables data to be differentiated in five dimensions defined by Biber through the analysis of 67 linguistic features and patterns of co-occurrence among them.

This method was originally developed for comparing spoken and written English, but it has been applied to other languages, such as Russian (Katinskaya & Sharoff, 2015).

- (k) vector-based semantic similarity measures (Huerta, 2008); this is a group of metrics in which the chosen units (sentences, words or others) are represented as vectors – in mathematics and physics these are quantities represented as directed lines having magnitude (which corresponds with the length of the line) and direction (which corresponds with the direction of the arrowhead attached to the line). The distance between vectors represents the semantic distance between the units.

The elementary vector-based measure is the cosine distance – the lower the distance between the vectors (or the higher the cosine value) the higher the similarity between the units. This method is domain and model free and in the case of units with low frequency it provides better outcomes than other methods when the task is to measure similarity of meaning across different linguistic expressions.

- (l) cluster analysis (Moisl, 2015); this method is the superordinate of the previous one and it is a “family of mathematically-based computational methods for identification and graphical display of structure in [large sets of] data” (Moisl, 2015, p. 153) organised in clusters based on their relative similarity. Cluster analysis deals with raw, not annotated data and treats words as strings of characters, therefore it is language-neutral, This method is used in Section 3.2.

While most of the methods described in this section can be used on raw, not annotated data, it is important to remember that for the most reliable results text should be at least lemmatised – each word should be assigned its base form – to avoid misleading conclusions. And while most of these methods could be exploited using simple calculating programs, such as Microsoft Excel, numerous tools are specially designed for such tasks.

It is not possible to count all of the tools available for corpus linguistics. The online portal *corpus-analysis.com* accounts for 111 independent tools, such as software with web- and computer-based interface or libraries for building programs in various programming languages, and these are for analysing English-language corpora alone. One also needs to remember about built-in tools for analysis of online corpora and selected functions of statistical programs, as well as tools designed for languages other than English. The possibilities are endless, however the range of available solutions differs depending on the language, and the relation between the number of tools and the status of a language is the same as in the case of available resources (see Chapter 1.4)

This section describes the tools and software designed to perform corpus analysis, indicating at the same time languages that are supported – English, Polish or Belarusian. The review starts with independent software, with both online and computer-based interfaces. Next, tools available with the most popular corpora are briefly described. Lastly the libraries available for a few programming languages are mentioned. The order and level of detail given for each group is not random, as only the independent software is used for exemplifying the use of the EPB corpus in Chapters 3 and 4.

One of the most popular toolkits for corpus analysis is AntConc (Anthony, 2018) and its popularity stems from the fact that it is not only freely available but also supports an open-source operating system, namely Linux. The program's website provides an extensive user's guide and video tutorials. AntConc contains several tools for corpus analysis: concordancer, concordancer plot (to illustrate the dispersion of results), n-grams, collocates, wordlist and keyword list. It allows for the analysis of raw data saved as plain text, as well as an annotated corpus in XML format. Even though the tools offered in the program are not very advanced, the wide range of options gives the researcher flexibility of analysis. AntConc does not require specifying the language of analysis and it works well with Belarusian, as long as the character encoding in files matches the one declared in settings.

Another well known computer-based toolkit is WordSmith (M. Scott, 2017). Its use is more restricted than in the case of AntConc – the toolkit operates on Windows only and a paid licence is required. WordSmith provides all tools that are available in AntConc and additional *utility programs*, such as *Aligner* to align translated texts or *CorpusChecker* that identifies anomalous text and helps clean the corpus, thus the toolkit is very useful for researchers building their own corpus. WordSmith supports languages in non-Latin alphabets, as long as an appropriate encoding is defined in the settings.

Probably the most comprehensive program is Sketch Engine (Kilgariff et al., 2014) which is not only a toolkit, but also a corpus management system. Its advantage is that it operates on an online basis so the operating system used by a researcher does not pose any problem, however a paid licence is required<sup>5</sup>. SketchEngine provides tools for corpus analysis and for corpus building, and it has a built-in POS tagger, in contrast to previously discussed toolkits. SketchEngine offers also some preloaded corpora and facilities to store corpora built for the purpose of the research. It supports over 90 languages, including Belarusian, although it does not provide POS tagging for it.

Apart from these three popular comprehensive programs for corpus analysis and corpus compilation, a range of tools specialise in only one task, such as annotation or concordancing. Among the largest providers of language analysis software one should name the CLARIN consortium (mentioned in Section 1.3). It gathers 21 member countries that work separately to develop tools and resources for their national languages and network for collaborative projects. Two portals summarise the work of CLARIN: Language Resource Switchboard (Zinn, 2018) which searches for tools to analyse text in a given language (not only from CLARIN resources but also from open-source tools accessible online) and VLO, or Virtual Language Observatory (CLARIN ERIC, 2018) which is a browser of resources available in the consortium and related communities.

Apart from such comprehensive toolkits, some tools are specifically designed for research in a given area. This section focuses on several of them that are used in the chapters presenting the applications of the EPB corpus.

---

5 The access to Sketch Engine for the academic institutions and exclusively for non-commercial use will be funded by the EU through the ELEXIS infrastructure project between 2018 and 2022.

An interesting program is WebSty (Walkowiak & Piasecki, 2018) – it is a free and on-line based application for analysis of texts' similarity, and in fact it combines three other tools with visualisations. The three tools comprising the base of WebSty are:

- Apache Tika (The Apache Software Foundation, 2018) – tool for detecting and extracting metadata and text from different file types;
- UDPipe (Institute of Formal and Applied Linguistics, Charles University, 2018) – tool for tokenisation, tagging, lemmatisation and dependency parsing; supports all three languages from the EPB corpus;
- SuperMatrix (Language Technology Group G4.19, 2018) – tool for automatic extraction of semantic relations.

WebSty offers four methods of analysis: authorship analysis on the basis of stylometric features, analysis of grammatical style, grouping on the basis of co-occurrences of particular words, and classic authorship analysis based on the co-occurrence of particular forms of words. The program offers a wide range of options regarding grammatical features one might want to include or exclude from the analysis, and options for calculating the similarity. The analysis provides results in eight formats: interactive dendrogram, heatmap, multidimensional scaling in 2D and 3D, schema ball, pie graph, importance of features (with additional settings) and .xlsx file. The richness of available options makes WebSty an excellent toolkit for literary stylistics researchers.

In the case of translation studies the Version Variation Visualization (VVV) web-based platform (Cheesman et al., 2012) provides tools for comparison of translation versions. Registered users (registration for non-commercial use is available for free) can upload their own corpora (saved as plain text), VVV automatically aligns the base text with the translations and the user is given a few options of analysis. Firstly, one can search through the text by the means of *alignment view*, which allows for exploring the aligned parts of text and a quick overview of possible omissions and re-arrangements in the translations. Another possibility is to compare translations using two measures of differences: Eddy, which determines how much one segment in a translation differs from all other versions of this segment, based on the words used, and Viv, which compares the set of Eddys assigned to translations of one source text segment to the set of Eddys assigned to another sources text segment. Visualising results on the graph presents a clear overview of the outcome of analysis.

A tool helpful in conducting discourse analysis and created within the CLARIN-PL consortium is the Polish-English Wordnet (Wrocław University of Technology, 2016). It illustrates the relations between words in a form of net and describes types of these relations (synonymy and equivalence, hyponymy and hypernymy, meronymy and holonymy) in both. Additionally, words are assigned one of eleven registers, and are annotated with emotional attitude (positive, neutral, negative), as well as basic emotions and universal values associated with this word (according to Puzynina, 1992). Even though the P-E Wordnet is still in the process of correction (specially in terms of emotional annotation) it provides researchers with an insight into the meanings associated with various terms used in a text and enables making comparisons between Polish and English languages. Currently CLARIN's wordnet contains no other languages, however a Belarusian

wordnet assembled as a dataset without user interface is available, and the possibility of expanding P-E Wordnet and making it trilingual is currently being discussed.

This section began with a short description of corpus linguistics methods arranged from the most basic to the most complex. It went on to discuss some comprehensive computer software incorporating these methods and then to present programs designed to deal with tasks related to particular research domains. This outline of corpus linguistics methodology will be elaborated on in Chapters 3 and 4 on the occasion of exemplifying the EPB corpus use in various areas of study.

## **1.6. Outline of the next chapters**

This introductory chapter is followed by four others. Chapter 2 describes the process of corpus compilation and is divided into five sections. The first section offers a detailed account of the corpus design and all issues stemming from it: justification of the purpose, consideration of the availability of the texts, as well as related legal issues, discussion on the topic of corpus size, representativeness and sampling. This section also presents possible solutions to the problems under consideration and explains the reasons behind the chosen course of action. The second section of the chapter focuses on text collection and the technical procedures used in the case of different types of texts – digital and analogue, Polish, English and Belarusian – and explains the means used to unify the data formats.

The third section of the Chapter 2 moves on to describe the types of encoding chosen for the corpus, available tools for annotation, their evaluation in respect to the three languages used in the corpus, and consequently justification of the final choice. Section four is a brief description of tools for multilingual corpus alignment. Similarly to the previous section, their evaluation and an explanation of tool choice is given. In the last section of the Chapter 2 the issue of corpus data storing and sharing is discussed, with a special attention to the outreach of the resource and its meaning for the Belarusian-speaking community.

Chapters 3 and 4 present a handful of case studies exemplifying applications of the EPB corpus. The main focus is on the translation, however other fields are exemplified too. Firstly, the theoretical explorations concerning translation universals, particularly simplification, levelling out and explicitation, are demonstrated. Using methods and indicators already utilised by other scholars it is possible to determine to what degree the claims of the universality of translation features are tenable. Secondly, Chapter 3 addresses the topic of DTS in a stylometric overview of the whole corpus. This overview then prompts a detailed investigation of one intriguing case. Next, the example of using the EPB corpus in the area of applied translation studies is provided. The case presented in this part concerns analysing phraseology for the purposes of training translators. Finally, Chapter 4 approaches the field of discourse analysis. By determining features of the discourse surrounding men and women in the three subcorpora and subsequent comparison with features identified in the corpora of original Polish and Belarusian texts, it is possible to estimate the influence of the original on the translation at the level of discourse.

Chapter 5 looks ahead and discusses the planned path of development for the EPB corpus and steps taken in order to secure steady growth of the resource. What is presented in this final part are the

pros and cons of possible solutions and the discussion of functionalities of particular infrastructures which are implemented for the purposes of the EPB corpus.

## Chapter 2: Corpus compilation

According to Kennedy (1998, p. 70) “there are three main stages in corpus compilation: corpus design, text collection or capture, and text encoding or markup”. Those issues are considered in this chapter alongside the alignment of the data – which is specific for parallel corpora. Additionally the topic of storing and sharing the resources is discussed in the last section. Firstly, the corpus design topic is broken into several interconnected problems raised by scholars: purpose, availability of texts, size and representativeness, legal issues and sampling. All of these are discussed in relation to the three languages included in this project, that is English, Polish and Belarusian. Next, the workflow of the text collection is presented. The focus in this section is on the sources for the EPB corpus, as well as on the formats and programmes utilised in that task. What follows is the discussion of encoding of the corpus, particularly adding metadata, tokenising, lemmatising and POS tagging with the indication of problems concerning the corpus languages in each of the tasks. Afterwards the questions of alignment are considered, particularly what is the basic unit of alignment, what types of relation exist and can be encoded, as well as what are the differences between various ways of aligning the texts. In the final section storing and sharing the data is covered, with special attention to the issues of choosing the right format, providing extensive documentation, finding a suitable repository and advertising the resource.

### 2.1. Corpus design

Designing a corpus involves analysing certain issues (such as purpose, feasibility or scope), and setting criteria for text selection based on the outcomes of such an analysis.

- a) Purpose (Crawford & Csomay, 2016, pp. 76–78; Kenning, 2010, pp. 489–490; Nelson, 2010, p. 54)

Building a new corpus should be preceded by a thorough search for already existing corpora. In the case of translational and translated Belarusian the obvious answer to the question of the project’s motivation is the scarcity of parallel corpora containing data in this language. Only few such corpora are available: Russian-Belarusian (Nacionalnyj korpus russkogo jazyka, 2017), which is a part of the Russian National Corpus and comprises almost 9.5 million words from literary and journalistic texts; Swedish-Belarusian (University of Gothenburg, 2016), which comprises over 400 thousand words of technical language; and partially Polish-Belarusian (BNkorpus, 2019) in the Biblical Corpus containing 16 Belarusian translations of various parts of the Holy Bible and one overall Polish translation, the language of which is archaised.

The Polish language is also somewhat under-resourced in this matter. European Language Resources Association (ELRA) catalogue contains 54 bilingual corpora for Polish, all of them paired with English. Out of these, source of 34 were bilingual websites, eg. Export Promotion Portal or the website of the Chancellery of the Prime Minister of Poland. 6 Polish corpora from ELRA catalogue are part of multilingual sources. The “Computational Linguistics in Poland” website lists 15 parallel corpora and translation memories (Ogrodniczuk, Maciej, 2020). Among these resources only some are paired with English, and, although they are considerably large (eg. Europarl –

European Parliament Proceedings Parallel Corpus 1996-2011 (Koehn, 2005) where Polish English subcorpus is over 28 million words), they contain texts written mainly in legal and technical language. The biggest source of Polish-English literary data is publicly available corpus Paralela (Pęzik, 2016), in which literary works comprise 3.1% of all data (over 5.3 million words). Out of that amount slightly over 30% consists of Polish literary works translated into the English language. As Paralela is available in the public domain it contains titles with no valid copyrights, which, in fact, means that these are mainly works from the 19<sup>th</sup> century and older.

One resource connecting all three languages, although not directly (in most cases) is InterCorp, developed at the Charles University in Prague. The core part of that multilingual parallel corpus is fiction, currently 390 million words in 41 languages, where each language's share differs significantly (from under 100 thousand to over 100 million). The governing principle in data selection for InterCorp is to have each text in Czech and at least one other language. In most cases each text occurs in more than three languages, and understandably some texts included in the database appear in English, Polish and Belarusian. The core part of InterCorp is aligned at the sentence level, it is proofread and complemented with rich metadata, however the directionality of translation is not marked in any way and the morphosyntactical annotation varies from language to language, although there is an ongoing effort to unify it according to Universal Dependencies (details of UD project in Chapter 2.3).

As discussed in the previous chapter, the Belarusian language is minoritised, therefore creating the EPB corpus not only enriches the resources of the language, but also strengthens its position in the international academic field. This concerns not only the Belarusian language, as Polish, even though a statutory national language, is, to some extent, under-resourced – the amount of available tools and resources is much closer to the Belarusian language than to English. Additionally, involving three languages of different status creates interesting opportunities for research, as the English language is considered as an indicator of the perceived importance of the translated literature in the translating culture, and that aspect can be investigated in later phases of the EPB corpus development (see Chapter 4).

As Even-Zohar asserts in his polysystem theory (1990), in the strong, developed cultures, such as English, the translated literature might be either a source of novelty or follow the existing norms and patterns. Extra-linguistic factors of text choice, such as the sociocultural and ideological contexts (Linn, 2006), are not to be forgotten. In the modern, capitalist world, only works of literature which are perceived to stand a chance of being profitable products are normally published, although a few small presses make a point of publishing work in translation from less widely known languages. This can be illustrated by Glagoslav Publications, an independent British-Dutch press – its primary focus is “to bring out translations that embody values that are uniquely Slavic in nature and celebrate universal values as reflected in diverse cultural demographics of Russia, Ukraine, Belarus and other nations in the region” (Glagoslav Publications Ltd., 2018). Another notable example is the publication series by the publishing house of The Jan Nowak-Jeziorański College of Eastern Europe entitled “Belarusian Library” (College of Eastern Europe, 2018).

The bigger the market the wider the range of potential readers, but also more candidates for translation. In the case of peripheral European literature, such as Belarusian, any literary text that has been translated into any lingua franca is regarded as successful.

The issue of the purpose of building a corpus is connected with the criterion of time span of the texts included in the database. As mentioned above, the number of parallel corpora including contemporary non-specialised uses of language is not sufficient. Literary prose provides or tries to provide this type of language, and prose, rather than plays or poems, was chosen for the corpus for a couple of reasons. First, prose is the most popular form translated from English into Polish and Belarusian. Only a small number of plays and poems are available in the three languages. Second, poetry and some plays operate a highly poetic language that cannot be translated but has to be actually rewritten, therefore it cannot be used to analyse how various structures and grammatical patterns are reflected in the languages in question. Therefore literary prose from the 20<sup>th</sup> and the 21<sup>st</sup> centuries was chosen for the EPB corpus.

b) Availability of texts (Crawford & Csomay, 2016, pp. 76–77; Kenning, 2010, pp. 489–490)

The first factor determining the availability of texts is their mode. Spoken texts are much more difficult to obtain and their preparation is extremely time-consuming. Nevertheless, Paralela does contain spoken text of a kind, namely subtitles. Reflecting the amount of English-language movies distributed in Poland, this dataset is large – over 63 million words (Pęzik, 2016). In the case of Belarusian, however, this type of data is not easy to obtain, for three reasons. The first, is the minorisation of Belarusian language and strong prevalence of Russian. The second is the position of English-language cinema, which is less significant in Belarus than in Poland, and therefore results in less English-language movies distributed in this country. The third reason is a strong tradition of dubbing and voice-over in Belarus – subtitles are simply very unpopular (Franco et al., 2013).

For these reasons written texts have been chosen as the core of the corpus. Additionally, literary works have been chosen, rather than non-fiction, because translations exist, enabling comparison, and because literary works tend to be rich in cultural items, and linguistically diverse. The process of obtaining and digitising (where needed) the texts is discussed in detail later in this chapter.

c) Size and representativeness (Kennedy, 1998, pp. 62–70; McEnery et al., 2005, pp. 13–19; Nelson, 2010, pp. 54–60; Weisser, 2016, p. 44)

Corpora are supposed to represent a language or a variety of it in a way that makes it possible to make generalisations. In the case of the EPB corpus the goal was to create a tri-lingual corpus, and that imposed the methodology of text selection – if a title was translated into the Polish language but not into the Belarusian, it was excluded from the database. And vice versa. Each title needs two counterparts.

Due to the scarcity of Belarusian-language material the number of titles available in the three languages is not enormous, however the time and workforce constraints of the project impose some restrictions in the data choice. Therefore in the initial phase of its existence the EPB corpus is designed to be uni-directional and comprise 20<sup>th</sup>/21<sup>st</sup> century English fiction translated into both the Polish and the Belarusian languages. In the case of multiple translations, only one version is chosen

– usually the first translation, otherwise the easiest to obtain. The number of works is slightly over 100 and the initial size of the corpus is almost 10 million words.

The issue of size and representativeness indicates some important criteria of text selection – verified translation and verified translator. In some parallel corpora designs (Hareide & Hofland, 2012) all texts which are not verified translations in accordance with Toury's definition of translation (Toury, 2012) – meaning, they are not presented or recognised as translations in the target culture – are removed from the collection. Similarly, all texts by translators not operating under their full name are excluded.

In the case of the EPB corpus a few items do not fall within these criteria. Firstly, three texts have been published on community websites and not in reliable sources, such as a publishing house or translation journal. Secondly, four Belarusian translators from the EPB database translate under pseudonyms. The identity of one of these translators is unknown, even though he or she published in various acknowledged sources and translated works by Virginia Woolf, William Golding and Aldous Huxley, among others. For the sake of keeping the EPB corpus truly representative these titles were included in the collection, nevertheless in the final outcome the corpus user should have the possibility of excluding these unverified translations and translators in the process of analysis.

Yet another criterion of text selection is the orthographic standard. As discussed in the previous chapter, the Belarusian language has two orthographic variants, *Taraškievica* and *narkamaŭka*. Only one of these variants is regarded as official and used in public affairs and, most importantly, in education (including teaching Belarusian abroad). However, precisely because of that difference in attitude towards the two standards the EPB corpus does contain translations in both *Taraškievica* and *narkamaŭka*. Being able to analyse translation trends combined with the orthographical variant of the text offers new insight in the area.

- d) Legal issues (Crawford & Csomay, 2016, p. 76; Kenning, 2010, p. 490; Weisser, 2016, pp. 45–46)

Copyright law in the United Kingdom is governed by the directives of the European Union. While it is difficult to predict how these laws will change after Brexit, one should not underestimate the current state of affairs. The full list of eleven directives and two regulations can be found on the website of European Commission (European Commission, 2018). Certain aspects of these laws are especially significant for a corpus compiler.

First, the duration of protection. Under European legislation, intellectual property is subject to copyright during the life of its author and 70 years after his or her death. This is a primary reason for including in corpora literary works from the 19<sup>th</sup> century and earlier years rather than the 20<sup>th</sup> and the 21<sup>st</sup>. In the case of national corpora external funds are usually available for purchasing the copyright or the licence for distributing a particular literary work, as well as an appropriate number of people and amount of time needed for contacting applicable institutions.

A second important aspect of EU copyright laws is database protection. Databases that are proven to be created by the means of substantial investment of human, technical and financial resources are protected under the Database Directive (The European Parliament and the Council of the European

Union, 1996). This directive, however, concerns the way in which the data is structured, not the content itself.

A third significant aspect of the EU regulations is *research exception* (Karapapa, 2017). This exception allows for using protected data for one's own research under the condition that these data are essential from the point of view of the research aim and that they are not used commercially. Additionally, the protected data can be available only to a limited number of users and an effective method of limiting the access should be implemented.

In the case of the EPB corpus every effort to share the resource with the public will be made. However due to copyrights the process is going to be twofold. First, an effort to gain the copyrights is being made so the corpus could be partially available to the public. Regardless of the results, the collection can be made accessible to students and researchers via the infrastructure of organisations such as CLARIN (mentioned in Chapter 1.4) or ELRA (European Language Resources Association). They offer login via educational institution of the user – this immediately limits the circle of people accessing the corpus to those who are conducting research in the member countries belonging to a particular organisation. Additionally, the corpus is going to be available through a specially designed interface allowing for querying the database and conducting simple statistical tests. Users will not be granted access to all data, only the elements which are necessary for their research, furthermore a method of technical protection against data misuse will be implemented.

The actions mentioned above should suffice for publishing the EPB corpus for researchers in compliance with EU legislation. CLARIN and ELRA specialist, Paweł Kamocki (CLARIN, 2018) will be consulted and will supervise the process.

- e) Sampling (Kennedy, 1998, pp. 74–75; McEnery et al., 2005, pp. 19–21; Nelson, 2010, pp. 57–59; Weisser, 2016, pp. 42–44)

Due to the assumption of gathering the whole population of the chosen type of data, the EPB corpus contains texts ranging from short stories, a few hundred words long, up to novels counting over 200 thousand words. The uneven size of the literary works included in the corpus may skew the results of some types of corpus analysis, in other words the prevalence of a feature in a lengthy text might obscure important features that occur in the shorter text. Therefore a sample of each text should be extracted and used to build a balanced sub-corpus, according to the assumption that “a sample is [...] representative if what we find for the sample also holds for the general population” (McEnery et al., 2005, p. 19), where “population” means the full text in this case. Details of the EPB sampling are discussed in Chapter 5.

After analysing the issues of corpus design, the database of titles was created (proportions of the data sources are discussed in Chapter 2.2). The preliminary assumption was that Belarusian translations of English literature are less numerous than Polish and therefore the search started from Belarusian sources available online. Three main open-source repositories of Belarusian-language literature are available online. First is Biełaruskaja palička (Biełaruskaja Palička, 2017), the Belarusian digital repository of literature, second is Kamunikat (Kamunikat.org, 2018), the Belarusian Internet library, and the third source is the website of *PrajdziSviet* ('PrajdziSviet -

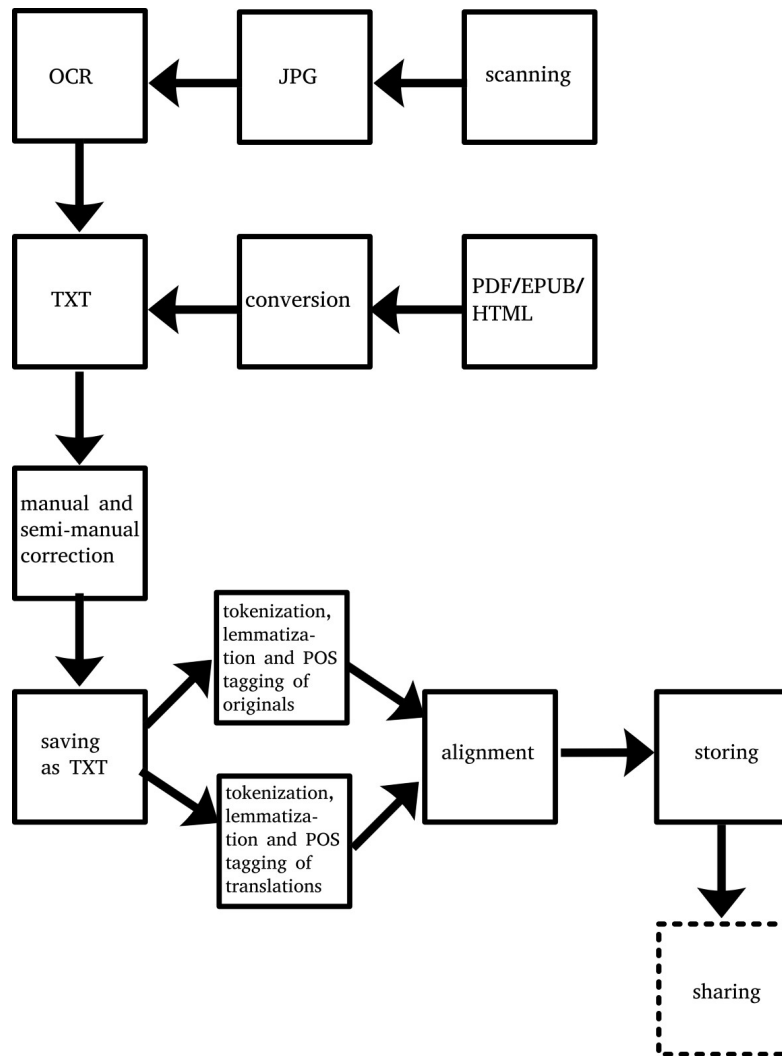
Časopis pierakładnoj litaratury [PrajdziSviet - Journal of translational literature]', n.d.), the Belarusian journal of translational literature. Searching through the catalogues enabled to create a database of English-language literary works translated into Belarusian. The database consisted of metadata – name of the author, title of the work, name of the translator, publication date, online version if available and the number of words. The database was further verified by confirming the existence of the corresponding Polish translations and supplemented with information obtained by means of manual search with the use of Google search engine. This in turn enriched the outcomes with useful sources, such as Čarniakievič's catalogue of the translations of international literature in the Belarusian journal *Krynica* (Čarniakievič, 2010).

The outcomes of searching through networks of library content, such as WorldCat (OCLC Online Computer Library Center, 2017) which claims to be the world's largest, were not satisfactory as they lack even the most obvious results, such the translations of Tolkien's *Lord of the Rings*. Preliminary searching in a Collective Digital Catalogue of Belarusian Libraries (Nacyjanalnaja biblijateka Biełarusi, 2016) proved that source to be insufficient and lacking translations published, e.g. in journals.

The final database contains 113 literary works by 11 female and 47 male authors (as of 2020; for details see Primary sources A in References).

## **2.2. Text collection**

The process of corpus building after it has been designed can be represented as follows:



*Illustration 2.2. Workflow of building the corpus.*

These elements are now discussed in detail, starting with data entry and text pre-processing. Depending on the source of the text, different methods of acquiring the data were implemented; either conversion of data in electronic format or adaptation by optical scanning (Nelson, 2010, p. 62). For obvious reasons, whenever it was possible the tasks were automated to save time and effort, but at the same time the high quality of the data was considered and various techniques have been implemented to enhance the output of text collection.

Some of the texts<sup>6</sup> were obtained already in a digital version: HTML (40%) and PDF or EPUB (25%). Some of the PDFs and all texts in HTML were freely available in sources gathering materials with no copyrights:

- online libraries: Project Gutenberg (*Project Gutenberg*, 2018); Project Gutenberg Australia (*Project Gutenberg Australia*, 2018); Bibliowiki (*Bibliowiki*, 2018); Wikisource (*Wikisource, the Free Library*, 2018); Google books (Google, 2018); Fadedpage (*Faded Page*, 2018); JSTOR (ITHAKA, 2018) (for English);

<sup>6</sup> A single text is understood in this context as one title (of a short story, novella or novel).

Wolne Lektury (Modern Poland Foundation, 2018) (for Polish);

The Internet Archive (The Internet Archive, 2018) (for English and Polish);

Kamunikat (Kamunikat.org, 2018); Biełaruskaja palička (Biełaruskaja Palička, 2017); CoolLib (*CoolLib*, 2018) (for Belarusian)

- repositories of the University of Adelaide (The University of Adelaide, 2018) and Federal University of Santa Catarina (Federal University of Santa Catarina, 2018) (for English)
- website of the Belarusian PEN Centre (Belarusian PEN-Centre, 2018) (for Belarusian)
- community websites (non-professional sources): *dziетки.org* (Našyja dziетки, 2010), *belpotter.by* (belpotter.by, 2018), *translatedby.com* (*Translated by Humans*, n.d.) (for Belarusian)
- online educational platform Lingualeo (Aynur Abdulnasurov, 2018) (for English)
- websites of journals: *New Yorker* (The New Yorker, 2018), *The London Magazine* (*The London Magazine*, 2018) (for English); *PrajdziŚviet* (*PrajdziŚviet—Časopis pierakładnoj litaratury [PrajdiSviet—Journal of translational literature]*, n.d.), *Dziejasłoŭ* (*Dziejasłoŭ*, 2018) (for Belarusian)
- catalogue Cyfroteka (Cyfroteka, 2013) (for Polish)
- music platform Genius (Genius Media Group Inc., 2018) (for English)

Texts in HTML format were gathered by a specially designed Python script which stripped them of HTML coding and put them in separate TXT files. With PDF and EPUB files, a web-based freeware converter was used ('Convert PDF to Text Online', 2018; 'Free EPUB to TXT Converter', n.d.) and the text files were then edited with the use of Linux command line and regular expressions, e.g. *sed* command for deleting page numbers and headers or *csplit* command for dividing big files into a series of smaller ones (usually novels divided into chapters).

All remaining texts (35%) had to be adapted by optical scanning. For financial reasons only freeware OCR software was used. Tests proved that LIOS (*Linux-Intelligent-Ocr-Solution*, 2017) after being trained to process Belarusian language, deals very well with all three languages (English, Polish, Belarusian). LIOS uses the Tesseract engine, development sponsored by Google, currently supporting over 110 languages (Algun, 2018). The downside of the Tesseract are high requirements as to the quality of the picture – it cannot handle skewed images or images with noisy background and low resolution, therefore pre-preparation of scans was required, which was performed in GIMP (The GIMP Development Team, 2017). Nevertheless this procedure was not needed too often as this project deals with contemporary data and in most cases the process of scanning was closely controlled by humans to provide acceptable quality. Texts processed with OCR software were proofread in text editors operating dictionaries – Microsoft Word (Microsoft, 2016) and Libre Writer (The Document Foundation, Debian and Ubuntu, 2017), since both of these programmes support all three languages (Belarusian dictionaries are available at the website of the BNkorpus). Using the text editing programmes it was possible to indicate discrepancies between the text and the dictionaries and consequently to speed up the process of digitising.

All texts were finally saved in TXT files, using 8-bit Unicode Transformation Format (UTF-8). UTF-8 is nowadays the most common character encoding in the World Wide Web (W3Techs, 2018) – it is used by nearly 93% of all the known websites. Its main advantage in the context of corpus building is its multi-lingual operability. Put in simple words, it supports all alphabets and special characters of all languages.

### 2.3. Text encoding

Encoding is as an umbrella term for the process of adding various types of data and it is a very important, if not crucial, element of corpus creation. The importance of text encoding has been discussed by a number of researchers (Kuebler & Zinsmeister, 2015; McEnery et al., 2005; Weisser, 2016).

At the highest level every text needs character encoding. As pointed out in Section 2.2, in the EPB corpus UTF-8 was chosen for encoding characters, mainly because of its popularity and on the grounds that it supports both Latin and Cyrillic alphabets, and Polish special characters. Character encoding is something that any text needs but in corpus linguistics text encoding is usually performed on more levels.

Firstly, the metadata, or the information about the text itself, is provided in the corpus. McEnery (McEnery et al., 2005, pp. 22–23) identifies two reasons why this type of mark-up is important. First is the restoration of the original context; second – the added value of extra-textual information which enables a broader range of research questions to be addressed. The most common scheme for corpus mark-up is provided by the Text Encoding Initiative (TEI).

The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation (TEI Consortium, 2020)

Finally, the linguistic data, on multiple levels, is encoded within a corpus. Analysing the raw text what is text encoding? Add a definition or description of what the term means in corpus linguistics allows for reaching some conclusions, however enhancing it with linguistic encoding is highly recommended, as it enables much more detailed analysis. Even though linguistic annotation is resource-consuming (especially in terms of time and manpower), one cannot forget its undeniable advantages: ease of information extraction, re-usability, multi-functionality and explicit record of linguistic analysis (McEnery et al., 2005, p. 30).

On the other hand, the disadvantages of annotation must be taken into account: annotation clutters the corpus, it makes it less accessible and expandable, it enforces one particular linguistic analysis and it cannot be entirely consistent (McEnery et al., 2005, pp. 31–32). Yet these disadvantages can be overcome by using annotation formats which enable separating the original text and the linguistic information, and ensure that the linguistic analysis imposed by the annotation does not prevent researchers from finding the phenomena they are interested in (Kuebler & Zinsmeister,

2015, p. 33). Consistency, in turn, much like anything else in corpus building, can never reach 100%, but some methods provide a high quality of annotation. These methods are best summarised in Leech's Maxims of Annotation (Leech, 1997, pp. 6–7):

- (1) It should be possible to remove the annotation from an annotated corpus in order to revert to the raw corpus.
- (2) It should be possible to extract the annotations by themselves from the text.
- (3) The annotation scheme should be based on guidelines which are available to the end user.
- (4) It should be made clear how and by whom the annotation was carried out.
- (5) The end user should be made aware that the corpus annotation is not infallible, but simply a potentially useful tool.
- (6) Annotation schemes should be based as far as possible on widely agreed and theory-neutral principles.
- (7) No annotation scheme has the a priori right to be considered as a standard. Standards emerge through practical consensus.

Having this in mind let us discuss the encoding used in the EPB corpus.

### 2.3.1. Metadata

Elements to be included in the metadata:

- Name of the author – refers to the name under which a text was published, even if it is a pseudonym
- Real name of the author – refers to a real name (if known) in the case an author published under a pseudonym; if unknown the field is left empty
- Title of the text – full title of a published piece of literature, no short versions
- Name of the translator (for Polish and Belarusian versions) – as with the 'name of the author', the published name is taken into account
- Real name of the translator – analogously to the 'real name of the author' refers to the cases in which a translator published under a pseudonym
- Translated title of the text – full title in Polish/Belarusian
- Author's/translator's gender – various studies indicate that gender is an important factor not only when writing literature but also when translating, as men and women use language in different ways
- Author's/translator's date of birth – recent studies suggest interesting relationships between the texts written by authors born in certain periods, as well as between translations made by translators born in different periods; this dataset is still being developed as the dates have not been found for all translators yet

- Year of the publication of the original/translation – relationship between the dates of publication of certain titles and dates of their translation might indicate specific sociopolitical issues
- Size of the original text/translation (in words) - significant differences (recurring among all types of texts) between the size of the text and its translation indicate systematic differences between languages' grammatical patterns, possibly also (self-)censorship. The differences in sizes of translations in two different languages belonging to the same family (such as Polish and Belarusian) might indicate differences at yet another level (since the grammar is very much alike)
- Written standard - this concerns English texts (differentiation between British, American and other written language standards) and Belarusian translations; as indicated in Chapter 1 Belarusian language has two orthographical standards the and although they are very close, using one or the another might be perceived as a political statement.

Metadata were initially stored as a spreadsheet and later transferred into a database in MySQL ('My' stands for the name of the co-funder's daughter and 'SQL' for Structured Query Language). MySQL is an open-source management system for operating databases in SQL programming language. This format for storing metadata is one of the most common in corpus linguistics for two reasons. First, it allows for querying the metadata in a very flexible way and gaining a deeper insight in the relations between particular elements. Second, owing to its popularity it guarantees high interoperability – it can be easily modified by users of various operating systems.

### 2.3.2. Tokenisation

Tokenisation is a process of converting the whole text into separate tokens. While the definition is simple, researchers frequently discuss what should be regarded as a *token* and various corpus designs understand a *token* in various ways. As pointed out in Section 1.4, Weisser (2006) indicates the problem of varied spelling of words (e.g. icecream/ice-cream/ice cream). If the token is defined as a string of characters between spaces and/or punctuation marks, 'ice cream' would be two tokens, even though they constitute a single semantic unit, and 'he's' might become one token, even though these are two semantic units. This is just one example but English, as well as other languages, poses many more challenges of this sort (Manning et al., 2008). Nevertheless, tokenisation is the easiest of several stages of linguistic annotation and tokenizers usually perform really well even for under-resourced languages. The table 2.2 presents results for the UDPipe v2.5 tokenizer (details of this programme are to be found in section 2.3.3) available via the natural language processing toolkit 'udpipe' (Wijffels, 2020):

metrics	Belarusian	Polish	English
precision	92.86%	100%	100%
recall	96.30%	100%	100%
F1 (weighted average of precision and recall)	94.55%	100%	100%

Table 2.2: UDPipe v2.5 tokenizer accuracy.

What is important to remember is that tokenisation issues are language specific and no solutions are universal. Usually the best results are obtained when using taggers developed specifically for each language. Tokenisation is always a first stage of lemmatisation, therefore the taggers for English, Polish and Belarusian are discussed in the following section.

The ultimate number of tokens obtained for the EPB corpus are as follows:

English	3 913 960
Polish	3 383 614
Belarusian	3 372 408
<b>total</b>	<b>10 669 982</b>

*Table 2.3: Number of tokens in the EPB corpus.*

### 2.3.3. Lemmatisation

Lemmatisation is the process of assigning each token to its basic linguistic form, lemma (e.g. ‘to be’ for ‘was’ or ‘tall’ for ‘taller’) and subsequently grouping words as multiple variants of one lemma and with regard to its meaning (e.g. ‘saw’ as a past form of ‘to see’ or as a noun). Lemmatisation should not be confused with stemming which is the process of reducing a word to its stem (base or root form), e.g., deleting the morphemes used for inflecting the word. The stem may not be identical with the morphological root (e.g. ‘studies’ is reduced to ‘studi’).

As mentioned at the beginning of this section, lemmatisation is an essential part of corpus building as it allows for more accurate language analysis. A simple example to explain this importance is frequency list analysis. If the verb ‘to be’ is represented in the corpus by all its varieties – be, am, is, are, was, were, been, being – and each of these variants is represented in the result as a separate item, this could easily obscure the actual results as each of these forms may be outnumbered by another word. But if one is to add up all the occurrences of just one lemma – to be – then this particular lemma might be the most common. This is why lemmatisation cannot be neglected even at the very basic level of corpus analysis.

Available tools for lemmatisation differ from language to language, in both quantity and quality. Below several freely available lemmatisers are described, and if their accuracy is known it is noted. Let us start with lemmatisers available for the English language.

One of the most versatile tools for English language in general is Python NLTK (Natural Language Toolkit) (NLTK Project, 2018a). It is a suite of libraries and programs for natural language processing, initially designed for English only but later enriched with data for other languages – currently NLTK provides interfaces for over 50 corpora and lexical resources (NLTK Project, 2018b). NLTK provides a lemmatiser based on the WordNet, which is a lexical database showing relations (such as synonymy or hyponymy) between words and thus designed specifically for natural language processing. While this function is available only for English one should note that it

is possible to develop a stemmer (programme for stemming) for non-English languages, including Belarusian and Polish (Hardeniya et al., 2016, pp. 500–501).

Another lemmatiser for English is CST Lemmatiser (Center for Sprogteknologi, 2018). It is based on set of rules and an optional dictionary expressing relations between the base form and particular word forms. CST Lemmatiser works for English and a dozen other languages, including Polish, but not Belarusian.

For Polish, one of the most comprehensive tools for lemmatisation is the morphological analyser Morfeusz (Woliński & Lenart, 2018). Its newest version is based on the Polimorf dictionary (Woliński et al., 2012) which, in turn, is a compilation of several other dictionaries. A substantial part consists of data from the SGJP dictionary which accumulated lexemes from a dozen paper-edition dictionaries published from the 1980's, as well as from Polish language corpora and from the Internet. Currently PoliMorf contains over 414,000 lexemes.

All the above-mentioned tools reach at least 90% accuracy in the process of lemmatisation. In the case of the Belarusian language however that level of performance is not achieved. Regarding the Belarusian language, only three tools for tokenisation and lemmatisation are known to be available for public use.

First is UDPipe (Straka, 2018)(Straka, 2018a), a programme developed at the Charles University in Prague, Czech Republic, as a CLARIN (mentioned in Chapter 1.4) project, available on-line. UDPipe is the only tool that allows for analysing all three languages involved in this project, though with different outputs. This is due to the difference in the input data. The input data is provided by Universal Dependencies, “a framework for cross-linguistically consistent grammatical annotation” that is created voluntarily by a community of 200 contributors from the field of computational linguistics (UD; Universal Dependencies contributors, 2019). The amount and structure of the data (in UDPipe version 2.7) necessary for the EPB corpus analysis is as follows:

- English: 9 treebanks (that is, syntactically annotated corpus), 648 thousand tokens, 14 types of language register (including fiction),
- Polish: 3 treebanks, 499 thousand tokens, 5 types of language register (fiction, news, non-fiction, social and spoken language),
- Belarusian: 1 treebank, 275 thousand words, 6 types of language register (fiction, legal, news, non-fiction, poetry, social).

It must be noted that the UD project is very dynamic, as the updates are being made every 6 months (although not every language is updated with each new release). Among those three languages the amount of data for Belarusian changed the most dramatically. It joined the project in 2017, it was merely 8 thousand in the version 2.3 (November 2018), then 13 thousand in version 2.4 (May 2019) until finally in the most recent version 2.7 it reached 275 thousand words (November 2020).

UDPipe v2.5 tested via NLTK ‘udpipe’ exhibits the following F1 (the weighted average of precision and recall) figures: 65.45% for Belarusian, 62.96% for Polish and 62.96% for English. However, a manual test conducted on a 1000-word sample from the EPB corpus demonstrates 86.93%

lemmatisation accuracy in version 2.4, the first one containing fiction (for details see Appendix 1). The UD algorithm does not use the tag ‘unknown word’, instead it guesses the lemma. It is an educated guess, therefore it might improve the performance of the programme.

Another lemmatiser for Belarusian is Lemmatizer (Speech synthesis and recognition laboratory, 2018) developed in the Speech Synthesis and Recognition Laboratory at the National Academy of Sciences in Minsk, Belarus, available on-line. This tool is based on a dictionary published in 1987 (Biryła, 1987). Tests on a 1000-word sample proved the Lemmatizer to be 86.91% accurate; 8.69% of words were unknown (these were mainly words written in Taraškievica and foreign proper names; details are to be found in Appendix 1) and in some cases two or more possible lemmas were given – in this test only the first one was taken into account.

A third available lemmatiser for Belarusian is available as a part of the NooJ application (Silberztein, 2018). The application has a set of built in dictionaries that can be developed with the use of the programme. The Belarusian dictionary is primarily based on the 1987 Biryła dictionary, however “new words from modern texts on literature, science and technology were added to the basic electronic dictionaries” (Lobanov et al., 2015). The authors do not provide any indication as to where these words come from. Tests on a 1000-word sample proved the NooJ lemmatiser to be 73.80% accurate with 6.12% unknown units (details in Appendix 1). The format of the output is not easy to operate on later, because the basic form is optional, that is in many cases the lemma is absent because it is identical to the word form. As a result, the number of fields in the output is uneven – sometimes there are two fields (word form + morphological features), sometimes there are three (word form + lemma + morphological features). To some extent, this hinders automatic processing of the annotated text.

#### 2.3.4. POS tagging

Part-of-speech (POS) tagging is the process of assigning to each token the information about what part of speech it is. As with previous procedures described in this chapter, regardless of the simple definition, the process itself is not easy. One of the most significant problems is defining the ‘part of speech’ and compiling the list of the available types of these parts, or the tagset. The shortest and most general list of POS tags that could be applied to most languages of the world is provided by the UD project (Universal Dependencies contributors, 2018c). This tagset includes six open class types (adjective, adverb, interjection, noun, proper noun, verb), eight closed class types (adposition, auxiliary, coordinating and subordinating conjunctions, determiner, numeral, particle, pronoun) and three other types (punctuation, symbol and X for words that cannot be assigned any other tag). The UD universal POS tagset supports over 70 languages, Indoeuropean, Afro-Asiatic, Sino-Tibetan, Turkic, Uralic and even Basque and Swedish sign language. Nevertheless more detailed tagsets exist. They distinguish between various types of a particular part of speech, such as personal and possessive pronoun in the Penn Treebank (A. Taylor et al., 2003) or include language-specific classes, such as ‘siebie’, a reflexive pronoun in the National Corpus of Polish (Przepiórkowski, 2011).

Regardless of the typology applied in the process of tagging, the accuracy of the programmes differs substantially between languages. English still occupies the place of the pioneer of language technology advances and, as we may expect, tools for its analysis perform the best. In fact, some researchers question the possibility of further developments for English, as POS taggers for this language achieve per-token accuracy at the level of human annotators – over 97% (Manning, 2011). It is noted, however, that this measure might not be entirely reliable because non-ambiguous tokens easily increase the score. If one is to test the taggers' accuracy in terms of the whole sentence, then the outcome is only 55-57% (Manning, 2011). Another significant problem is the choice of training data – well performing taggers tested with input of different topic, writing style or time of publication produce much poorer results.

One study of nine POS taggers (Horsmann et al., 2015) reveals that all of them reach an average accuracy of between 84.1 and 89.5%. The most accurate of them, Clear NLP tagger, does not have a user interface, it is available only via command line. Next comes the Arktools tagger that is primarily designed to process tweets, and third, the Stanford tagger which is freely available (Toutanova et al., 2003). Even though it was developed over 15 years ago, it is still supported and new versions are being released. Stanford tagger uses Penn Treebank and offers various extensions, such as a model for Twitter tagging and tagging languages other than English, including Polish. No model for the Belarusian language is available, the programme can however be trained to deal with other languages. It is suggested that the accuracy of the programme can be improved by improving the taxonomic basis of the linguistic resources from which this tagger is trained, that is by providing clearer classification into grammatical categories and minimising inconsistencies (Manning, 2011).

Polish language tagging has proven to be more difficult and currently the average level of accuracy is around 91% (Kobyliński & Kieraś, 2018). The reason for that is a combination of the rich morphology of the language (however taggers for similar languages, e.g. Czech and Slovene, prove to perform much better) and the quality of training data. English taggers are tested on the Penn Treebank which consists of Wall Street Journal articles, while Polish training data comes from the National Corpus of Polish which is much more differentiated in terms of language register.

The Institute of Computer Science at the Polish Academy of Science lists nine publicly available POS taggers for Polish (Wawer, 2018), most of which were evaluated in a recent study by Kobyliński and Kieraś (Kobyliński & Kieraś, 2018). All tested taggers perform in the range from 87 up to 91%, and three programmes with the best results are Concraft-pl (Waszczuk, 2018), Wrocław CRF Tagger (Walkowiak & Kaliński, 2013) and Wrocław Memory-Based Tagger (Radziszewski & Śniatowski, 2013). Even though none of these programmes come with an interface, the WCRF algorithm is used in the CLARIN Morpho-syntactic tagger (CLARIN PL, 2018).

In terms of the Belarusian language, the three lemmatisation tools discussed above are also equipped with POS tagging algorithms. Table 2.4 showcases the F1 scores for tagging from plain text in UDPipe v2.5:

(Speech Synthesis and Recognition Laboratory, 2018a)

type	Belarusian	Polish	English
upostag (universal POS tags)	18.18%	22.22%	29.63%
feats (features; additional lexical and grammatical properties of words, not covered by the POS tags)	0.00%	11.11%	3.70%

*Table 2.4: Tagging with UDPipe v2.5 accuracy.*

A test on the 1000-word sample from the EPB corpus reveals that UDPipe v2.4 achieves 81.80% accuracy (details in Appendix 1).

In the POS tagger provided by SSR lab (Speech Synthesis and Recognition Laboratory, 2018a) the analysis is based on a wider range of data, including not only the 1987 Biryła dictionary but also one newer publication and an internet-based list of frequent words compiled in 2016. SSR lab tagger uses a bigger tagset based on the Penn Treebank but enriched with language specific tags. The meaning of the tags can be found in the Voiced Electronic Grammatical Dictionary (Speech Synthesis and Recognition Laboratory, 2018c). In a test on a 1000-word sample from the EPB corpus, the SSR tagger achieved 87.62% accuracy with 5.91% unknown words (for details see Appendix 1). The last of the three taggers provided by NooJ application achieved 81.63% accuracy and did not recognise 6.12% of units in the test on a 1000-word sample (for details see Appendix 1).

Multiple aspects of usage and performance of the above-mentioned programmes to varying degrees influenced the ultimate choice of the software used later on in preparing the EPB corpus. Despite the accuracy test results it was UDPipe that was implemented in this case and the comprehensive discussion of the reasons behind that selection can be found in Chapter 5.

## 2.4. Alignment

Alignment is a process of assigning to a unit from the source text a corresponding unit in the target text. This definition is not as obvious as it might seem; the procedure itself is complicated and poses several questions (these are discussed later in this chapter), nevertheless it is hard to deny the value of alignment.

- a) What is the basic unit of alignment?

When using a parallel corpus we usually look for the equivalents of words and many people expect that texts are aligned at this level. This is however really difficult to do with the use of automatic methods. Semi-automatic and manual ways of aligning the texts at the level of words are much more accurate but very time-consuming. Apart from that choosing an equivalent of a word is often problematic as the word might be omitted in the translation or, on the contrary, translated with the use of two or even more words. For this reason it is sentence alignment that is chosen most frequently for parallel corpora, including the EPB corpus described in this thesis.

- b) What are the types of relation between the source text units and the translation units?

Three obvious types are ‘one to one’, ‘one to many’ (alternatively ‘many to one’) and ‘one to none’ (‘none to one’). These types of relations are implemented in the vast majority of aligners (whether automatic or manual) and they refer to the number of units being aligned, for example, in the case

when the alignment unit is a sentence the relation ‘one to many’ mean that one original sentence correspond to many (two or more) sentences in the target text. However these types of relation are not sufficient in the case of literary texts. The common problem with literature is that translators very often change the order of sentences, paraphrase or even omit short fragments. Such cases can completely disrupt automatic alignment of the text and make manual alignment more difficult or impossible if the program does not allow for other relations.

c) What are the ways of aligning the texts and which one is the most efficient?

Three basic types of the algorithms for sentence alignment are distinguished – length-based, lexicon-based and the combination of the two. A classic example of a length-based algorithm is the Gale-Church method (Gale & Church, 1993). It is based solely on the idea that all matching segments are similar in length and therefore the length, or number of characters of the sentences, are the only input data. Two main advantages of this algorithm are language independence and simplicity with concurrent high accuracy (over 90%). Other length-based algorithms, such as Brown (Brown et al., 1991), are based on the word count, this however proved to be less accurate than algorithms using number of characters (Singh & Husain, 2005).

The Champollion method (Xiaoyi, 2006) is a useful example of a lexicon-based algorithm. This method is based on the tf-idf (term frequency – inverse document frequency) value which reflects the importance of a word (or a lemma if the text is lemmatised) in a document that is a part of a bigger collection, such as corpus. More occurrences of a word in a single text mean high tf-idf value, however more occurrences of a word in many texts of the collection mean low tf-idf value. Put another way, the words that occur frequently in a single or just a few texts are the most important ones. By calculating the weights the algorithm aligns sentences. The Champollion algorithm is more precise than Gale-Church (close to 100%) it is however more time-consuming. Among lexicon-based algorithms, the cognate matching method should be mentioned. In this algorithm cognates, words of common origin, are searched for and used for finding matching sentences. This method, however, might be efficient mainly in closely related languages and even in this case false friends would cause worse results.

Hybrid methods of sentence alignment, such as Fast-Champollion (Li et al., 2010), take the advantages of the length-based and lexicon-based methods but have fewer disadvantages. Fast-Champollion has slightly lower precision than Champollion, it is however much faster because it analyses the text in smaller chunks that are first selected by means of the length-based method. Both Champollion and Fast-Champollion algorithms are wholly or partially lexicon-based and therefore are dependent on lexical data. It follows then that they are most efficient for languages with extensive resources, but can under-perform for low-resource languages.

Recently a fourth type of algorithm has emerged, and while it is still being developed and is not a well-established method, linguists have high hopes for it. This algorithm is based on semantic embeddings – “vectors whose relative similarities correlate with semantic similarity” (Gavagai AB, 2015). They have been used in language modelling, however traditionally features used to create word embeddings were introduced manually. Using artificial neural networks (ANN) allowed for doing so automatically.

ANN is “a bio-inspired mechanism of data processing, that enables computers to learn technically similar to a brain and even generalise once solutions to enough problem instances are taught” (Kriesel, 2007). ANN converts word embeddings to a numerical form, namely a vector in space, which makes this method language-independent. These vectors are dependent on text features such as length, score based on punctuation, and cognate score values (Fattah et al., 2006). Additionally, with this method it is possible to analyse vast amounts of data available on the Internet, no manually annotated data is needed, contrary to traditional statistical language modelling. ANN create huge possibilities for computational linguistics, however linguists need to wait for ANN theory and terminology to become better established and more consistent.

Having discussed three main issues of the text alignment – what to align, how and in what manner – let us now consider the tools. Programs for aligning pairs of texts can be roughly divided into stand-alone programmes and modules of other programmes. Among aligners which are a part of another programme one can name:

- translation alignment tool in SDL Trados Studio (SDL Trados, 2018)
- Microsoft Excel/Libre Calc – this spreadsheet is an accepted format of data for creating a parallel corpus in the SketchEngine

Standalone programmes for aligning texts include:

- CORAL (Corpus Aligner) (TakeLab, 2014)
- Translation Corpus Aligner (Hofland & Johansson, 1998)
- Montreal Forced Aligner (Montreal Corpus Tools Revision, 2018)
- NOVA Text Aligner (Supernova-soft, 2014)
- LF Aligner (Farkas, 2018)

These are just few examples of publicly available text aligners. The main problems with these programmes are lack of possibility of simultaneous processing of three languages (except for the LF Aligner) and insufficient choice of types of relations in the case of manual alignment.

A rare example of a trilingual aligner developed for the purposes of aligning literary texts is TRACE-Aligner (Zubillaga et al., 2015) which allows for tagging, automatic alignment and then manual refinement of the alignment. The outcome of the work in TRACE-Aligner is a MySQL database which can later be used to carry out complex searches with the texts which can be additionally filtered based on metadata. The authors claim that they “added different editing options to the program, such as “combine”, “add cell”, “edit” or “split”, to make the tool as versatile and as easy to use as possible” (Zubillaga et al., 2015, p. 83). However these options are not sufficient for the purpose of literary translations alignment.

Creators of Paralela Corpus (CLARIN-PL project, 2014) encountered this problem when aligning the texts. An automatic aligner was accurate enough for all texts but literary. This is why the Mantel application (Pęzik, 2016) was developed. It is a purely manual aligner, without automatic processing module, however it is enriched with interesting types of relations. Apart from one-to-one

(*simple* in Mantel) and one-to-many (*Merge/Split*) relations, it is possible to mark a segment as inserted (with no equivalent in the original) or deleted (with no equivalent in the translation). Four further options are:

*Crosslink* – used to mark equivalent sentences separated by one or more intervening segments

*Composite* – used to mark many-to-many segment blocks with overlapping sentence to sentence equivalence relations

*Compression* – used to mark complex mergers where several sentences are translated into significantly fewer sentences

*Paraphrase* – a last resort marker used to mark significant adaptations or paraphrases in the translation which could not be reasonably mapped at the level of individual sentences. (Pezik, 2016, p. 70)

Such an array of possibilities is highly recommended for aligning literary texts. The application is online-based which makes it independent from the operating system and enables it to be used on any computer, the interface is simple and user-friendly and the relations are saved as a relational database (RDB) which can later be exported and queried in Structured Query Language (SQL).

The main advantage of using automatic alignment is the time necessary for the process – it is many times shorter than in the case of the manual alignment. The drawback of such a solution is increased noisiness of the data, especially in the corpora of low-resource languages, because these, as stated before, provide less input necessary in most automatic text alignment methods. As a result, there are more zero-to-one/one-to-zero relations, or, in other words, omissions, than expected. Some sources report between 5 and 10% level of such a noise (Xiaoyi, 2006), and many researchers believe it is a fair trade-of for the speed of preparing the data.

It is however doubtful that much time could be saved with a manual check of automatically aligned texts, as in most researchers' experience automatic alignment is not sufficiently accurate when it comes to the literary texts. In other words, the alignment is distorted to a degree that prevents the researcher from drawing meaningful conclusions from the data analysis. That was the main rationale for creating Mantel, which offers a means of dealing with significant structure interferences, moreover, it is possible to do so in three languages simultaneously which saves time.

For these reasons it is Mantel that has been chosen to align the EPB corpus. The outcome of the process is the RDB containing information about types of relations in each text and each language. After a year of part-time work by one human annotator (namely the present author) a 1-million-word subcorpus of the EPB was aligned. It was the maximum amount of data manageable for one person to align while this PhD project was carried out. The rest is gradually being supplemented and it is expected that by the end of 2023 the full parallel corpus will be available online (more details on this topic are to be found in Chapter 5). The summary of the alignment types used in the process is presented in the table below:

Type of alignment	Count in 1-million-word corpus
simple	37 508

merge	3 575
split	7 138
crosslink	46
composite	1 184
compression	0
paraphrase	189

*Table 2.5. Count of various types of alignment in 1-million-word EPB subcorpus.*

These numbers apply to segments in English. As expected, the *simple* type, that is one-to-one alignment is the most popular, there are however twice as many segments with the original text that were split rather than merged in order to align the translation. This number suggests that Polish and Belarusian translators generally use shorter sentences than the English authors.

The information about alignment types grouped with the metadata can be used for further analysis of theoretical aspects of English-Polish and English-Belarusian translations, such as the level of interference in the translation depending on the year of publication or the gender of the translator, both in macro- and micro-scale.

## **2.5. Storing and sharing**

Storing and sharing the data, even though frequently disregarded in the process of building a corpus, is vital for the success of the project. After having devoted much effort and time to compile a corpus and annotate it so it would be ready to analyse, many researchers do not take proper action in order to save their work. The benefits of sharing the data, especially in the open mode, are numerous, varied and widely recognised. Within academia, it allows for verification and evaluation of the claims made by researchers, it helps to avoid duplication of a study, and it gives the possibility of combining the data from different sources. Whenever possible, open data sharing is also beneficial for businesses, which can re-use the data to create or improve products and services, and for governments to implement or improve their evidence-based policies. Facilitating some or all of these activities becomes easier after completing the following stages.

The first step is choosing the right format for the data, because “saving files in a format that is not compatible with the tools that will be used for analysis will result in many extra hours of work” (Reppen et al., 2010, pp. 33–34). Most tools for corpus analysis operate on TXT files (or plain text). Even if the corpus is linguistically annotated, the annotations themselves are stored in separate files, or the files with raw (unannotated) texts are stored alongside the annotated ones. That procedure fulfils Leech’s First Maxim of Annotation. Various tools for annotation use various ways of placing the annotations within the text.

One fairly common way is placing the annotation after an underscore. The result is the plain text file with no significant changes, except the tag after each word. This is used by CLAWS Tagger, e.g. ‘hospitality\_NN’. NN is a tag for singular common noun. This annotation is easily removable with a simple text editing command that deletes every character after the underscore. Similarly, the annotation itself can be extracted and this is in accordance with the second Maxim of Annotation.

However, it becomes complicated when more linguistic information, such as word accent, is added. This is a case in the Belarusian POS Tagger, e.g. `дзялі+ў_VIIPM`. Removing the POS tags will leave the text with + symbols after every accentuated syllable.

Another way for encoding the annotation is by organizing the data in the form of a table and consequently separating the word and the annotation with a tab. This solution is used in the UDPipe tool and the result of annotation is saved as CONLLU file (Universal Dependencies contributors, 2018a), which is also a plain text. Unlike in the first solution, the output of UDPipe annotation is structured and divided into three types of lines:

- word lines containing the annotation of a word/token in ten fields separated by single tab characters; in these lines each word is assigned an ID, base form, POS tag and other information
- blank lines marking sentence boundaries
- comment lines starting with hash (#); these comments contain information about the beginning of the document, beginning of sentence, ID of the sentence and the sentence itself in the raw form.

Such a structure of annotation allows for much more flexibility of data processing. Firstly, it is easy to extract the raw text by simply extracting all lines following the comment `#text=`. Secondly, it is easy to extract only the annotations and even create and edit a dictionary from the annotated words. This fulfils the second Maxim of Annotation.

Yet another standard of encoding the annotation is provided by TEI (TEI Consortium, 2020)(Text Encoding Initiative, 2018b). What that consortium proposes is using XML (Extensible Markup Language) for representation of texts in digital form. With the use of tags (all tags are contained within angle brackets `<>`) it is possible to encode not only the linguistic annotation on any level but also metadata of any kind. The TEI website provides extensive guidelines for using this type of encoding (Text Encoding Initiative, 2018)(Text Encoding Initiative, 2018a). As with previous examples, this format enables easy retrieval of the raw text and the extraction of the annotations themselves. It is also fairly easy to convert XML into other formats required in various programmes.

After having collected all the data and selected the right format for it, the second step before sharing the fruits of one's work is creating an extensive documentation. Such a documentation is essential to anyone who would like to use the corpus, or any dataset within it, and the possibility of using the data outside of the field it has been used in should be taken into account. Information about all the decisions taken and the reasons behind those decisions must be known to the potential user. Details reflecting the steps taken to create the corpus have to be explicit: design – the purpose behind the dataset, its size, representativeness, legal status and sampling method; secondly the process of text collection – particularly the exact sources of the texts and their character encoding; apart from that the type of linguistic encoding – metadata, tokenisation, lemmatisation, POS tagging and any higher level encoding; finally, in the case of parallel or multilingual data, type of alignment. The more details provided, the more reusable the data becomes. Both the procedures performed and

the tools used cannot be omitted – sometimes even the differences in operating systems cause seemingly small problems, which nevertheless result in waste of time and effort.

The third step of data storing is finding the right repository. Obviously, one's own computer or cloud storage is the first choice and actually it might be much more convenient for work in progress. However, a corpus that is ready to be shared with the academic community is better placed in a repository that guarantees the stability of services on the one hand, and the visibility among researchers on the other hand. Nowadays the number of repositories is fairly big, most institutions want to have their own common storing space and to promote it among other researchers. Open research specialists list Zenodo, Figshare, Dryad, 3TU.Datacentrum or UK Data Archive (Fenrich et al., 2016). There are however some more commonly known, international and interinstitutional data storages:

- GitHub (GitHub, Inc., 2019) – it is actually a development platform intended for cooperation in software development, that is hosting and reviewing the code. However it is commonly used for storing linguistic data. Notably the project Universal Dependencies hosts its treebanks and accompanying documentation in GitHub (including English, Polish and Belarusian data). GitHub's advantages include the open status of each project – it is easy to complement data whenever needed, keeping track of all changes at the same time, and the changelog serves as a part of documentation. Additionally, GitHub is used for showcasing one's work, as many companies look into this repository's profiles when looking for recruits.
- ELRA (mentioned in Section 2.1) – as a partner of the Open Language Archives Community (OLAC) ELRA deposits spoken, written and terminological resources. The advantage of using the services of ELRA is a good technological and legal support throughout the whole process of sharing the data. Additionally, data deposited in the ELRA catalogue is assigned one of a few categories, e.g. research or commercial use, and depending on the nature of the resource, its creator may decide to put a price tag on it or to distribute it for the academic community only, or to make it available to the public in general.
- CLARIN (mentioned in Section 1.4) – this international consortium offers multiple repositories created by its various members (CLARIN ERIC, 2019), each repository serving slightly different functions and hosting different type of linguistic data. All resources available in CLARIN repositories are visible in the VLO (mentioned in Section 1.5). CLARIN offers generally good support (slightly varying from country to country) and restricts access to the data for people having institutional logins, that is logins of research institutions belonging to the consortium. Logging via institutional login guarantees that the data is going to be used for academic purposes, thus copyrights will not be violated.

Regardless of the type of the repository chosen, the owner of the data needs to remember about the key issues, already signalled in this chapter, that is – the stability of the services (in particular the length of time guaranteed for storing the data), the support at all stages of depositing the data, the accepted format and maximum size of the data (textual formats are universally accepted but this is not a rule in the case of voice data), the format and amount of metadata and the possibility of

faceted searching (this is extremely important in sociolinguistic studies where information concerning gender or age is extremely important for the research).

Last but not least, when the corpus or any other dataset is available for the academic community, even in a widely known repository, it is important to advertise it among other researchers. One really good way is easily available for anyone – an e-mail list, notably the CORPORA list (Hofland, 2007) administrated by the University of Bergen, Norway. This particular list can be accessed via its website containing instructions for subscription, as well as information about preferred content of e-mails distributed among the list users, for example “information and questions about text corpora [...] all types of discussion with a bearing on corpora [...] conference and book announcements RELEVANT TO CORPORA” (Hofland, 2007; emphasis original). It is a perfect channel for communicating with the academic community, not only to ask about resources necessary for a researcher but also to promote one’s resource among its potential users. Secondly, one can publish a note on a new corpus in a peer-reviewed journal, such as the *International Journal of Corpus Linguistics* published by John Benjamins. Lastly, one’s own social channels, such as profiles on ResearchGate.net or Academia.eu might prove useful in spreading the word about our projects.

This chapter has described corpus compilation in all its stages: design, text collection, encoding and alignment, as well as storing and sharing the data. In the first section the issues of the purpose, availability of the texts, size and representativeness, legal status and sampling have been considered in the context of designing the corpus. The second section presented a detailed workflow of the text collection, while the third section touched on the topic of text encoding, that is the tokenisation, lemmatisation, POS-tagging and assigning the metadata. What followed in the fourth section was the discussion of alignment, its types and methods for aligning the text, as well as available programmes dealing with this task. Lastly, the fifth section went on to describe the elements of the process of storing and sharing data, such as choosing the right format, creating the extensive documentation, finding the best repository and finally, promoting the new resource among fellow researchers. What needs to be stressed here is that each decision made on each step should reflect what is the best for one particular dataset under consideration. There are no universal solutions and this chapter is by no means an exhaustive account of the topic of corpus compilation.

### **Chapter 3: Applications of the EPB corpus (A): Translation studies**

This chapter showcases applications of the EPB corpus in translation studies, specifically in three branches, that is theoretical, descriptive and applied studies. The first section of the chapter considers three translation universals (TU), that is explication followed by simplification and levelling out. Using methodologies applied in existing theoretical research, it is possible to explore the topic within language pairs that have not been extensively analysed from this angle. Moreover, some indicators (such as the size of the corpora) are available to be inspected without full access to the corpus and therefore allow for comparison between other pairs, especially English and other languages from the Slavonic group.

Next, in the domain of descriptive translation studies this chapter contains an overall stylistic analysis of the corpus. Utilising certain techniques, namely dendrograms, general stylistic features of the EPB corpus are determined and scrutinised in order to select one group of texts for further detailed analysis. To finish the topic of exploiting the resource in translation studies, a few examples of querying and analysing phraseology for training purposes are demonstrated.

#### **3.1. Theoretical translation studies: explication, simplification and levelling out in Polish and Belarusian translations of English literature**

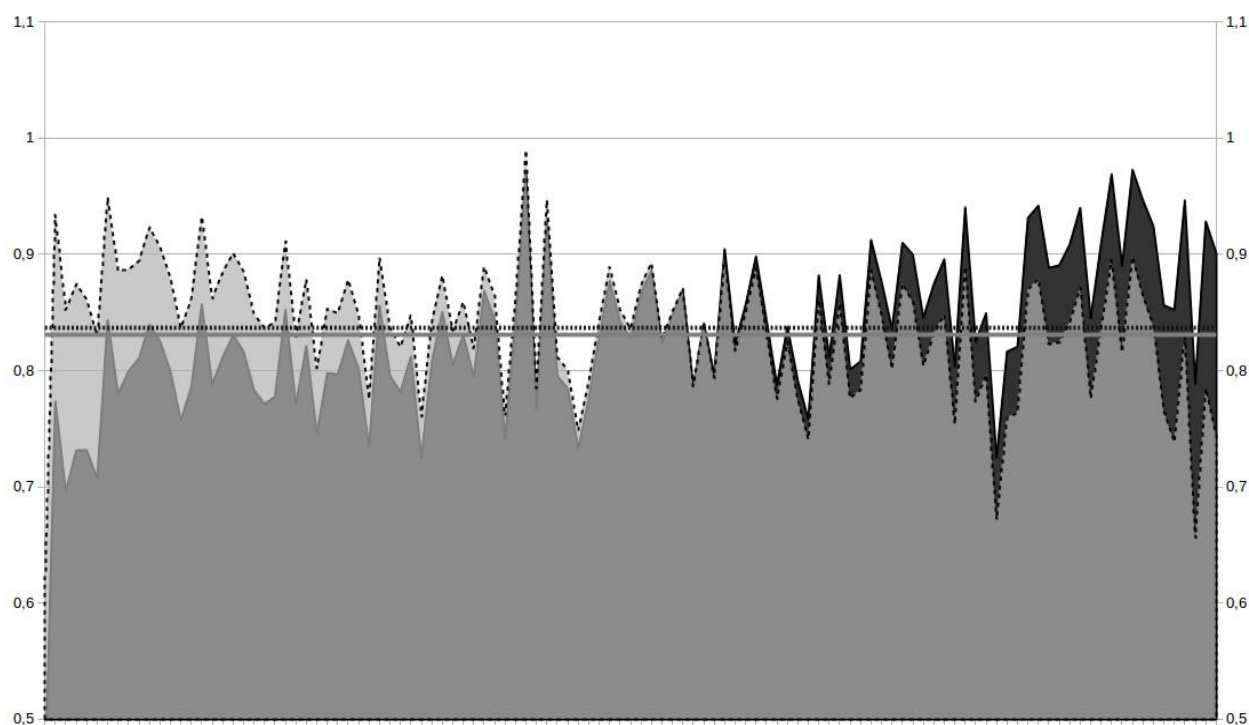
As explained in Section 1.1, the aim of theoretical translation studies is to identify general principles that can be used to explain and anticipate certain phenomena in the process of translation. These principles are most commonly referred to as translation universals. Even though some scholars regard this approach as the right way towards high-level generalisations (Chesterman, 2004), multiple studies undermine the concept, for example via studies of collocations (Maurannen, 2000), or lexical use and syntax (Jantunen, 2001, 2004). Certain studies explain translation universals with lexicogrammatical differences and pragmatic factors (Becher, 2011), or even editing (Bisiada, 2017). Moreover TU proposed by Baker have been described as shallow because they do not take into account higher levels of linguistic organisation and because there is too little language variety to effectively test the TU theory (Teich, 2003) – Baker and her colleagues worked only on English as a target language. Teich suggests that “translations have specific properties that distinguish them from the original texts” but they “are not assumed to be universal” (2003, p. 30) and therefore we should rather focus on particular language pairs when reporting such properties.

Following Teich’s thought this chapter adds new data to the discussion concerning two of the most popular TUs: simplification and explication, as well as levelling out. This material might be significant for several reasons. Firstly, it involves two Slavonic languages which are not often the subject of theoretical translation studies. Secondly, it involves the Belarusian language, which due to its status, is rarely the subject of translation studies in general. And thirdly, it contradicts some of the premises of translation universals, thus tipping the scales in favour of the TU sceptics. As Teich rightly noted, the properties of translation have a lot to do with particular language properties, and this shows in the data concerning English, Polish and Belarusian.

Explication, put simply, is a phenomenon of translating an implicit message in an explicit way. It can be detected on various levels of linguistic analysis – in lexis (as explicit signals of

clausal relations), in syntax (as explicitation of optional syntactic choices) or in discourse (via explicitating shifts in lexical cohesion, conjunctive explicitness and explicative reformulation) (Zanettin, 2013). Among many markers indicating explicitation, it has been reported that translated texts tend to be longer than the source texts (M. Baker, 1996; Frankenberg-Garcia, 2009) and that the translation conveys information explicitly on the syntactical and lexical levels. Especially optional syntactic features of English have been studied in depth, e.g. using ‘that’ in reported speech (M. Baker, 1996), or ‘to be’ in a complement clause, modal ‘should’ in a ‘that’ complement, complementiser ‘to’ and the phrase ‘in order’ (Olohan, 2001). The case of explicit ‘that’ in translation has recently been disproved (Kruger, 2018; Kruger & De Sutter, 2018)(H. Kruger, 2018; H. Kruger & De Sutter, 2018), similarly the length of the translation vs. original text is being discussed and researchers note that it is not always the case that translation is longer. It happens so because along some information being explicitated there is also some portion of text being conveyed in an implicit way and this tendency intensifies in translation constrained by time or space, such as audiovisual translation or websites localization (Gumul, 2020).

Nevertheless, the text length is rarely being scrutinised in explicitation studies concerning written translation and the earlier statements – about translation being longer than the original text – are often taken for granted. Therefore the explicitation analysis in the EPB corpus shall start with examining the translated text sizes. These are represented in the graph below as a relation of the translation size to the original size:



*Illustration 3.3. Differences in contraction factors of translated texts across the EPB corpus.*

The target text sizes vary from just a few hundred to over 200 thousand words therefore it is best to compare them in relation to a particular value. In this case the size of each translated text ( $S_t$ ) is compared to the size of the original ( $S_o$ ) giving the contraction factor ( $F_c$ ):  $F_c = S_t / S_o$ ; one must

remember that this might be a counter-intuitive idea because the higher the contraction factor, the less compressed the text actually is.  $F_c=95\%$  means that translation size is 95% of the original size.

The contraction factors across the EPB corpus are compared and presented in Illustration 3.2. The type chosen for representation of the data is area chart. The X axis (bottom) shows values of the contraction factor for Polish and Belarusian translations corresponding to the texts of the corpus. The data is sorted according to the difference between Polish and Belarusian contraction factors. The light grey colour marks the texts for which the contraction factor is higher in Polish, and black colour marks a bigger contraction factor in Belarusian translations. Two horizontal lines, light for Polish and dark for Belarusian represent the average contraction factors for those languages. One should note that the EPB database contains over 100 titles and therefore in print the names cannot fit in the description of the X axis. Additionally, the scale on the Y axis is shifted – it starts with 0.5 instead of 0 for clearer representation of the data.

What the graph immediately shows is that all translations are shorter than the source text. This contradicts the findings of studies on explicitation. The general conviction is that one of the indicators of explicitation is the extension of the target text. However studies supporting this notion take into account only Germanic and Romance languages. Few studies consider languages from outside of these groups. One example is Cheong's study of English-Korean translations (Cheong, 2006) which contradicts the long-held notion of target text expansion and reveals that in the case of Korean the text is contracted. Researchers report a similar phenomenon also in the case of Romanian-Hungarian translations (Heltai, 2005). Another common source reporting varying contraction factors (both below and over 1) are translation agencies (Andiamo!, 2019; Kwintessential, 2019; Media Lingo, 2019) that publish tables comparing text sizes for various pairs of languages.

The phenomenon of text contraction in the case of the Belarusian and Polish translations is not restricted only to the EPB corpus, it is also supported by the data from other corpora, such as the Open Parallel Corpus (OPUS) (Tiedemann, 2012):

corpus	en tokens	be tokens	Contraction factor
<b>GNOME</b>	1.8M	1.7M	0.94
<b>KDE4</b>	0.8M	0.4M	0.50
<b>Ubuntu</b>	0.5M	0.2M	0.40
<b>EUbookshop</b>	0.2k	0.1k	0.50
<b>total</b>	<b>3.1M</b>	<b>2.3M</b>	<b>Average=0.59</b>

*Table 3.6. Size of English-Belarusian translation data in the OPUS corpus (Tiedemann, 2012).*

corpus	en tokens	pl tokens	Contraction factor
<b>DGT</b> v2018	71.1M	56.2M	0.79
<b>JRC-Acquis</b> v3.0	63.8M	57.6M	0.90
<b>ParaCrawl</b> v1	25.2M	22.7M	0.90
<b>Eubookshop</b> v2	21.0M	18.3M	0.87
<b>Europarl</b> v8	17.3M	14.9M	0.86
<b>GNOME</b> v1	2.8M	2.5M	0.89
<b>TED2013</b> v1.1	3.0M	2.3M	0.77
<b>Tanzil</b> v1	2.8M	2.4M	0.86
<b>ECB</b> v1	1.9M	1.7M	0.89
<b>KDE4</b> v2	1.7M	1.6M	0.94
<b>GlobalVoices</b> v2017	1.1M	1.0M	0.91
<b>GlobalVoices</b> v2015	1.0M	0.9M	0.90
<b>Ubuntu</b>	0.7M	0.5M	0.71
<b>PHP</b>	0.5M	0.2M	0.40
<b>EUconst</b>	0.2M	0.1M	0.50
<b>total</b>	<b>225.8M</b>	<b>197.6M</b>	<b>Average=0.83</b>

Table 3.7. Size of Polish-English parallel data in the OPUS corpus (Tiedemann, 2012).

In both cases the overall number of English tokens is higher. A number of corpora in the OPUS corpus contain technical language (e.g. GNOME, KDE4, Ubuntu) and in the case of the Polish-English corpora the majority of the datas come from non-professional sources (e.g. subtitles created by an Internet community, not necessarily professional translators) or sources without information about the translation direction (e.g. in the case of the data from Tatoeba educational platform). Such resources were not taken into account, although they do follow the pattern emerging from Tables 3.2 and 3.3. One must also notice that almost half of the sources contain legal or business language data translated from English (JRC, ParaCrawl, Europarl, ECB), similarly the technical documents are translations into Polish, while in some cases (such as DGT translation memories or articles from the GlobalVoices) the directionality is hard or impossible to determine.

Returning to the subject of text contraction in the EPB corpus, Illustration 3.2 gives us several other pieces of information. Firstly, the contraction factor varies from just 45% up to 97%. Secondly, the average contraction factor for all Polish and Belarusian texts is the same – 82%, only the median is slightly different – 82% for Belarusian and 83% for Polish. Thirdly, in about half of the cases the Polish translation is longer than the Belarusian and inversely in the other half of cases. Put simply, the differences in contraction factor between the Polish and Belarusian translations fall within the normal distribution. Additionally, when the Polish translation is longer than the Belarusian the difference between contraction factors tends to be higher (up to 23% versus a maximum of 15% when it is the other way round).

Following from these preliminary results, three basic questions need to be answered:

1. Why are all the target texts shorter than the originals?
2. Do the contraction factors of Polish and/or Belarusian translations depend on any of the metadata collected for this corpus? Factors that potentially influence the contraction factor are: publication date of the original/translation, date of birth of the author/translator, gender of the author/translator, length of the original.
3. Does the difference between contraction factors of the Polish and Belarusian translations depend on any of the metadata?

Considering the first question, several factors can possibly cause the translations to always be shorter. First, the structural differences between English (a Germanic language) and Polish and Belarusian (Slavic languages). The most prominent of such differences is the existence of articles in English. ‘The’ and ‘a’ are among the five most common words in English (Leech et al., 2001), but in Polish and Belarusian translation they are usually omitted (with the exception of some cases when ‘the’ might be translated with a demonstrative pronoun). Another word from the top five is ‘of’ which is commonly used for expressing possession of items or features. In Polish and Belarusian this type of prepositional phrase (e.g. ‘possession of items’ – three words) is replaced with a noun phrase where the object is inflected (e.g. ‘posiadanie rzeczy’/‘уладанне рэчамі’ [uładannie rečami] – two words). Also the personal pronoun ‘I’ and the preposition ‘to’ are commonly omitted in both Polish and Belarusian. These phenomena are visible in nearly every sentence of every text (see Appendix 3 for detailed analysis of a few excerpts from the EPB corpus).

Structural differences lie as well in the syntactic properties of Slavic languages. Usage of the conditional in Polish is a good illustration of this issue. The English conditional comprises two words (would + verb; e.g. ‘would try’) whereas in Polish this grammatical property is conveyed by adding a suffix to the verb (verb+suffix; ‘próbowałby’). Querying the first ten segments of a 1.5-million-words subcorpus<sup>7</sup> of the EPB corpus turns up a handful of such instances but it also reveals that other uses of ‘would’ are translated with a different number of words:

No.	Source text in English	Polish translation	F <sub>c</sub>
1	And then, one Thursday, nearly two thousand years after one man had been nailed to a tree for saying how great it <b>would be</b> to be nice to people for a change, one girl sitting on her own in a small cafe in Rickmansworth suddenly realized what it was that had been going wrong all this time, and she finally knew how the world could be made a good and happy place. (word count = 72)	I oto nagle pewnego czwartku prawie dwa tysiące lat po tym, jak pewien człowiek został przybity gwoździami do drzewa za to, że mówił, jak to <b>byłoby</b> świetnie być dla odmiany miłymi dla siebie nawzajem, pewna dziewczyna siedząca samotnie w małej kawiarni w Hickmansworth nagle zrozumiała, co przez cały czas szło źle i wiedziała już, jak można sprawić, aby świat stał się dobrym i szczęśliwym miejscem. (wc = 65 )	0.90

<sup>7</sup> A subcorpus of sentence-aligned samples from 56 corpus texts, where each language part contains approx. 500 thousand words.

No.	Source text in English	Polish translation	F <sub>c</sub>
2	This time it was right, it would work, and no one <b>would have</b> to get nailed to anything. (wc = 18)	Tym razem wszystko by się zgadzało, wszystko by zadziało i nikt nie <b>zostałby</b> do niczego przybity. (wc = 16)	0.89
3	It <b>would sort</b> itself out, he'd decided, no one wanted a bypass, the council didn't have a leg to stand on. (wc = 23)	Samo się jakoś <b>zalatwi</b> , doszedł do wniosku. Rada nie ma nic do gadania. Komu potrzebna jest autostrada? (wc = 17)	0.74
4	It <b>would sort</b> itself out. (wc = 5)	Samo się <b>zalatwi</b> . (wc = 3)	0.60
5	He tried to make his eyes blaze fiercely but they just <b>wouldn't do</b> it. (wc = 15)	Spróbował zmusić swoje oczy do miotania błyskawic, ale nic z tego <b>nie wyszło</b> . (wc=13)	0.87
6	They often wish that people <b>would</b> just once and for all <b>work</b> out where the hell they wanted to be. (wc = 20)	Ci z punktu C chcieliby, żeby ludzie <b>zdecydowali</b> się raz na zawsze, gdzie, do cholery, chcą być. (wc = 17)	0.85
7	He <b>would have</b> a nice little cottage at point D, with axes over the door, and spend a pleasant amount of time at point E, which would be the nearest pub to point D. (wc = 34)	<b>Miałby</b> tam ładny domek z toporami nad drzwiami i spędzałby dowolną ilość czasu w punkcie E, który byłby najbliższym pubem. (wc = 20)	0.59
8	Have you any idea how much damage that bulldozer <b>would suffer</b> if I just let it roll straight over you?" (wc = 20)	— Panie Dent — powiedział — czy wie pan, jakiego uszczerbku <b>doznałby</b> ten buldożer, gdybym na przykład wpadł na pomysł, aby pozwolić mu przejechać po panu? (wc = 23)	1.15
9	For instance he <b>would</b> often <b>gatecrash</b> university parties, get badly drunk and start making fun of any astrophysicist he could find till he got thrown out. (wc = 26)	<b>Miał</b> na przykład w <b>zwyczaju pojawiać się bez zaproszenia</b> na przyjęciach uniwersyteckich, gdzie upijał się, a potem wyśmiewał z każdego napotkanego tam astrofizyka tak długo, aż go wyrzucano. (wc = 28)	1.08
10	Sometimes he <b>would</b> get seized with oddly distracted moods and <b>stare</b> into the sky as if hypnotized until someone asked him what he was doing. (wc = 25)	<b>Wpatrywał się</b> wtedy w niebo jak zahipnotyzowany, aż wreszcie ktoś go pytał, po co to robi. (wc = 16)	0.64

Table 3.8. 'Would' and its Polish translations in the EPB corpus (first ten hits).

Examples 3, 4 and 6 are cases of the future tense, and they are translated with just one word which, again, results in a shorter target text. Additionally, example 6 is a particular case of a wished future where the construction is translated into Polish with the use of a past tense. Examples 5, 9 and 10 illustrate the past tense and, unlike other examples they are translated with a varying amount of words. Example 5 contains a single-word translation (with accompanying negation, as in the original example), ex. 10 is two-word because the Polish equivalent is a reflexive verb and there is need for the reflexive pronoun. However ex. 9 contains an English verb with no equivalent in Polish therefore a semantic explicitation is needed.

Interestingly, even though the conditional construction in Belarusian consists of two words (unlike in Polish, the conditional suffix is always separate from the verb) the cases of past/future tense uses of 'would' are conveyed, similarly to Polish, with a lower number of words. Moreover, even in such a small sample it is visible that Polish translation is usually shorter than English source text.

Another example of syntactic difference in the case of Belarusian that causes the translation to contract is omission of the verb ‘to be’ in the case of predicative expressions following the verb ‘to be’ (‘to be’ has almost exclusively auxiliary function in Belarusian). As a result, the English sentence, such as ‘I am hungry’ is usually shorter in Belarusian – ‘Я галодны’ [Ja hałodny].

No.	Source text in English	Belarusian translation
1	What the hell <b>am I doing</b> in the pub, Ford?”	Дык якой халеры, Фольксвагене, <b>я раблю</b> ў карчме?!
2	I worked hard to get where <b>I am</b> today, and I didn’t become captain of a Vagon constructor ship simply so I could turn it into a taxi service for a load of degenerate freeloaders.	Я выжыльваўся з усяе моцы, каб апынуцца на тым месцы, дзе <b>знаходжуся</b> цяпер, і пасада капітана Воганаўскай будаўнічай эскадры не далася мне нагэтулькі лёгка, каб я дазволіў ператварыць свае караблі ў таксоўку для бадзячых дармаедзін.
3	”The argument goes something like this: ‘I refuse to prove that I exist,’ says God, ‘for proof denies faith, and without faith <b>I am nothing.</b> ’	Довад гучыць неяк так: «Бог кажа: - Я адмаўляюся даказваць, што я існую, паколькі любы доказ адмаўляе веру, а без веры <b>я нішто</b> ».
4	”Hell,” he said, ”how <b>am I going to operate</b> my digital watch now?”	Як жа <b>мне</b> цяпер <b>насіць</b> мой электронны гадзіннік?!
5	” <b>I’m not getting</b> you down at all <b>am I?</b> ” he said pathetically.	- <b>Я ж вас не прыгнятаю?</b> - поўным невымоўнае тугі голасам спытаў ён.
6	”I want you to know that whatever your problem, <b>I am here</b> to help you solve it.”	- Я хачу, каб вы ведалі: якая б праблема вас ні дапякла, <b>я тут</b> , каб дапамагчы.
7	Er, excuse me, who <b>am I?</b>	«Э-э, выбачайце, хто <b>я такі?</b> »
8	Why <b>am I here?</b>	«Чаму <b>я тут?</b> »
9	” <b>I am</b> simply the second greatest <b>computer</b> in the Universe of Space and Time.”	- <b>Я</b> проста другі найлепшы <b>камп’ютар</b> ўсіх Галактык, часоў і народаў, і кропка.
10	I think therefore <b>I am</b>	«Я мыслю, а знакам тым <b>існую</b> »

Table 3.9. 'Am' and its Belarusian translation in the EPB corpus (first ten hits).

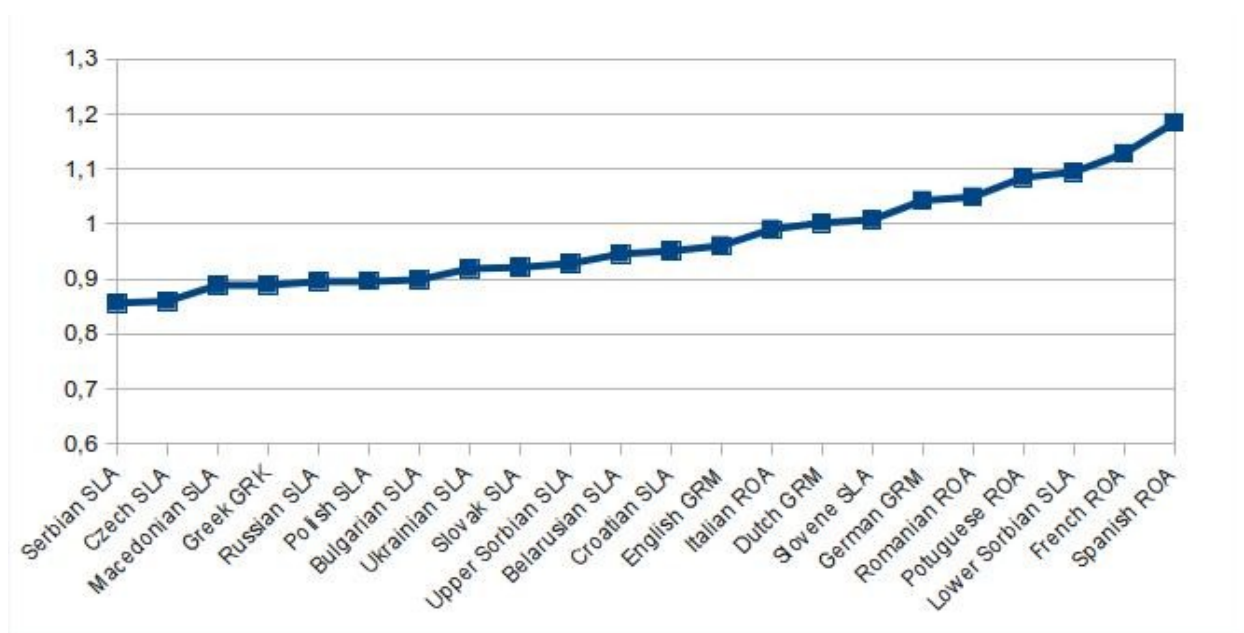
Examples 3, 6, 8 and 9 demonstrate predicative expressions in which the Belarusian translation consists of the pronoun and predicative without the verb. Ex. 1 and 5 are cases of the present continuous tense which does not exist in Belarusian and is translated with one verb corresponding to the present simple tense. Similarly, the construction ‘to be going to’ (presented in ex. 4) is shorter when conveyed in Belarusian. Additionally, the question tag from ex. 5 is reflected by a conditional particle, which also results in shortening the original two-word construction to just one word. Ex. 2, 7 and 10 demonstrate cases of using the verb ‘to be’ in its main function and in two cases they are translated with the use of a synonymous verb without the pronoun, and in one case there is a two-word translation due to the application of a word having a pragmatic function, namely underlining the confusion of the speaker.

Other examples of explicitation motivated by lexicogrammatical inconsummerability (Corness, 2014) (put differently, lack of common basis) include mainly the translation of the Past Perfect tense – a construction that does not exist in Polish, nor in Belarusian, and is rendered as Past Simple

with occasional addition of modifiers pointing to a particular point in time. According to Corness (2014), Polish translations contain the following renditions:

- had heard → already heard earlier
- before X had been made → before X was made
- had driven... taken... established → drive... took... established
- had carried → once... carried

The hypothesis that structural differences influence the contraction factor seems to be confirmed by analysis of other linguistic pairs, e.g. Swedish (Germanic language) paired with a number of Slavonic, Germanic and Romance languages:



*Illustration 3.4. Contraction factor of Swedish language translations in the ASPAC corpus (University of Gothenburg, 2018). Language family codes according to ISO “Codes of representation of names of languages”, capitalised for better visibility.*

Out of 13 Slavic languages paired with Swedish, 11 have a contraction factor lower than 100%, and Greek also falls into that group. This might indicate that highly inflected languages are more likely to produce contracted texts when used for translating from a language such as Swedish or English.

Tables 3.2 and 3.3 based on the OPUS corpus look quite different for other language pairs. Let us first compare English with a different Slavonic language, this time from the Southern group (as opposed to West, represented by Polish, and East, represented by Belarusian):

corpus	en tokens	sr tokens	Contraction factor
<b>GNOME</b>	3.2M	3.5M	1.09
<b>KDE4 v2</b>	0.5M	0.5M	1.00
<b>Ubuntu v14.10</b>	0.7M	0.3M	0.43
<b>EUbookshop v2</b>	83.4k	58.1k	0.70
<b>GlobalVoices v2017q3</b>	0.7M	0.4M	0.57
<b>total</b>	<b>5.1M</b>	<b>4.7M</b>	<b>Average=0.76</b>

*Table 3.10. Size of English-Serbian parallel data in the OPUS corpus (Tiedemann, 2012).*

Even though one source contains more tokens in the Slavonic language and one is of equal size, the tendency is the same – the Serbian part is shorter than the English and the target text has an average 0.76 contraction factor. Let us move now to comparison within the same group. To do that we should take a look into Czech-Polish (both West Slavonic languages) data from the OPUS collection:

corpus	cs tokens	pl tokens	Contraction factor
<b>DGT v2019</b>	93.5M	95.4M	1.02
<b>JRC-Acquis v3.0</b>	55.5M	56.3M	1.01
<b>QED v2.0a</b>	4.2M	4.0M	0.95
<b>Eubookshop v2</b>	15.6M	15.6M	1.00
<b>Europarl v8</b>	14.8M	14.9M	1.01
<b>GNOME v1</b>	2.5M	2.5M	1.00
<b>Tanzil v1</b>	0.3M	0.3M	1.00
<b>ECB v1</b>	1.1M	1.1M	1.00
<b>KDE4 v2</b>	0.6M	1.0M	1.67
<b>Ubuntu v14.10</b>	0.4M	0.5M	1.25
<b>PHP v1</b>	0.1M	0.2M	2
<b>EUconst</b>	0.1M	0.1M	1.00
<b>total</b>	<b>188.7M</b>	<b>181.9M</b>	<b>Average=1.16</b>

*Table 3.11. Size of Czech-Polish parallel data in the OPUS corpus (Tiedemann, 2012).*

In the case of these two Slavonic languages there is no consistent pattern. Out of 12 sources five are of the same size, one is bigger in the Czech language and six are bigger in Polish. The contraction factor stays low, except for the subcorpora under 1 million words. In these cases the bigger difference is clearly caused by their small relative size, because as little as 100 thousand words might account for as much as 50% of the corpus.

This short experiment on data from various language groups suggests that the differences in text length that are traditionally associated with the qualities of translation might have more to do with the characteristics of each language itself. As a start of this line of investigation an analysis of the English part of the EPB corpus was performed in the AntConc tool. Out of almost 2 million words (1.97 million tokens) the ten most frequent are:

Rank	Frequency	Word
1	114682	the
2	63170	and
3	45262	of
4	44596	a
5	44068	to
6	35932	he
7	29930	in
8	28670	I
9	28629	was
10	28461	it

*Table 3.12. Most frequent words in the English texts in EPB.*

The article ‘the’ alone constitutes almost 6% of all tokens. Summarizing the count for words mentioned earlier in this chapter (the, a, of, to, I) gives 14.05% of all tokens (27.73 thousand tokens). Even if all these words were omitted in the Polish and Belarusian translations it still does not fully explain the size of the translational part of the EPB corpus, because the average contraction factor is 82% rather than 86%. In addition, it is not always the case that ‘the’ or ‘of’ are omitted. Occasionally the article is translated as a demonstrative pronoun (e.g. *ten/гэты* [hety], en. *this*) and the preposition may be used, e.g. in describing the production material (*wykonany z marmuru/ выраблены з мармуру* [vyrableny z marmuru], en. *made of marble*).

To investigate the impact of syntactic features in the contraction factor two main indicators mentioned in previous sections have been tested. Firstly, auxiliary ‘would’ which can influence the Polish translations was searched in the English sub-corpus. Its frequency is 9111 which is less than 0.3%, definitely not enough to affect the contraction factor. Secondly, ‘to be’ and its present tense forms were searched for; they amount to less than 2% of all tokens (nearly 58 thousand hits) and therefore cannot explain the contraction factor in Belarusian translations.

One more fact remains puzzling – according to researchers most translational texts contain not only syntactic explicitation which is classified as ‘obligatory’ (Klaudy, 2009), but also ‘optional explicitation’, which is motivated by stylistic preferences. It is not vital for grammatical correctness, however its lack causes the text to sound unnatural and clumsy. Apart from that there is *pragmatic explicitation* (Klaudy, 2009) or *explicitation motivated by a sense of cultural distance*

(Corness, 2014), illustrated by example for *Fir Bolgs* translated as *Fir Bolg tribe* or *Maros as the river Maros*. Moreover translators may use ‘over-interpretation’:

the arbitrary addition of information which is not made explicit in the given sentence of the source text because it is clearly derivable from the context, being explicitly mentioned elsewhere in that text, and therefore redundant (Corness, 2014, p. 17).

All above-mentioned phenomena can contribute to the extension of the target text, which suggests there must be an additional factor causing the contraction. A possible explanation of the difference in the texts sizes is linguistic simplification, which is also asserted to be one of the translation universals. The exploration of this possibility is presented later in this chapter when the indicators of simplification proposed by Mona Baker (1996) are tested.

Returning to the question of the contraction factors of Polish and Belarusian translations, a series of tests was conducted to find possible relations between the relative sizes of the target texts as well as the difference between them, and the metadata. Two types of tests were conducted:

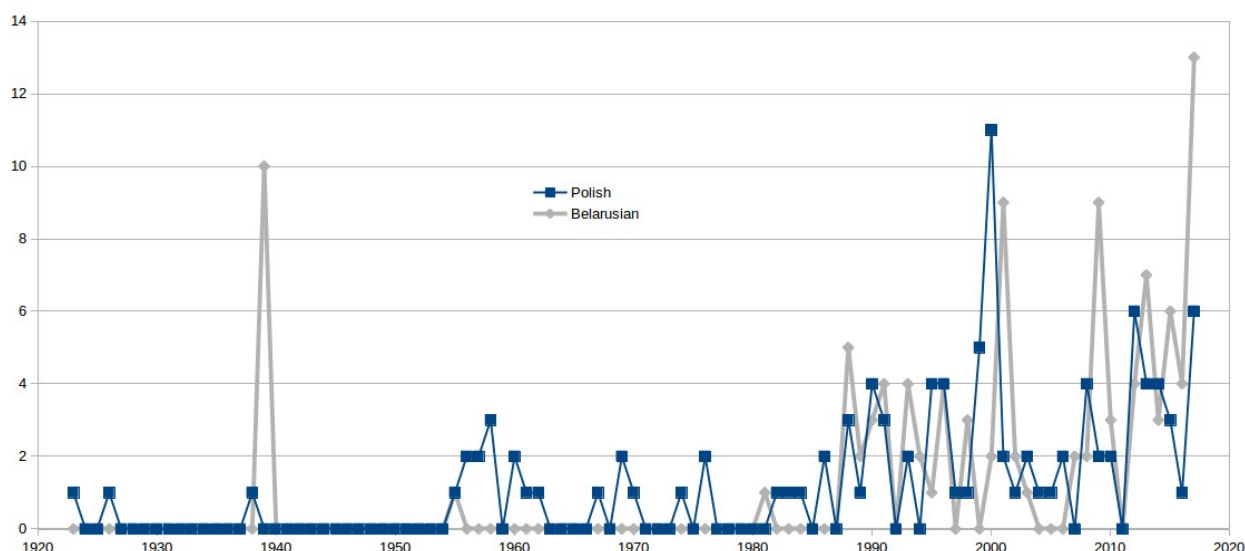
- Pearson correlation coefficient (PCC): it can have a value between -1 and +1, where -1 means total negative correlation, 0 means no correlation and +1 means total positive correlation. The value of PCC indicates the strength of the correlation: if it is between -0.5 and +0.5 the correlation is regarded as weak, anything below -0.5 and over +0.5 is regarded as strong correlation. PCC only indicates the correlation between variables, it does not indicate which one is the dependent variable – put differently, it does not indicate which variable influences the other one;
- p-value (p): probability value. This indicates how probable it is that the correlation between the data is due to chance; any correlation that has  $p > 0.05$  is regarded as random.

A table in Appendix 2 represents the results of the test of correlation between 16 external factors and the contraction factors of Polish and Belarusian translations as well as the differences between them. Each result contains correlation coefficient (first row), probability value (second row) and the size of sample tested (third row). The size of sample varies due to gaps in the database (the metadata is still being supplemented). Grey colour is used to mark all results with a p-value indicating a statistically significant result.

As evident from the table, only one factor is potentially correlated with the contraction factor: the difference between the dates of birth of the author and the Belarusian translator exhibits weak correlation with the contraction factor of Belarusian translations, as well as with the difference between Belarusian and Polish contraction factors. This result may be disregarded, firstly, because the PCC value is really low, and secondly, because correlation does not mean cause-and-effect relationship. In other words, the correlation of two factors may be, in fact, caused by a third factor that has not been accounted for during the tests.

What does not show in the results, despite being often regarded as influencing the text, is the date of publication, or more precisely, the level of censorship present in each period. Post-war Poland and Belarus were well known for their well developed censorship apparatus and yet this is not reflected in the EPB correlation tests. The reason for that might be twofold. Firstly, the EPB database is not a

complete database of English-Polish or English-Belarusian translations. As explained in Chapter 2.1, only the texts translated both into Polish and Belarusian were included; moreover the corpus might be described as opportunistic, because only one, more accessible text was chosen whenever more than one translation existed. Nevertheless, the collected data reflects the fact that, since 1956, that is, since Stalin's death, there was a rise in the number of translations from English and from 1989 the market in Anglophone literature started flourishing. This phenomenon is illustrated with the graph below:



*Illustration 3.5. Distribution of the number of works published in years 1920-2020 in the EPB database.*

The X axis of the graph shows the dates of publication of the translations (dark thin line and square points represent Polish, light thick line and rhombus represent Belarusian), the Y axis represents the number of publications per year. What is visible is the temporary increase in the late 1950s and a steep rise in the number of publications in the 1990s and the 21<sup>st</sup> century. The spike in Belarusian publications in 1940 is due to Kipling's *Just So Stories* – a collection of short stories, each of which has been treated as a separate text, hence the high number in this period.

A second reason for the lack of correlation between the text size and its year of publication might be due to overestimation of the effects of censorship. As Looby notes, “no censor ever changed a text as much as its translator” (2015, p. 7). Looby's extensive study of censorship in the Polish translations of English literature between 1944 and 1989 shows that the censors intervened mainly in terms of stylistics and that the translators frequently applied self-censorship in order to increase the chances of a translation being published. The issue of translator style is investigated further in Chapter 3.2.

To conclude this section, three issues related to explicitation have been analysed, the difference in length of the original texts and the translations being the first one. Evidence both from academic and non-academic sources indicates that the contraction factor is strongly related to the type of languages, precisely to the language group it belongs to. As illustrated by data from various corpora,

in the case of Polish and Belarusian, the translations from English tend to be shorter which is to a great extent (however not entirely) caused by structural differences between these languages and the English language. Another potentially significant factor is linguistic-semantic simplification which is explored in the following pages. Secondly, as shown by correlation tests, none of the metadata considered in this study – publication date of the original/translation, date of birth of the author/translator, gender of the author/translator, length of the original – influence the contraction factor of Polish and Belarusian translations, nor the difference between them to a significant extent. It should be concluded then, that other external factors may be important, such as the style of the author and translator. These possibilities are explored in Chapter 3.2.

Simplification, or the “tendency to simplify the language used in translation” (P. Baker, 2012, p. 250), has been investigated by a number of scholars (Blum-Kulka & Levenston, 1978; Klaudy, 1996; Toury, 1995) and more systematically, with the use of corpora, by Mona Baker (1993, 1995, 1996, 2004), who identified general indicators of simplification, such as lower average sentence length, lower lexical density (ratio of the number of content words to the number of grammatical words) and lower lexical variety (type-token ratio). These outcomes have been confirmed by other scholars (Corpas Pastor et al., 2008; Laviosa, 2002; Scarpa, 2006; Xiao, 2010), however, the same studies revealed as well that the direction of translation, or languages involved, may impact the simplification indicators. Another problem commonly emerging in TU research is the discrepancy in methodology, that is, using too wide a range of measures or using measures which are understood in different ways. The result of such practices is that researchers are unable to efficiently compare results concerning the same languages, registers etc.

In order to address this issue, the methodology applied in this section is largely based on procedures proposed by Laviosa (1998), and later implemented by Grabowski (2013) in his study of TU in translational literary Polish. The measures used in this section are as follows:

- (a) standardised type/token ratio (M. Scott, 2010) and its standard deviation,
- (b) proportion of high-frequency words measured with frequency profiles (Baroni, 2009),
- (c) mean sentence length and its standard deviation.

The material collected for this research contains only non-translational English and translational Polish and Belarusian, therefore this study uses additional datasets to compare the above-mentioned measures across the three languages. In the case of English, the results are compared to results obtained by Laviosa in her 1998 study. With respect to Polish, the comparison is made with a sample by Grabowski (2013) which consists of seven 20<sup>th</sup> century novels of total length over 352 thousand words. As the author notes, “the data analyzed in this study, which are not representative of all contemporary literary translations into Polish, only warrant restricted claims as to the investigated universalist hypothesis” (L. Grabowski, 2013, p. 259)(Grabowski, 2013, p. 259). An additional source for comparison is Grabowski’s study of two Polish translations of Nabokov’s *Lolita* juxtaposed with a custom-made 478-thousand-word corpus of original Polish language (2012). Regarding Belarusian, a corpus of seventeen novels and short stories from the 20<sup>th</sup> and 21<sup>st</sup>

century, of over 632 thousand words has been compiled for the purposes of comparison (see Primary sources B in References for further details).

The first measure implemented in this study is an improved and more advanced version of the traditional type/token ratio (TTR). TTR, proposed by Baker and later by Laviosa, is size sensitive, meaning that a larger text will have lower TTR value and shorter text – higher. Moreover, as the type is usually understood as a different form, not a different lemma, highly inflectional languages, such as Belarusian and Polish, will contain many more types, e.g. any noun in English can take up to two forms (one singular and one plural), while the same noun might have up to eleven different forms in Belarusian and thirteen in Polish, due to declension. Hence standardised TTR (STTR), proposed by Scott (2010) is used in this study. The difference lies in the fact that the STTR is counted every 100 words (or any other amount chosen as the STTR base), rather than for the whole text at once, and then the average value for all 100-word chunks accounts for STTR. Within these shorter chunks the differences between the number of declension forms is less prominent, as only some of them have a chance to occur, unlike in the full text where one can usually find any given noun or verb in all available forms. Similarly to studies by Laviosa and Grabowski, all tests were run in WordSmith and their outcomes are presented below, in Table 3.9. Following the previous studies, apart from STTR value itself the standard deviation (SD) is also calculated. SD indicates how varied is the population in the sample, low SD means that the values lie close to the mean, and a high SD points to more dispersion in the values. SD is used to measure statistical confidence and therefore is it a common practice to report it alongside test results.

<b>Language</b>	<b>Corpus</b>	<b>STTR</b>	<b>STTR standard deviation</b>
<b>English</b>	Non-translational EPB	72.52	27.78
<b>Polish</b>	Translational EPB	83.65	16.58
	Translational-1 Grabowski (2012)	66.07	32.91
	Translational-2 Grabowski (2012)	70.03	28.82
	Translational Grabowski (2013)	61.95	36.41
	Non-translational Grabowski (2012)	60.40	39.75
	Non-translational Grabowski (2013)	59.43	38.48
<b>Belarusian</b>	Translational EPB	82.66	19.33
	Non-translational (App. 4)	82.52	18.38

*Table 3.13. Standardised type/token ratios across corpora of English, Polish and Belarusian (calculated with WordSmith).*

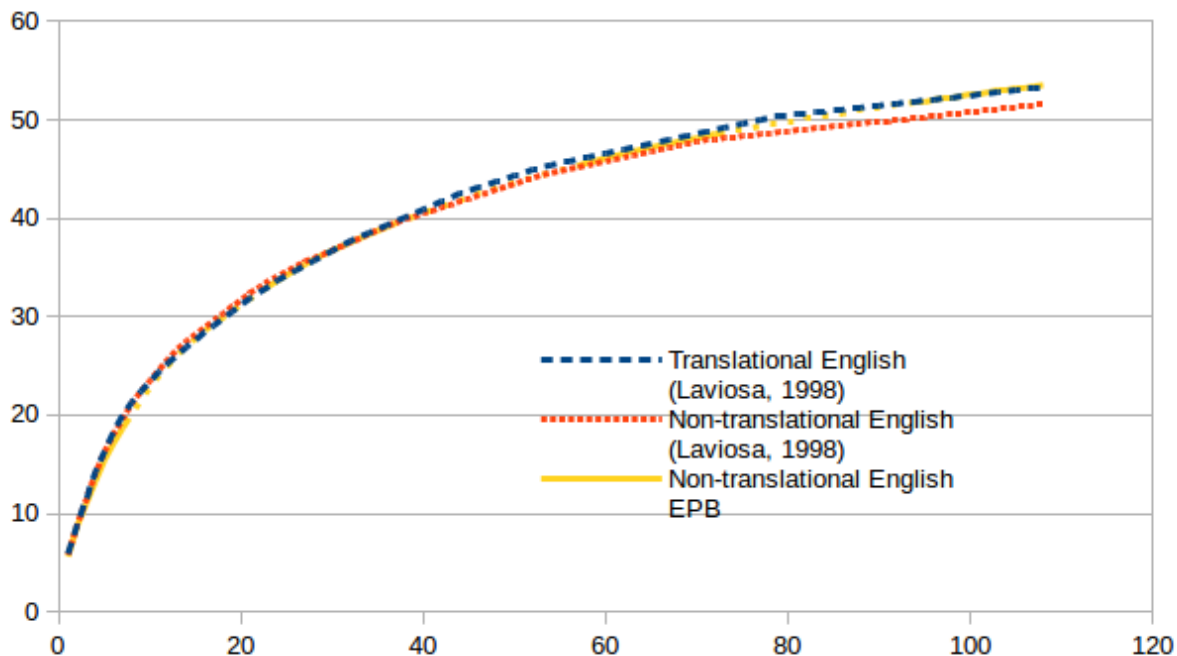
According to the universalist hypothesis, translations are less lexically rich than the texts originally written in a given language. The STTR measure for samples by Grabowski shows, however, that the non-translational Polish texts' STTR value range is 59.43–60.40, while translational Polish has higher values, varying from 61.95 to 70.03, thus invalidating the hypothesis of lexical simplification. The results for the EPB data are even higher – 83.65. The difference between

translational and non-translational Belarusian is much less prominent, however translated texts are, similarly to the case of Polish, more lexically rich in terms of STTR. To sum up, using the standardised type/token ratio as the measure of lexical richness does not confirm the simplification hypothesis in the Polish and Belarusian languages.

The high frequency words proportion is regarded as another measure of simplification. Laviosa (1998) asserts that the proportion of high-frequency words to the total word count is higher in translational corpora, thus indicating less variety. Here this proportion is measured with frequency profiles (Baroni, 2009): for the first 200 most frequent words the proportion of first, then first 10, first 50, first 100 and first 200 words is calculated and compared. The results, based on the AntConc calculations, are represented in Table 3.10 and in Illustration 3.5:

rank	% of all tokens		
	Non-translational English EPB	Non-translational English (Laviosa, 1998)	Translational English (Laviosa, 1998)
1	5.62	5.8	5.9
10	22.85	23.5	23.4
50	43.61	43.5	44.3
100	52.55	50.8	52.5
108	53.53	51.6	53.3

*Table 3.14. Frequency profiles of translational and non-translational corpora of English.*



*Illustration 3.6. Frequency profiles in corpora of English.*

What should be noted first is that in the course of data preparation an error has been found in Laviosa's study. According to the data from the Translational English Corpus (TEC) provided in the

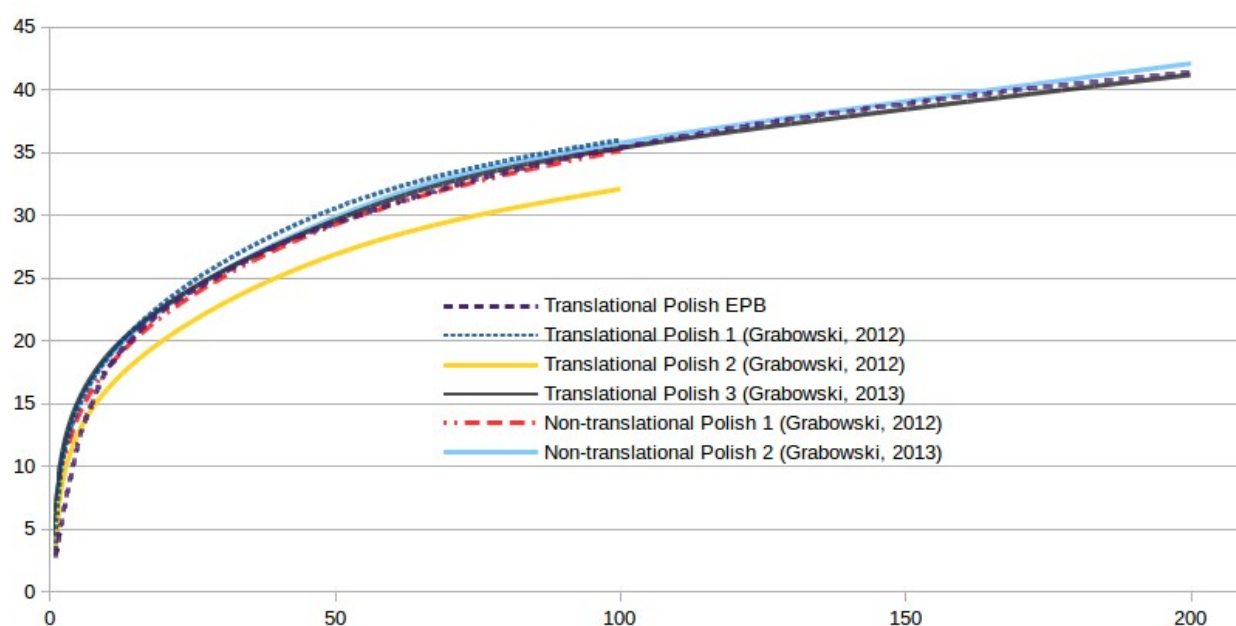
appendix to her article, the list-head of the translational component of the corpus comprises 53.3 rather than 56.20%, thus significantly reducing the difference between the translational and non-translational sample.

According to the universalist hypothesis, the high proportion of the most frequent words in the whole corpus indicates low lexical variety. In the case of English, available data allows comparing 108 most frequent words. In Laviosa's study, according to her prediction, non-translational texts are more lexically varied (as illustrated by the lowest value in Table 3.10 and the lowest positioned line in the graph from Illustration 3.5). However, the frequency profile of non-translational texts from EPB is almost a perfect match for the translational sample by Laviosa. Therefore the simplification hypothesis cannot be confirmed based on the frequency profiles for English.

There are more samples available for testing the Polish language, however let us note that the samples of translational Polish by Grabowski from 2012 contain only one text (a translation of Nabokov's *Lolita*) and therefore they should be treated judiciously. Table 3.12 and Illustration 3.7 present frequency profiles data for the available Polish samples:

	% of all tokens					
rank	Translational Polish EPB	Translational Polish 1 (Grabowski, 2012)	Translational Polish 2 (Grabowski, 2012)	Translational Polish 3 (Grabowski, 2013)	Non-translational Polish 1 (Grabowski, 2012)	Non-translational Polish 2 (Grabowski, 2013)
1	2.76	3.36	3.01	2.98	3.17	3.02
10	17.73	18.61	16.14	18.74	17.80	18.58
50	29.40	30.61	26.90	29.57	29.28	29.86
100	35.40	36.07	32.11	35.34	35.16	35.77
200	41.45	N/a	N/a	41.17	N/a	42.10

*Table 3.15. Frequency profiles of translational and non-translational corpora of Polish.*



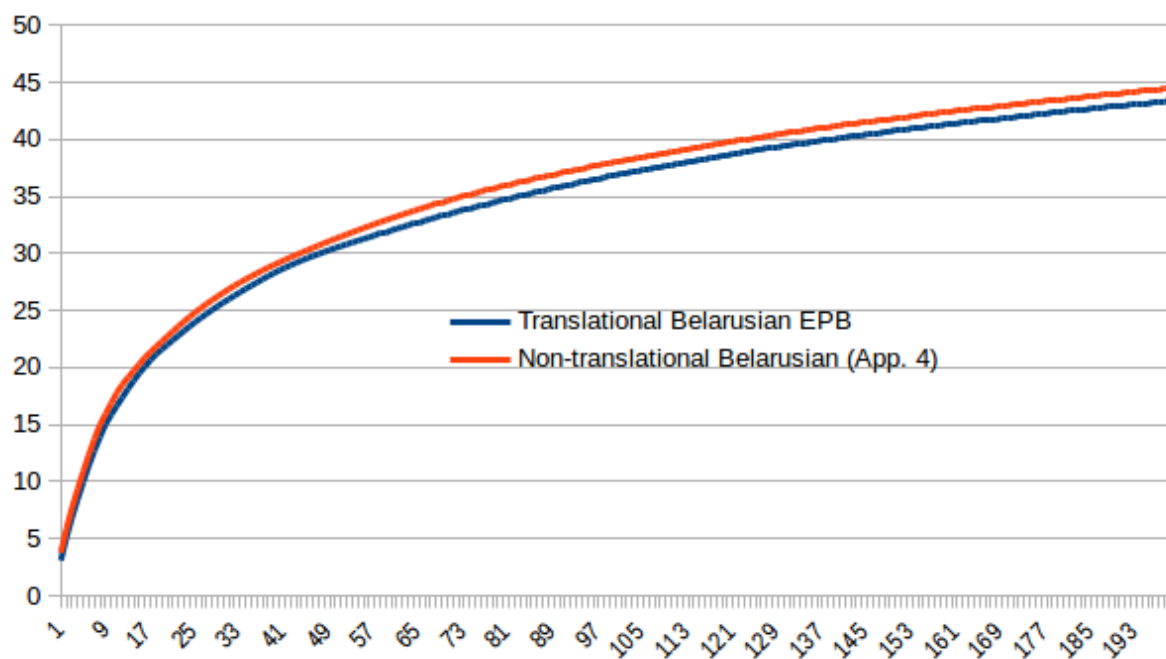
*Illustration 3.7. Frequency profiles in corpora of Polish.*

Assessing by the 100 most popular words, the translational sample of Polish from the EPB corpus falls between two non-translational samples by Grabowski, however at the rank of 200 the EPB sample exhibits more lexical variety. Similarly to the STTR measures samples of Polish do not give conclusive evidence proving the simplification hypothesis.

The Belarusian-language samples are the most sparse, however based on available data it can be determined that the translational language exhibits more lexical variety than non-translational, thus disproving the simplification hypothesis. This is illustrated with Table 3.12 and Illustration 3.7:

	% of all tokens	
rank	Translational Belarusian EPB	Non-translational Belarusian (App. 4)
1	3.09	3.77
10	15.87	16.87
50	30.49	31.30
100	36.88	38.04
200	43.42	44.56

*Table 3.16. Frequency profiles of translational and non-translational corpora of Belarusian.*



*Illustration 3.8. Frequency profiles in corpora of Belarusian.*

To conclude, the study of frequency profiles gives no significant evidence confirming that translational texts have lower lexical variety. The only significant difference indicating such a relation are samples by Laviosa (1998) which have been proven to be erroneous at least in one respect. Comparing Polish and Belarusian samples of translational and non-translational texts does not reveal major differences, and when differences do occur, they mostly contradict the simplification hypothesis.

The third measure considered to indicate simplification is the mean sentence length (MSL). Laviosa (1998) observed in her corpus significantly higher mean sentence length for the translational component (24.09) compared to the non-translational (15.62). Results for the three languages obtained from cross-examination of Laviosa's and Grabowski's studies and supplemented with tests run on the EPB corpus are demonstrated in Table 3.13:

Language	Corpus	Mean sentence length (MSL)	MSL standard deviation
English	Non-translational EPB	12.71	11.03
	Non-translational (Laviosa, 1998)	15.52	1.88
	Translational (Laviosa, 1998)	24.09	11.73
Polish	Translational EPB	10.35	8.63
	Translational-1 (Grabowski, 2012)	17.96	18.97
	Translational-2 (Grabowski, 2012)	17.35	17.75
	Translational 3 (Grabowski, 2013)	9.65	7.64
	Non-translational-1/2 (Grabowski, 2012)	11.14	23.37
	Non-translational-3 (Grabowski, 2013)	10.07	9.80

<b>Belarusian</b>	Translational EPB	10.49	8.55
	Non-translational (App. 4)	10.47	8.76

Table 3.17. Average sentence length in EPB corpus components and corresponding corpora (calculated with WordSmith).

Results for the English part of the EPB corpus are consistent with Laviosa's findings – the mean sentence length is lower in non-translational language. This result, according to Laviosa herself, is contrary to her predictions. In the case of Polish, results are inconclusive. In the 2012 study by Grabowski two translational samples have much higher MSL than the non-translational counterpart, while in his 2013 study it's the other way around. The result for the Polish part of the EPB corpus falls between his two values – when compared to the 2012 sample the EPB result is lower, however it is higher when compared to the 2013 sample. The difference between the Belarusian samples, only 0.02, is negligible – it neither confirms nor disproves the simplification hypothesis. In summary, only the results for the English language exhibit a consistent pattern. This pattern however is evidence of the greater lexical richness of the translational language, rather than non-translational, and thus contradicts the hypothesis of simplification.

With full access to corpora of Belarusian translational and non-translational languages one more universalist hypothesis can be investigated. Levelling out, or convergence, predicts that the translations are more similar to each other than to other texts originally written in a particular language. To test this hypothesis, the methodology of Grabowski (2012) is followed, that is the Principal Component Analysis (PCA) technique is implemented. Simply speaking, PCA turns a huge amount of information extracted from the input data into statistically independent components and organises them according to their importance. The process of PCA was completed with the Stylo package (already mentioned in Chapter 1.3). Stylo, even though a complex R script, offers a simple user interface for conducting basic operations. There are five sets of parameters to be manipulated and for this test the following settings were selected:

- INPUT&LANGUAGE: *plain text* (as opposed to other available options – xml and html), and *other* (as Belarusian is not specified among ten available in the list) encoded in UTF-8
- FEATURES: *words* (as opposed to single characters) with n-gram size 1 (to analyse single words rather than pairs or even bigger combinations), 1000 most frequent words (MFW)
- STATISTICS: PCA with Classic Delta measure
- SAMPLING: *no sampling* (in order to treat each text as a separate entity rather than a set of smaller samples)
- OUTPUT: *in the program window to be exported as JPG*.

The results of the PCA for the Belarusian literature (both original and translated) are presented in the graph. In Stylo the metadata is included in the filenames – any string of characters followed by an underscore becomes a class identifier and each class is assigned a different colour on the graph. In the case of this study two classifiers were used: BNT for non-translational Belarusian (red colour

in the graph, 17 samples) and BT for translational Belarusian (green colour, 113 samples). More dispersion stands for more differences between the elements.

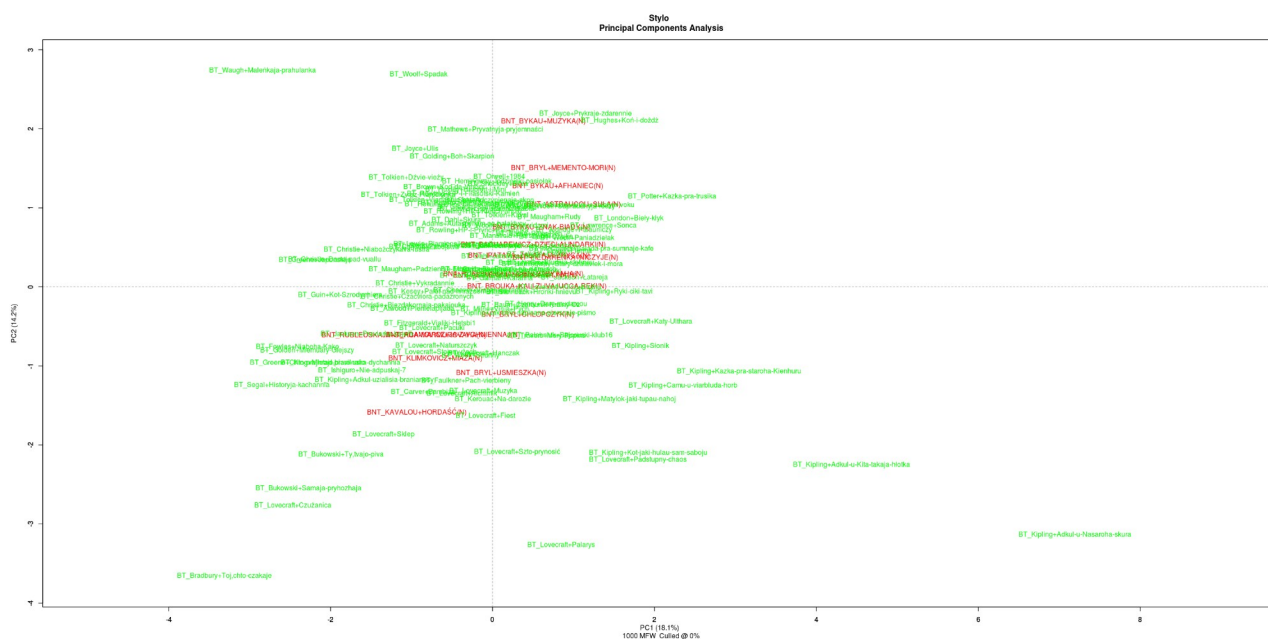


Illustration 3.8. PCA for translational and non-translational Belarusian texts.

Contrary to the results obtained by Grabowski and the levelling out hypothesis proposed by Baker, in the Belarusian sample it is non-translational texts that are more homogeneous. Understandably, this result may be explained by the disproportion between the number of translational and non-translational samples, however the corpus of original Belarusian texts was compiled with diversity as a leading principle and it contains short stories, novellas and novels on a multitude of topics (the only exception being stories for children which are only present in the translational Belarusian corpus). Therefore it is a matching counterpart for the Belarusian part of the EPB corpus and it may be concluded that there is not sufficient data supporting the levelling out hypothesis.

To sum up this section, its first part focused on testing the simplification hypothesis with the use of three measures, STTR, frequency profiles and MSL. Consistent methodology allowed the results obtained from the EPB corpus to be compared with other researchers' outcomes. STTR measured across different samples proved to be inconclusive. Similarly, the frequency profiles for the Polish samples did not give a clear answer as to whether the simplification hypothesis might be true or not, but the same measure disproves this hypothesis when applied to the English and Belarusian samples. The third measure, MSL, is inconclusive in the case of Polish and Belarusian, however it disproves the simplification in the case of English. No sufficient evidence confirming the simplification hypothesis proposed by Baker has been found among the data, rather some evidence disproving the hypothesis has been demonstrated.

In the second part of this section the main focus was on the levelling out hypothesis stating, that the translated texts are less linguistically diverse than original texts in a given language. This hypothesis was tested across Belarusian samples with the use of the PCA technique implemented in

the Stylo package and it was revealed that the assumption of levelling out in translation is not supported by the available data.

### 3.2. Descriptive translation studies: translator style vs. author style.

Stylometry, or the use of statistical methods in the analysis of literary style, can be traced back to the 19<sup>th</sup> century (D. I. Holmes, 1998), however its application in digital humanities is relatively recent. Nevertheless, stylometry has already proven to be a tremendous tool for authorship attribution, verification and profiling, for dating texts (stylochronometry), and for detecting and dealing with deception in writing style (adversarial stylometry, Brennan, Afroz, & Greenstadt, 2012; Neal et al., 2017). This chapter will deal with the first type of stylometrical subtask, that is with authorship identification by detecting stylistic similarities. Authorship identification is based on a set of textual features characterising an author's idiosyncrasy and, even though the existence of such features has never been categorically proven and therefore the field is accused of lacking a strong theoretical base (Oakes, 2014), stylometry cannot be denied multiple successes. In particular, studies of words serving grammatical functions show statistically meaningful differences between authors, and it is very doubtful that humans, even literature creators, give much conscious deliberation to which function words they use and how often. Stylometry has triumphed in some high-profile author attribution cases, such as J.K. Rowling and the novel "The Cuckoo's Calling" (Juola, 2013).

Analysis of style does not only concern the author of the original text but also of the translation. Translator style can be approached in two ways hugely determining the method of investigating it. From one point of view, translation is simply a replication of the author style and as such it does not need a thorough examination, the focus is on the source text. On the other hand, a growing group of researchers treats translation as an active and creative process and argues that the translator style can and should be investigated. Most notably, Baker (2000) and Saldanha (2011b) propose a corpus approach to such an endeavour. Baker's framework is based on the idea of stylistic features as translator's fingerprints which can be identified via comparison of type/token ratio, average sentence length or the reporting structures in the work of various translators. Saldanha, on the other hand, adds the source text to the equation in order to determine the origin of particular linguistic features. Moreover, she focuses on the way translators present the speech and thought, for example via italics or the connective *that*. Both researchers use their frameworks to compare translations of different source texts and even though this approach is burdened with a handful of problems, it can be used for effective investigation of translator style.

Importantly, over the years stylometric methods have been refined and are now well established in the academic community. Tools such as WebSty (mentioned in Chapter 1.5) or Stylo (mentioned in Chapter 1.3 and 3.1) have made it easier for researchers to study stylistic differences in texts of interest. For the sake of flexibility, in the case study presented in this section it is the latter tool that has been employed. Stylo successfully combines text processing, feature extraction, statistical analysis and the visualisation of the results. A tremendous advantage of Stylo, especially in a multilingual research setting, is the usage of language-independent algorithms, that is extracting the n-grams on the level of tokens and characters, rather than lemmas. Nevertheless, the package supports work with annotated corpora as well.

In the case of the EPB corpus, which is varied and fairly big (over 10 million tokens), it is difficult to identify the points of interest right away and therefore a holistic view of the interweaving styles is needed. What provides such a view is a dendrogram – a visualisation of one of the multivariate analyses, namely cluster analysis (mentioned in Section 1.5). In general, multivariate analyses evaluate frequency data, which is normalised in various ways depending on the statistical measure chosen to rate the similarity. In this case the settings used to perform the analysis play a tremendous role. Stylo’s settings (described in detail in Section 3.1) have been manipulated in the following way.

Firstly, the number of words analysed has been reduced by performing 20% culling. “A culling value of 20 indicates that words that appear in at least 20% of the texts in the corpus will be considered in the analysis” (Eder et al., 2019). This manipulation to the wordlist should prevent any text from being singled out on the basis of using proper nouns which are not present in other texts. Secondly, the tests were run with the use of two rather than just one statistical measure, namely the Classic Delta and Eder’s Delta. Classic Delta is a widely acknowledged statistical measure for attributing authorship, and in terms of doing simply that it works better for languages such as English and German but worse for inflectional languages, such as Polish and Belarusian (Rybicki & Eder, 2011). Eder’s Delta, on the other hand, was specifically designed for highly inflected languages and evidence show that it does perform better on languages such as French (Evert & Proisl, 2015). This measure achieves its goal by using “a ranking factor that reduces the weight of less-frequent words’ z-scores” (Evert et al., 2017), that is the frequencies normalised in a way that reduces the influence of the relatively (in terms of the corpus size) most popular words. Due to the differences in the way different statistical measures work, it has been concluded that it might be more revealing to compare results obtained with both Classic and Eder’s Delta. The rest of the settings, most notably 1000 MFW and words as a base unit of analysis, stayed the same as in the PCA analysis from Section 3.1.

The first dendrogram presents all English texts organised in clusters accordingly to the stylistic similarity calculated based on the Classic Delta measure. The second one is based on the Eder’s Delta. The same colour represents the same author and the spatial distance represent similarity in style:

Stylo  
Cluster Analysis

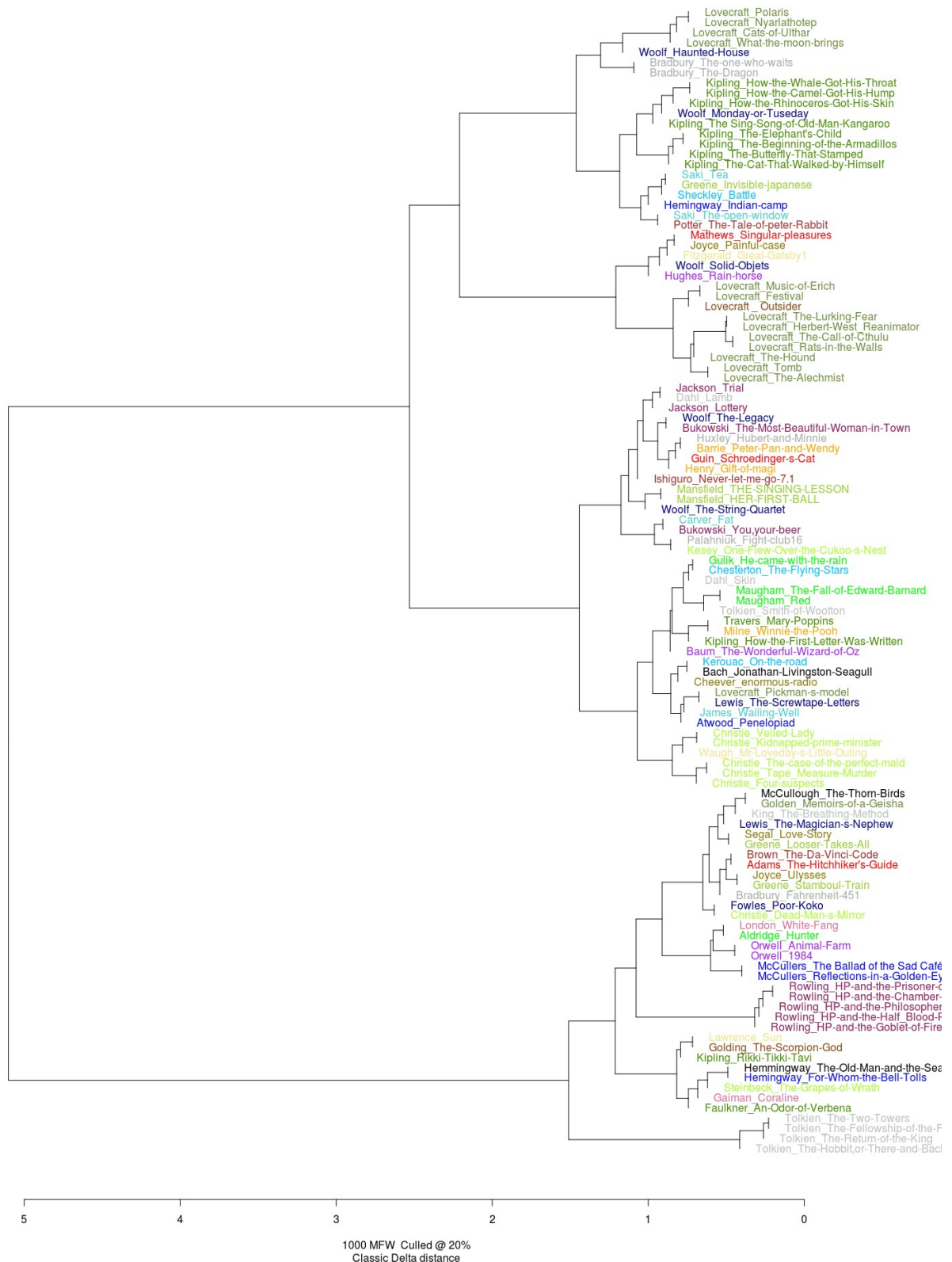


Illustration 3.9. Dendrogram of English texts in the EPB corpus (Classic Delta measure).

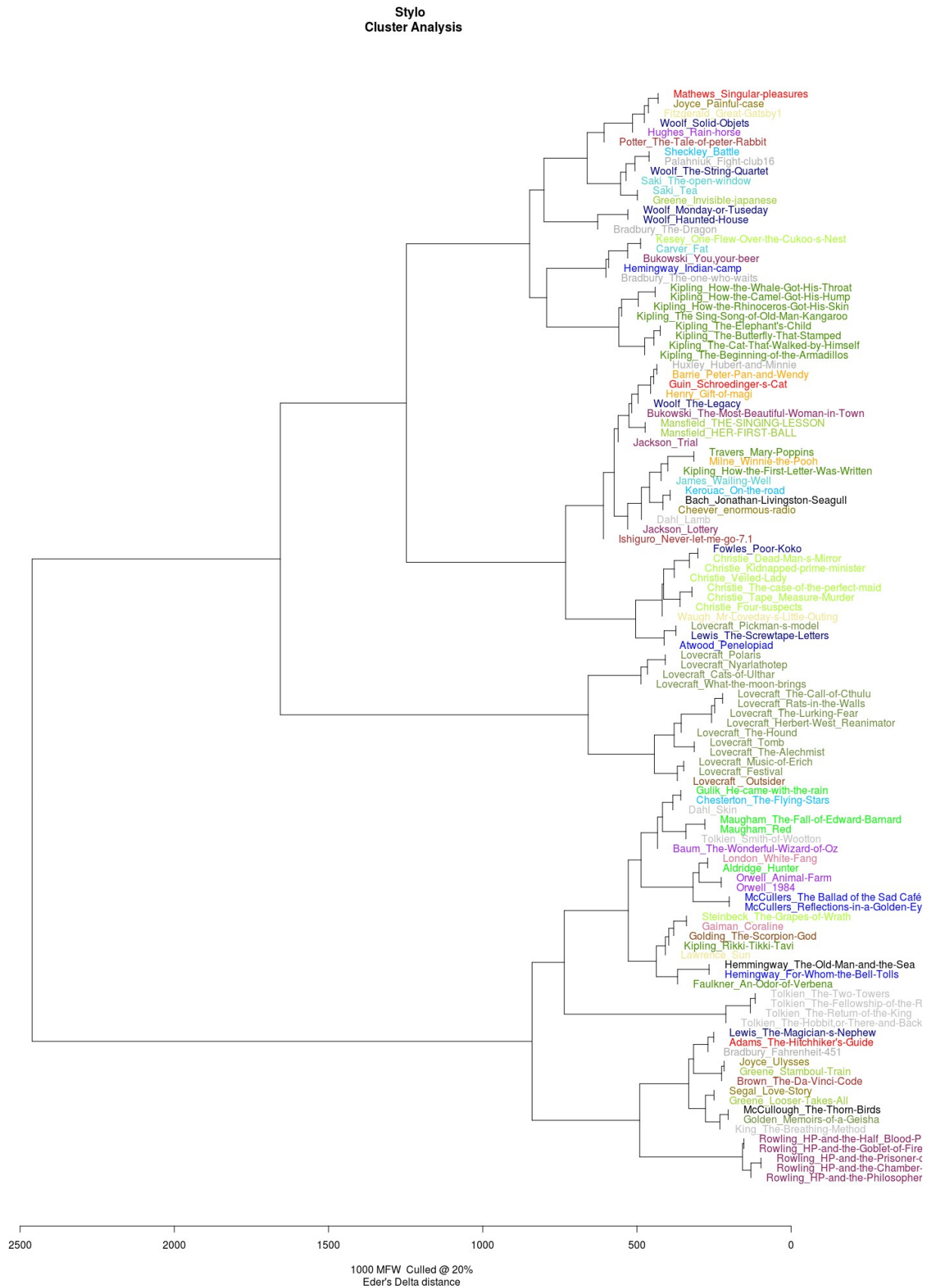


Illustration 3.10. Dendrogram of English texts in the EPB corpus (Eder's Delta measure).

As one can easily note, the scale on both diagrams is different and therefore all results for the tests with the Eder's Delta measure are normalised to the scale on the Classic Delta measure diagram, that is they are divided by 500. The most similar texts appear closest to the beginning of the scale, at the value of 0, and the most dissimilar around the value of 5. Analysis of the results obtained with two measures reveals no major differences in the composition of the clusters, only variation in the distance. The observations concerning English texts are as follows:

- (a) most of Lovecraft's short stories (14 out of 15) are grouped in two clusters connecting at the distance of 2.17 (Classic) or 1.03 (Eder's), one text however is an outlier, that is *Pickman's Model* joining the rest at 2.5 (Classic) or 3.03 (Eder's) distance (measured from the beginning of the scale)
- (b) most of Kipling's short stories (8 out of 10) are grouped in one cluster, but there are two outliers – *How the First Letter was Written* at 2.5 (both Classic and Eder's) common point with the main cluster, and *Rikki Tikki Tavi* at 5.08 (Classic) or 4.92 (Eder's) distance in a common point for all Kipling's texts
- (c) most of Tolkien's novels are clustered together with the exception of the *Smith of Wotton Major* connecting to the main cluster at 5.08 (Classic) or 1.68 (Eder's)
- (d) Woolf's short stories are spread across many clusters, all of which connect at 2.5 (Classic and Eder's) distance from the beginning of the scale
- (e) the rest of the texts are grouped predominantly according to their genre, that is short stories or novels and novellas, including a cluster of short stories by Agatha Christie and novels by J.K. Rowling.

These outcomes are interesting on their own but their deeper meaning can be revealed when contrasted with the results for the translations of these texts. Dendrograms for the Polish part of the EPB corpus (created with the same settings as described above) are as follows:

Stylo  
Cluster Analysis

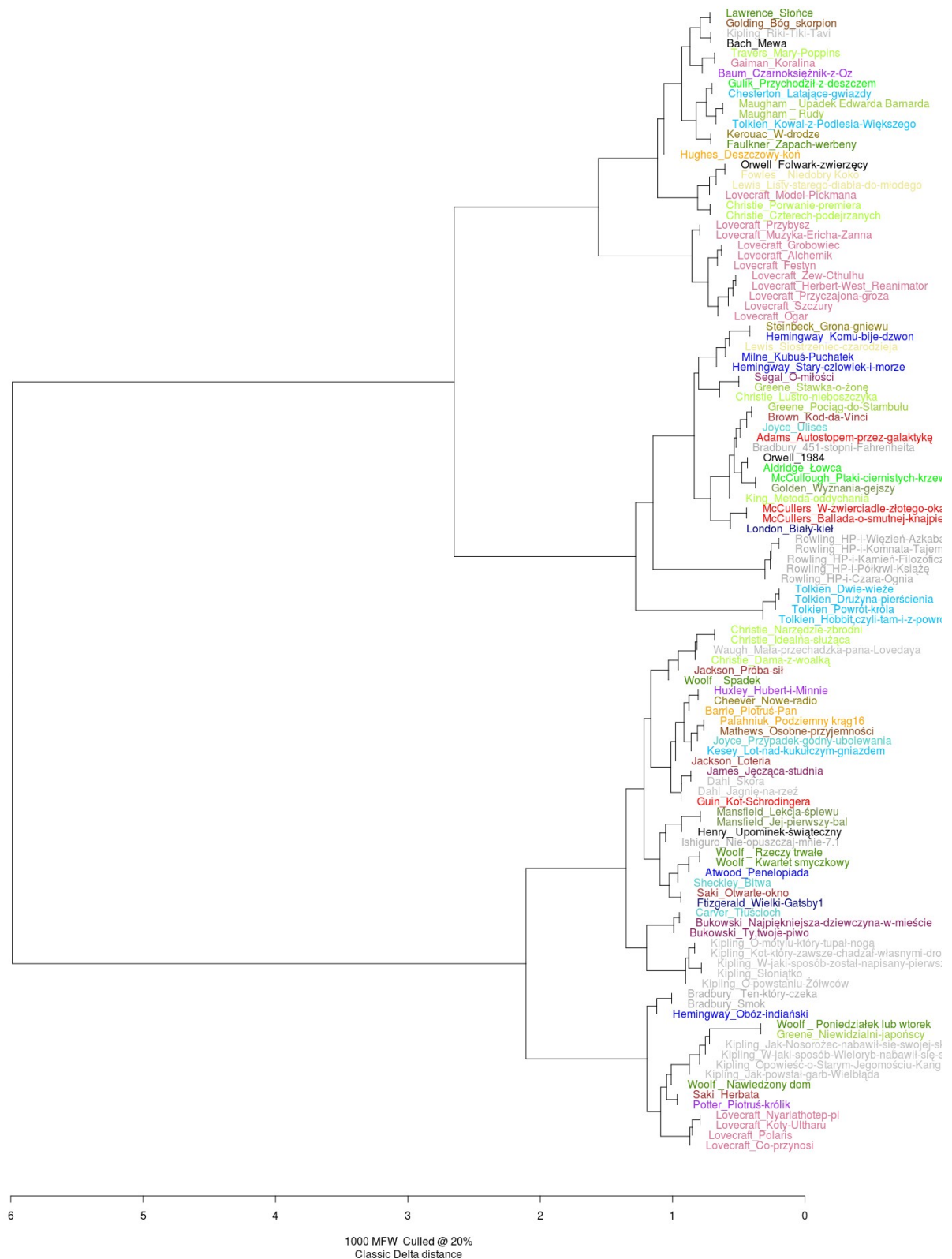


Illustration 3.11. Dendrogram of Polish texts in the EPB corpus (Classic Delta measure).

Stylo  
Cluster Analysis

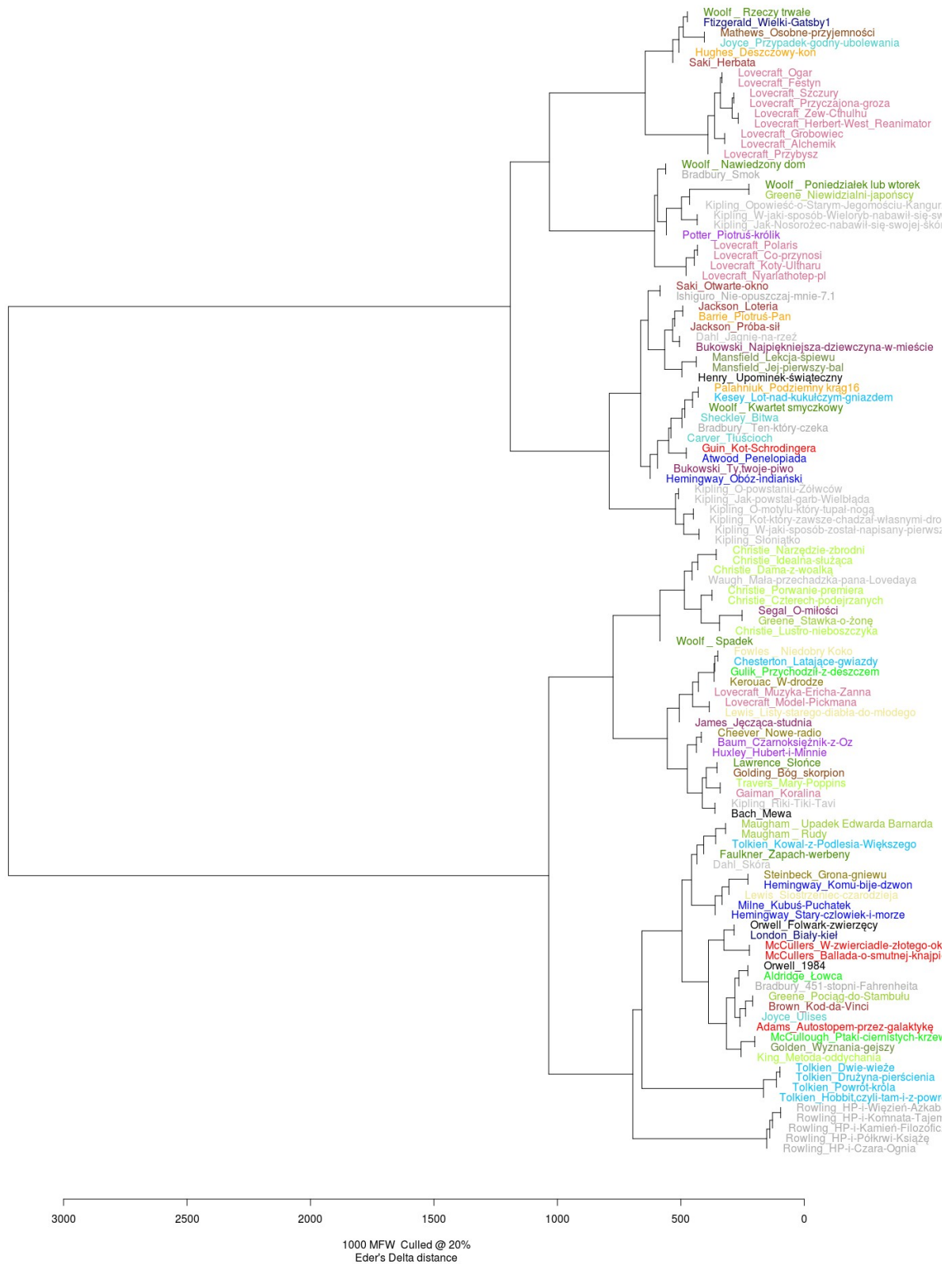


Illustration 3.12. Dendrogram of Polish texts in the EPB corpus (Eder's Delta measure).

Unlike in the case of English, using two different statistical measures for the Polish language produces more varied results. Let us consider clusters following the order from the results for the English texts:

- (a) Lovecraft's texts are divided in two clusters with one outlier – *Pickman's Model* connecting to one cluster at 1.6 (Classic measure) or at 6.5 (Eder's) together with *Music of Erich Zann*. The most prominent difference between the two dendrograms is however the distance between the two main clusters. In the case of Eder's Delta the clusters connect at 2.3 and in the case of Classic Measure it is as much as 6.0, which signifies a much bigger stylistic difference. However the higher value is assigned to the Classic Measure, which performs worse, and as the composition of both of the clusters is the same in both Polish dendrograms and English dendrograms, it might be concluded the two groups of texts are written differently and the translations simply convey these differences.
- (b) Short stories by Kipling, unlike their original versions, are grouped in two clusters connecting at 2.1 (Classic) or 2.2 (Eder's) distance. The exception is, as in the case of English, *Rikki Tikki Tavi* joining the main group at 6.0 (Classic) or 6.5 (Eder's) distance from the beginning of the scale.
- (c) Tolkien's novels are grouped in two clusters, one consisting solely of *The Smith of Wootton Major*, and the other of the rest of the texts. The two clusters connect at 2.7 (Classic) or 2.1 (Eder's) distance which is consistent with their position in English dendrograms.
- (d) Short stories by Woolf are scattered among one big cluster of mostly short stories, all at 2.1 (Classic) or 2.2 (Eder's) distance, which is consistent with the results obtained for the English versions.
- (e) Rowling's novels, as in the case of English, are grouped closely together and they belong to one big cluster of longer forms, such as novels and novellas.
- (f) Texts by Christie differ depending on the measure: in the case of the Eder's Delta, they belong to one cluster, however in the Classic Delta dendrogram they are in three clusters, two of them connecting at 2.7 and three of them at 6.0 distance; but it must be noticed that that result is obtained with the use of the measure that is less reliable for Polish, so this finding confirms the greater reliability of Eder's Delta for Polish.

Finally, the third set of dendrograms represents data from the Belarusian part of the EPB corpus:

Stylo  
Cluster Analysis

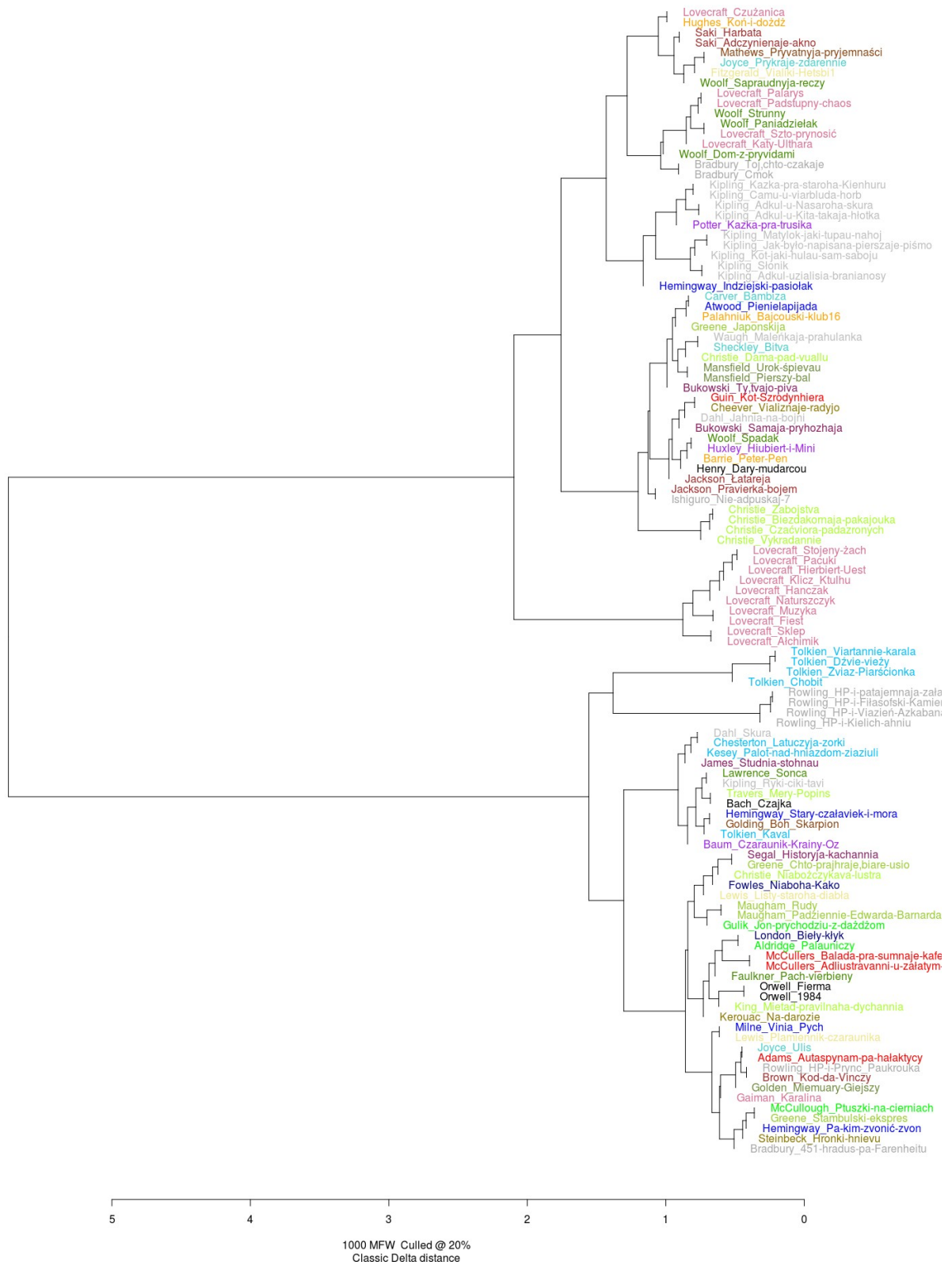


Illustration 3.13. Dendrogram of Belarusian texts in the EPB corpus (Classic Delta measure).

Stylo  
Cluster Analysis

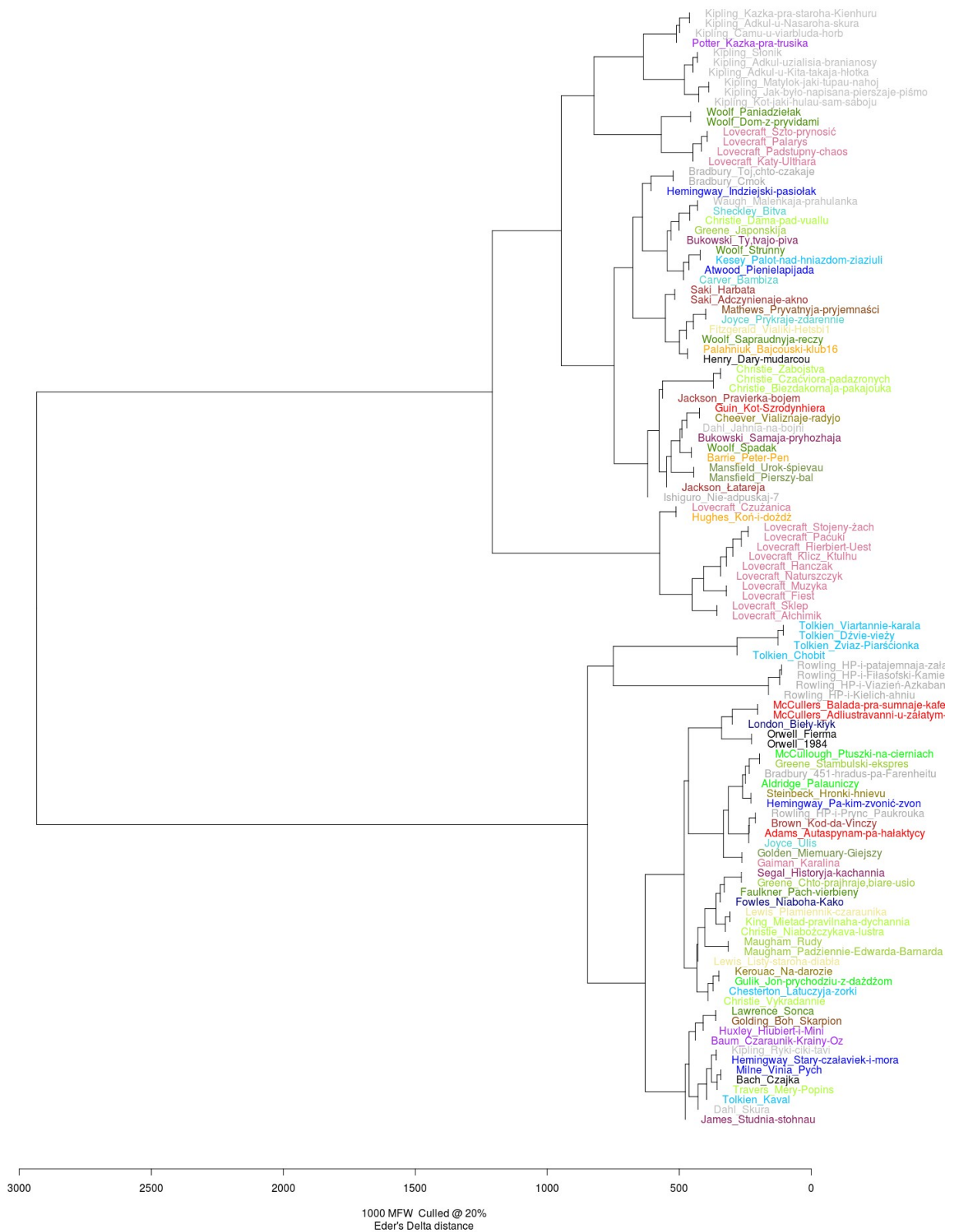


Illustration 3.14. Dendrogram of Belarusian texts in the EPB corpus (Eder's Delta measure).

The Belarusian dendrograms created based on the two statistical measures are the most unified of all languages.

- (a) Lovecraft's texts are divided in two groups the same way their Polish and English counterparts are. However the short story *Outsider* is singled out. It connects to the cluster at a fairly close distance and thus cannot be considered distinctive.
- (b) Short stories by Kipling are grouped in a cluster, with the exception of *Rikki Tikki Tavi* connecting to the main cluster at 5.75 (Classic Delta) or 5.87 (Eder's Delta) distance, making it very consistent with the results obtained for the English and Polish versions.
- (c) Tolkien's novels are grouped in one cluster and the novella *Smith of Wootton Major* connects to the rest at 1.5 (Classic) or 1.7 (Eder's) distance, which is almost exactly the same as in the English and Polish dendrograms.
- (d) Just as in the English and Polish dendrograms, texts by Woolf are scattered within one bigger cluster of short stories and the distance between them is 1.74 (Classic) or 2.45 (Eder's) which makes this result consistent across the three languages.
- (e) In the case of Harry Potter series, one book, *HP and the Half-Blood Prince*, is visibly different from the rest, although still close – at 1.5 (Classic) or 1.7 (Eder's) distance. This split is easily understandable as this is the only indirect translation in the collection of Rowling's novels. It was translated from Russian rather than English (to date there is no direct Belarusian translation of this particular part of the series).
- (f) Christie's texts are the only ones that show significant variations. Three short stories translated by one translator, Čudaŭ, are clustered together in both Classic and Eder's Delta dendrograms. One more short story by another translator lies close to the main cluster. The difference is in the short stories *The Kidnapped Prime Minister* and *Dead Man's Mirror*. In the case of the Classic Delta measure, the first story is grouped with the main cluster and the second one is at 5.75 distance, which is consistent with the results obtained for English. However, in the case of Eder's Delta the two texts are close together and connect to the rest at 5.9 distance, even though *The Kidnapped Prime Minister* was translated by Čudaŭ whose translations lie in the main cluster.

Generally speaking, it can be concluded that the dendrograms of Polish and Belarusian translations are fairly consistent with those of the original texts. This suggests that the author's style dominates over the style of the translators in the analysed sample. Apart from that, the overall analysis indicates directions for detailed investigation with the use of 'classical' corpus linguistics methods. To exemplify such an examination, the Belarusian translation of the *Kidnapped Prime Minister* is to be compared with the rest of the stories by Christie in the Belarusian corpus (except *Dead Man's Mirror* which is positioned far from the main cluster regardless of language and the statistical measure). The aim of this case study is to determine what features distinguish this particular text and might cause its isolated position in the Belarusian dendrogram based on the Eder's Delta measure.

In the first step of examining the texts in question a keyword list (mentioned in Chapter 1.5) has been applied. Keyword lists are usually used for determining the main topic of the text and they have been implemented in such a fashion in various studies, for example by Mahlberg and McIntyre (2011) to analyse *Casino Royale* by Ian Fleming. Other researchers applied keyword analysis in investigating characters' talk (Culpeper, 2009), in identifying creative language in translations (Kenny, 2001), or in comparing different translations in order to characterise translator style (Winters, 2009). These case studies prove the versatility of this particular technique.

SketchEngine enables the user to create a keyword list using as a reference corpus beTenTen (mentioned in Chapter 1.4). The programme creates a list ordered according to the score based on the simple maths method (Kilgarriff, 2009; Lexical Computing Ltd., 2015), which is basically comparing the normalised frequencies of particular words in two given corpora. Results obtained for the Belarusian version of *The Kidnapped Prime Minister* and other short stories by Christie contain among the top 50 words mostly proper nouns and titles, namely *lord*, *milord*, *mister*, *sir* in the case of the single text, and additionally *miss*, *missus*, *monsieur* and *lady* in the case of the rest of the stories. Obviously, the first set of words is exclusively male, while the second includes additionally female appellations. This pattern becomes more prominent when analysing content words. While nouns and adjectives relate directly to the events described in the stories (such as *chloroform* and *gag*), the verbs differ substantially between the two samples:

<i>The Kidnapped Prime Minister</i>				Other short stories			
Term	Score	Freq.	Ref. freq.	Term	Score	Freq.	Ref. freq.
усклікнуў (shouted, masculine)	364.470	5	89	кіўнула (noded, feminine)	205.300	5	31
паківаў (noded, masculine)	351.670	3	25	спыталася (asked, feminine)	161.280	7	118
кіўнуў (noded, masculine)	324.290	5	110	папыталася (asked to, feminine)	153.830	3	9
знаеце (you know, plural)	296.550	2	3	пастукала (knocked, feminine)	152.130	3	10
пакруціў (shook, masculine)	267.870	3	58	уздыхнула (sighed, feminine)	140.450	4	50
ўкралі (stole, plural)	258.530	3	63	бачыце (you see, plural)	117.910	7	191
пачуўся (was heard, masculine)	232.500	2	26	падазраяце (you suspect, plural)	110.240	2	3

Table 3.18. Highest scored verbs on the keyword lists of Christie's short stories in Belarusian.

Out of first seven verbs on the keyword lists from *The Kidnapped Prime Minister* five are past tense, singular, 3<sup>rd</sup> person and male; the two remaining (at 4<sup>th</sup> and 6<sup>th</sup> position) are plural, which makes them universal gender-wise. In the other short stories by Christie, five out of the seven first verbs are past tense, singular, 3<sup>rd</sup> person and female; the two remaining (at 6<sup>th</sup> and 7<sup>th</sup> position) are plural, similarly to the results for the text which was singled out in the stylometric analysis. These two regularities visible within the first hundred keywords identified for the two samples might indicate that the reason for the stylistic differences between *The Kidnapped Prime Minister* and the rest of Christie's texts arises from the disproportion between male and female figures in both groups. This hypothesis can be effectively confirmed by close-reading of the concordances of the verbs presented in Table 3.14. It is also visible in the plot of the stories – in *The Kidnapped Prime Minister* there are solely male figures talking about other males, while in the other short stories most of the main and side characters are female. Moreover two out of four investigations are led by Miss Marple rather than Poirot.

Another level of analysing translation and translator style are n-grams, that is word sequences. They have been applied in text analysis (as indicated in Chapter 1.5) for various purposes, such as identifying sequences used to characterise a novel's protagonist (Fischer-Starcke, 2010) or for determining major themes in the text (Mahlberg, 2013). Based on the same principle as keywords analysis, clusters can be investigated in comparison to another text or corpus. Such a 'key cluster analysis' is implemented by Mastropierro (2018) in the study of Italian translations of Lovecraft's short story. The reason behind using clusters rather than individual words is that n-grams can affect the meaning to a higher degree.

As with the keywords, SketchEngine was implemented in creating key cluster lists – one in comparison to beTenTen corpus, and another one in comparison to the rest of Christie's short stories translated into Belarusian. Following Mastropierro's argument, that "having a larger pool of items to compare increases the chances of obtaining more relevant data for the analysis" (2018, p. 246), the widest possible span was chosen, that is 2-word to 6-word clusters. The minimum frequency of each n-gram was set to five. The result of the query is 25 key clusters when compared to other short stories and 26 results when compared to beTenTen corpus:

reference corpus – Christie's other short stories						reference corpus – beTenTen					
Keyword	Freq	Rel.	Freq. ref.	Rel. ref.	Score	Keyword	Freq	Rel.	Freq. ref.	Rel. ref.	Score
ў Францыі (in France)	6	919.800	0.000	0.000	920.800	са Скотланд-Ярда (from Scotland Yard)	6	919.800	0.000	0.000	920.800
са Скотланд-Ярда (from Scotland Yard)	6	919.800	0.000	0.000	920.800	містэр Додж (mister Dodge)	5	766.500	0.000	0.000	767.500
яго ў (him in)	5	766.500	0.000	0.000	767.500	другі аўтамабіль ([the] other car)	5	766.500	0.000	0.000	767.500
у Францыю (to France)	5	766.500	0.000	0.000	767.500	мой друг (my friend)	5	766.500	5.000	0.100	722.600
містэр Додж (mister Dodge)	5	766.500	0.000	0.000	767.500	з галоўнай (with [the/a] main)	5	766.500	120.000	1.500	308.100
мой сябар (my friend)	5	766.500	0.000	0.000	767.500	мой сябар (my friend)	5	766.500	125.000	1.600	300.600
мой друг (my friend)	5	766.500	0.000	0.000	767.500	у Францыю (to France)	5	766.500	145.000	1.800	274.000
з галоўнай (with [the/a] main)	5	766.500	0.000	0.000	767.500	паглядзеў на ([he] looked at)	6	919.800	309.000	3.800	190.300
другі аўтамабіль ([the] other car)	5	766.500	0.000	0.000	767.500	ён павінен (he should)	5	766.500	273.000	3.400	174.800
ён павінен (he should)	5	766.500	0.000	0.000	767.500	некалькі хвілін (few moments)	5	766.500	614.000	7.600	88.900
ён не (he + 'no' particle)	5	766.500	2.000	113.400	6.700	ў Францыі (in France)	6	919.800	1070.000	13.300	64.400
некалькі хвілін (few moments)	5	766.500	2.000	113.400	6.700	як вы (as you [plural])	5	766.500	1110.000	13.800	51.900
калі ён (when he)	5	766.500	2.000	113.400	6.700	але я (but I)	5	766.500	2152.000	26.700	27.700
як вы (as you [plural])	5	766.500	3.000	170.000	4.500	што яго (that him)	6	919.800	3287.000	40.800	22.000
што яго (that him)	6	919.800	4.000	226.700	4.000	калі ён (when he)	5	766.500	3769.000	46.800	16.000
і я (and I)	5	766.500	4.000	226.700	3.400	і што (and that)	5	766.500	4992.000	62.000	12.200
паглядзеў на ([he] looked at)	6	919.800	5.000	283.400	3.200	яго ў (him in)	5	766.500	5273.000	65.500	11.500
на яго (at him/his)	7	1073.100	6.000	340.100	3.100	і я (and I)	5	766.500	5493.000	68.300	11.100
і што (and that)	5	766.500	5.000	283.400	2.700	нічога не (nothing + particle 'no')	8	1226.400	8862.000	110.100	11.000
але я (but I)	5	766.500	6.000	340.100	2.300	на яго (at him/his)	7	1073.100	8042.000	99.900	10.600
ніхто не (nobody + particle 'no')	5	766.500	7.000	396.700	1.900	я не (I + particle 'no')	8	1226.400	9390.000	116.700	10.400
што ён (that he)	8	1226.400	12.000	680.100	1.800	ён не (he + 'no' particle)	5	766.500	6314.000	78.500	9.700
я не (I + particle 'no')	8	1226.400	16.000	906.800	1.400	ніхто не (nobody + particle 'no')	5	766.500	6948.000	86.300	8.800
нічога не (nothing + particle 'no')	8	1226.400	17.000	963.500	1.300	што ён (that he)	8	1226.400	12071.000	150.000	8.100
						і не (and + particle 'no')	5	766.500	34692.000	431.100	1.800

Table 3.19. Key n-grams in Belarusian translation of Christie's 'The Kidnapped Prime Minister'.

The term 'Keyword' refers to a key n-gram in this case, 'Freq' and 'Freq. ref.' are real frequencies in the focus corpus and in the reference corpus respectively, 'Rel.' and 'Rel. ref.' are relative frequencies in the focus corpus and in the reference corpus respectively. Relative frequency is calculated in the following manner – the number of hits is multiplied by a million and then divided by the corpus size. Thus the smaller the corpus the higher the relative frequency. Finally, score, as in the case of keywords, is calculated using the simple maths procedure (discussed above).

The two lists are very similar. The difference lies in just one phrase which suggests that the Belarusian translation of *The Kidnapped Prime Minister* stands out not only compared to other short stories but also in terms of general linguistic characteristics. As in the keyword list, some key

clusters indicate a specific topic – there are mentions of Scotland Yard (*са Скотланд-Ярда*) and of France (*ў Францыі, у Францыю*) – and give some hints about the characters, such as mister Dodge (*містэр Додж*) and my friend (*мой сябар, мой друг*). However, there are many more clusters consisting of function, rather than content words, and they point to some interesting discourse features of this short story.

Firstly, there are six key clusters suggesting a male character being an object of conversation. The centre of these n-grams is the word *he* in various declension forms (*яго ў, ён не, калі ён, што яго, на яго, што ён*). Additionally one of the clusters containing content words – *ён павінен* (*he should*) – points to the same fact. Secondly, first-person narrative is significantly more common in this text; it is indicated by clusters containing the pronoun *I* (*і я, але я, я не*). Finally, n-grams including negation, that is the particle *не*, are singled out as key clusters (*ён не, ніхто не, я не, нічога не*).

All in all, the keywords and key clusters analysis indicates that the Belarusian translation of Christie's *The Kidnapped Prime Minister* is distinct not only from translations of other short stories, but it also contains linguistic features that make it stand out in the stylometric analysis conducted at the beginning of this chapter. The investigated text contains significantly more verbs in male grammatical gender, it frequently treats male characters as objects rather than subjects, and it includes a significant amount of first person discourse as well as negation.

Before moving on to the next chapter an additional comment must be made about the specific sociolinguistic features of the Belarusian corpus versus the Polish corpus (within the EPB) and about how these feature relate to the translator versus author style. The correlation between social factors and the way individuals use the language is undeniable. Specifically gender has been pinpointed as an important factor in linguistic attitudes differentiation. Researchers identified a handful of characteristics allowing to recognise the author's gender (Simaki et al., 2017). It is therefore justified to assume that the sociolinguistic features of each corpus comprising the EPB corpus would have an impact on the collective analysis of the data. Among the EPB metadata gathered for the purposes of this project there are two important social categories that should be discussed in more depth: age and gender.

Chapter 3.1. discusses a series of tests conducted to find possible relation between the relative sizes of the target texts and the metadata gathered alongside the EPB corpus. However, there might be connections between these metadata and other properties of the text, such as syntactic complexity or the vocabulary richness which do influence the stylometric analysis, but have not been taken into account in previous chapters. The sociolinguistic composition of particular parts of the EPB corpus differs and can influence the analysis when examined collectively. Firstly, the proportion of male versus female authors among English creators and Polish and Belarusian translators is quite different:

language	male authors	female authors
English	46 (81%)	11 (19%)
Polish	30 (54%)	26 (46%)
Belarusian	42 (68%)	20 (32%)

*Table 3.20: Gender composition of texts authors (EPB corpus).*

As evident from table 3.20 male authors dominate the English texts, the proportion of male to female creators is about 80/20, where as in the Belarusian part the disproportion reduces to about 70/30 and it almost evens out in the Polish part of the EPB corpus. Similarly, differences are observed when comparing the age of the Polish and Belarusian translators. Polish are on average over 83 years old, where as the mean age of Belarusian translator is slightly over 57. Both groups of translators are also, on average, younger than the authors of English originals. Therefore the sociopolitical and cultural discrepancies between generations, as well as different proportion of male and female authors, might account for some differences between the translations and the originals, not only in the stylometric context but also in the way the discourse is conveyed, as discussed more broadly in Chapter 4.

In summary, this chapter tackled the topic of stylometry and its use in translation studies exemplified on the EPB corpus data. The analysis was conducted on two levels. Firstly, the package Stylo was applied to investigate all texts collectively via dendrograms, that is visualisations indicating stylistic distance between particular titles. To achieve the most reliable results, two statistical measures were used – Classic Delta, which works better with the English language, and Eder’s Delta which is designed for languages such as Polish and Belarusian. With few exceptions, samples from the English part of the corpus were divided by the Stylo algorithm according to the type of the text (short forms, that is short stories, vs. long forms, such as novels) and according to the authorship.

The dendrograms for Polish and Belarusian translations appeared to show very similar features. Among the few exceptions was the Belarusian translation of *The Kidnapped Prime Minister* short story. In the second part of this chapter this particular text has been studied to determine possible reasons for the discrepancies. By applying keywords and key clusters analysis both to the source text and to the translations the following distinguishing features were identified: use of negation, first-person discourse, use of masculine verbs and putting male characters in the position of the object of discourse. These features are unusual in Christie’s writing style (Tsuchimura, 2016) and are a reminder that her writing is more diverse than often thought. In these cases stylistic features of the original were realistically reproduced in the translations, indeed, exaggerated.

### **3.3. Applied translation studies: how the corpus can improve the training of translators dealing with the Belarusian and Polish languages**

Parallel corpora have proved to be very useful in applied translation studies, e.g. in creating exercises for translation students (as mentioned in Section 1.3). The main advantage of parallel corpora is that they present native-like solutions in both source and target languages aligned, and therefore they can speed up the process of and enhance the quality of translation. With systematic use of parallel corpora, translators, especially inexperienced ones, can increase their awareness of general patterns functioning in a given language. In other words, parallel resources provide information about norms in two languages combined, therefore enabling understanding of these norms. This is not entirely possible in the case of using exclusively dictionaries, as they are limited in volume and scope. Similarly, text reading alone does not give a fair overview of most preferred patterns, as in this case they might be obscured by biases or by too little data that humans can take in. Therefore extracting and analysing phraseology in parallel datasets is an invaluable part of a translator's training.

This topic is especially interesting in the context of Belarusian, as phraseological dictionaries including this language are very scarce. In fact, there is only one such resource connecting Belarusian either with English (Artsiomava, 2014) or with Polish (Aksamitaŭ, 2000). Resources available for Polish are, of course, much more numerous, and include both general (Arabski, 2010; Jaworska, 2002; Kakietek, 2004) and specialised (Barycka et al., 1997; Domański, 1992) language, yet the newest dictionaries contain only the material in Polish and its English translations (opposite to what the EPB corpus offers) and due to printed sources' properties, they present the most popular examples, dismissing rare cases which are often challenging for literary translators.

Judging from the above-mentioned circumstances it can be concluded that phraseology is an appropriate starting point for showcasing how the EPB corpus can improve the training of translators dealing with the Polish and Belarusian languages. In order to conduct a phraseological study this chapter will be referring to a subcorpus of the EPB corpus of 1.5 million words (approx. 500 thousand in each of the three languages) aligned at the level of sentences with the use of the Mantel application (discussed in Section 2.4). The infrastructure utilised for aligning the data in the EPB corpus was the same as in the case of the Paralela Corpus (mentioned in Section 1.4 and 2.1) and the output is stored a database of an identical type, therefore both corpora can be searched using the same query syntax.

SlopeQ, as the above-mentioned query syntax is called, enables searching for single words, loosely defined phrases and lexico-grammatical patterns. According to Pezik, this “query syntax is expressive enough [...] to facilitate the investigation of subtle bilingual phenomena such as the idiomacity of translation and the incidence of phraseological equivalence” (2016, p. 75). Phraseological equivalence (PE) occurs in a case when translators consistently favour one target language unit over another in translating the corresponding source language phraseological unit. The existence of a preferred equivalent can be determined via analysis of the frequency of particular translation choices. According to Pezik, a low level of PE may have negative implications, namely more cognitive effort on the part of the reader and more ambiguity. On the other hand, it can

constitute a distinct translator's style and determine the unique form of the target text. Despite the ultimate choice of the translator, the key to success is the awareness of possible solutions and the possible consequences of particular choices. The EPB corpus and its infrastructure enables the data to be investigated in multiple ways in order to achieve that awareness and facilitate translator training.

With regard to the available methods of teaching translation, the most obvious procedure is to use the parallel corpus to look up unknown phraseologisms or check those whose translation is ambiguous due to the context or due to the phrase idiomacity where no established analogue exists. An example of an English idiomatic phrase which has no equivalent in Polish or Belarusian is *neck of the woods*. According to Merriam-Webster dictionary it means simply "the place or area where someone lives" (2019). In the examples from the EPB corpus various ways to render the English phrase occur:

English occurrences	Polish translations	Belarusian translations
"Since when," continued his murine colleague, "we have had an offer of a quite enormously fat contract to do the 5D chat show and lecture circuit back in our own dimensional <b>neck of the woods</b> , and we're very much inclined to take it." ( <i>Hitchhiker's Guide to the Galaxy</i> )	— No i wtedy — ciągnął jego myśl towarzysz otrzymaliśmy naprawdę niesamowicie korzystną propozycję kontraktu na wywiad w 5D i cykl wykładów w naszych <b>okolicach</b> (eng. area, neighbourhood) i mamy wielką ochotę ją przyjąć.	- І стуль, - працягваў яго пацукаваты сябрук, - мы атрымалі проста раскошную прапанову зрабіць у пяцівымеры аўтарскае ток-шоў і медыя-турнэ па нашым родным <b>вымярэнні</b> (eng. dimension). Няма чаго казаць, тлустая прапанова, ад якой немагчыма адмовіцца.
So as the hour grew dangerously near to dawn, we did as we had done with the others—dragged the thing across the meadows to the <b>neck of the woods</b> near the potter's field, and buried it there in the best sort of grave the frozen ground would furnish. ( <i>Herbert West – Reanimator</i> )	Gdy świt był już niepokojąco blisko, przenieśliśmy ciało przez pola, na <b>skraj lasu</b> (eng. the edge of the forest) opodal cmentarza i pogrzebaliśmy tam, w najlepszym grobie jaki byliśmy w stanie wykopać w zmarzniętej ziemi.	Тым часам звонку пачынала днець, і мы, не хочучы болей выпрабоўваць лёс, зрабілі тое, што рабілі з усімі папярэднімі экзэмплярамі, — адцягнулі паддоследнага ў <b>лес</b> (eng. forest) каля могілак і пахавалі ў яме, спехам выкапанай у мёрзлай зямлі.
"So, what brings you to this <b>neck of the woods</b> , Minister?" came Madam Rosmerta's voice. ( <i>Harry Potter and the Prisoner of Azkaban</i> )	— Więc co sprowadza pana aż na sam <b>skraj puszczy</b> (the edge of the forest), panie ministrze? — zabrzmiał głos madame Rosmerty.	- І які лёс прынёс вас да нашых <b>глухмяняў</b> (eng. wilderness), міністр?- спытаўся голас мадам Размерты.

Table 3.21. Occurrences of 'neck in the woods' phrase in the EPB corpus.

In the case of Polish translations the first example uses simple *okolica* (which can be back-translated also as *neighbourhood* or *area*). Next two, *skraj lasu* and *skraj puszczy* literally mean *the edge of the forest*, where *puszcza* is a type of forest which is much more dense, older and usually bigger. In both cases other than *okolica*, the translation of the phrase *neck of the woods* was clearly influenced by the context. *Herbert West* is a story with a spooky atmosphere and the most fitting location for an unofficial grave seems to be the edge of the forest than simply to the neighbourhood of the cemetery. In *Harry Potter* series the pub run by Madam Rosmerta is located

on the edge of the Forbidden Forest, translated into Polish as *Czarna Puszcza* (literally, *black forest*).

There is even more variation in the Belarusian translations of this fragment. In the first excerpt a word *вымярэнне* (dimension) was used and it refers to the adjective *dimensional* added to the phrase. In this case the context has been actually broadened as the phrase indicating one particular area (*our own dimensional neck of the woods*) was translated as *our own dimension*. The translation of the second excerpt, similarly to the Polish version, uses the word *forest*, which, again, broadens the context. In the excerpt from the Harry Potter series the phrase is translated with the word *глухмень*, which is a colloquial term for wilderness. It is a reference to the remote location of Hogwarts, which is underlined in the series on many occasions.

This simple exercise gives a translator not only an overview of possible translation solutions but also the insight into a specific context in which each translation appeared, because the output of the query is the whole segment (in this case – a sentence) in which the phrase occurred. This variety in a translation shows that embedding a text's context in a translation of a phrase can potentially make it more attractive to the reader. Practising translators do not necessarily keep only to standard solutions, but more often try to incorporate variation which is also associated with a translator's style, and their understanding of narrative context.

Next, as an example of a rare English phraseologism, namely *donkey's years (ago)*, is searched. It occurs only 20 times in the BNC corpus and 183 in enTenTen corpus. However its explication can be found in Merriam-Webster dictionary and such an explanation is already a good hint for a translator. What is much better though is an insight into an actual translation solution already utilised in real life. These can be found in the EPB corpus and are presented in the table below:

English	Polish	Belarusian
[...] my mother taught me <b>donkey's years ago</b> not to automatically believe people who tell you glibly to 'drop by anytime' [...] (King, <i>The Breathing Method</i> )	[...] matka <b>za młodu</b> (eng. when I was young) nauczyła mnie nie wierzyć ludziom, którzy gładko mówią "wpadaj, kiedy tylko chcesz" [...]	[...] маці <b>даўным-даўно</b> (eng. a long time ago) навучыла мяне ніколі не верыць абсалютна ўсяму, што кажуць людзі, асабліва, калі гэта фраза нахшталт "заходзь у любы час" [...]
I recollect oh, <b>donkey's years ago</b> -- I used to sometimes go to 'Yde Park of a Sunday afternoon [...] (Orwell, 1983)	Pamiętam, jak <b>wiele, wiele lat temu</b> (eng. many, many years ago) zaglądałem czasem do Hyde Parku w niedzielne popołudnia [...]	Памятаю, як... <b>колькі ж гэта гадоў</b> (eng. how many years ago) ужо? а, не ведаю. Часам я хадзіў у Гайд-парк вечарамі ў нядзелю [...]

Table 3.22. Occurrences of 'donkey's years' phrase in the EPB corpus.

There is no guarantee that these particular translations are the best possible but if they are not, they are a perfect starting point for further enhancement. A student may observe, apart from the actual translation solutions, also the techniques used for conveying particular information. The essence of the meaning of *donkey's years* is the emphasis on the length of a period of time and this is what all

the examples from the table followed in three different ways. Firstly, the Polish translation *za młodu* (*in my youth*) uses a reference to an early period of life that in the case of any adult character has obviously ended a long time ago. Secondly, the Polish translation *wiele, wiele lat temu* (*many, many years ago*) and the Belarusian *даўным-даўно* (*once upon a time*) use repetition for the same purpose. The first example is the case of a simple repetition of an adverb, however the other phrase is an instance of a phenomenon typical for Belarusian only, that is tautology as emphasis (Barszczewska, 2004, pp. 37–38). *Даўно* is an adverb meaning *long time ago*, while *даўным* is an adjective formed from the adverb, therefore it is not a simple repetition but the repetition of a common core in the form of two different parts of speech. Finally, the example *колькі ж гэта задоў* (*so many years ago*) showcases another feature characteristic for Belarusian language (which is also present in Polish, although not as frequently used) that is the usage of the particle *ж* for underlining the preceding word.

A short analysis of the examples available in the corpus enables the user not only to find ready to use translation solutions but also to learn overarching rules that the translators followed. This in turn allows the corpus user to come up with a new, distinct way of conveying a particular unit in the target language.

The procedure illustrated in the previous paragraphs can be used in investigating phraseological equivalence. PE, mentioned at the beginning of this chapter, can be argued to be a type of lexical equivalence. It requires special attention due to the fact that some phrases, even though conventionalised in language use, might be difficult to recognised as such. Systemic use of the open and restricted collocations is much more subtle than in the case of idioms or proverbs. Therefore the correct identification of a source phraseological unit is as important as finding the right target language equivalent. Unlike terminological equivalence, PE may take the shape of one-to-many, many-to-one, many-to-many or even null relation. It does not always depend on the existence of a target phraseological unit. In many cases the choice of the translation is motivated by the context. A parallel corpus is a perfect tool for determining such factors.

To showcase the corpus application in investigating PE, the phrase *bad luck* and its translations are analysed. The phrase itself occurs 29 times in 16 English-language samples from the EPB corpus. The table below presents the summary of the Polish and Belarusian translations found in the data:

Technique used	Polish translation	Number of occurrences	Belarusian translation	Number of occurrences
dictionary equivalent and its synonyms	pech (noun) /pechowo (adjective)	11	няшчасце	5
	nieszczęście	11	няўдача	2
	niepowodzenie	3	няшчасці і нягоды	1
			нешанцаванне	1
			бяда	1

Technique used	Polish translation	Number of occurrences	Belarusian translation	Number of occurrences
to wish (one) ill	niech (coś) szlag trafi	1	хай спруціць	1
other terms with similar meaning	licho, marny los	2	ліха, цяжкія удары, злавесны, горай/горш, шкада, трасца	10
different context	krak	1	крумкач	1
negation of the opposite term			не пашанцавала, не пашчасціла, не павезла, да дабра не давядуць, ўдача павярнулася тварам да нас	5
irony			поспехы	1
omission			—	1

Table 3.23. Translations of 'bad luck' in the EPB corpus.

There is a clear-cut difference between the results in the two languages. Polish translators clearly prefer two equivalents that can be found in the dictionaries and which can be easily back-translated as *bad luck*. These two equivalents occur in 22 out of 29 cases, in other words they comprise 75% of all Polish translations of the analysed phrase. Even though there is no cut-off point for determining the existence of PE, the regularity visible in this data suggests that one can argue for the PE in the case of Polish translations of *bad luck*.

The Belarusian translations, on the other hand, show no prevailing trend. The applied solutions are extremely varied – there are 21 distinct translations of *bad luck*, as opposed to only 8 in the Polish part of the EPB corpus. There are also more techniques used in translating this phrase, most notably using negation of a term of the opposite meaning, that is instead of describing someone as *having bad luck* the translators frequently mention a character who *didn't get lucky*. These results definitely prove that in Belarusian there is no PE for the phrase *bad luck*.

Another interesting aspect of training translators are phrasal verbs. They are very frequent in the English language; some sources indicate that phrasals comprise as much as 80% of all English verbs (Language Over Internet, LLC, 2019). Their sheer quantity combined with the multitude of meanings is often overwhelming for English language learners at all stages. Studies suggest that even trained interpreters refrain from using phrasal verbs and find it difficult to access them in situations of high cognitive effort, such as simultaneous interpreting (Cresswell, 2018). All these factors prove the importance of including phrasal verbs as a vital element of training future translators.

Parallel corpora have been proven to be an efficient tool in exercising translation, e.g. into Malay (Awal et al., 2014). In Malay, similarly to Polish and Belarusian, most of the phrasal verbs are translated as a single word. In rare cases they contain a common core and differ in prefixes which in turn can be associated by the students with the corresponding prepositions. Nevertheless, in most cases the meanings of each phrasal verb need to be learned separately, taking into account idiomatic occurrences. Parallel corpora are the perfect source of possible translation solutions in context.

In the above-mentioned study researchers chose the verb *set out* as the object of the experiment. According to the Merriam-Webster Dictionary it has four distinct meanings:

transitive verb

1a: to arrange and present graphically or systematically

b: to mark out (something, such as a design): lay out the plan of

2: to state, describe, or recite at length

3: to begin with a definite purpose: intend, undertake

intransitive verb

4: to start out on a course, a journey, or a career (Merriam-Webster Inc., 2020)

This phrasal verb occurs 123 times in 28 samples from the EPB corpus, and it appears 6 times in the aligned subcorpus, all of the occurrences available in the aligned trilingual corpus are presented in table 3.24:

No	English	Polish	Belarusian
1	Then he lost everything in the War like everybody else, all hope of descendants too (his son killed his daughter's fiancé on the eve of the wedding and vanished) yet he came back home and <b>set out</b> singlehanded to rebuild his plantation. (Faulkner, <i>An Odor of Verbena</i> )	A potem stracił wszystko jak inni podczas wojny, a także stracił nadzieję, jakie pokładał w dzieciach (syn jego zabił narzeczonego siostry w przeddzień ślubu i zniknął), a jednak wrócił i <b>zabrał się</b> sam jeden <b>do</b> podnoszenia z ruin swojej plantacji.	Потым, як і ўсе іншыя, ён усё страціў падчас вайны, у тым ліку і адзінага сына (той знік, застрэліўшы жаніха сваёй сястры напярэдадні вяселля), аднак Сатпэн вярнуўся на сваю сядзібу і адзін <b>пачаў</b> аднаўляць разбуранае.
2	He had no friends to borrow from and he had nobody to leave it to and he was past sixty years old, yet he <b>set out</b> to rebuild his place like it used to be [...] (Faulkner, <i>An Odor of Verbena</i> )	Nie miał przyjaciół, od których mógłby pożyczyć, i nie miał komu tej plantacji po sobie zostawić, przekroczył już sześćdziesiątkę, a jednak <b>zabrał się do</b> przywracania swej posiadłości do dawnego stanu.	Сяброў у яго не было, і пазычыць грошы ні ў кога ён не мог, спадкаемца, каму пакінуць зямлю, у яю таксама не было, Сатпэну ішоў ужо семы дзесятак, і ўсё-такі ён <b>узяўся за</b> адраджэнне плантацыі.
3	What <b>had set out</b> as a walk along pleasantly-remembered tarmac lanes had turned dreamily by gate and path and hedge-gap into a cross-ploughland trek [...] (Hughes, <i>The Rain Horse</i> )	Spacer, <b>zaplanowany</b> po wygodnych, pokrytych makademem drogach, zamienił się jak we śnie, po przejściu przez bramę, ścieżkę i dziurę w ogrodzeniu, w wędrówkę na przełaj po zoranyrn polu [...]	Пасля таго, як ён выйшаў за браму на суязынку і пралез праз дзірку ў агароджы, тое, што было <b>задумана</b> як вандроўка ўздоўж багатага прыемнымі ўспамінамі гасцінца, непрыкметна ператварылася ў пераадоленне ўзараных палеткаў.

No	English	Polish	Belarusian
4	One morning they <b>set out</b> , the pair of them, with the few things that belonged to the girl, and walked along a grassy path under the coconuts, till they came to the creek you see. (Maugham, <i>Red</i> )	Dość, że jednego dnia rano <b>opuścili chatę</b> zabrawszy z sobą tłumoczek rzeczy stanowiących własność dziewczyny. Szli zieloną ścieżką pod palmami kokosowymi, aż znaleźli się tutaj, nad tą rzeczką, którą pan widzi.	[...] так ці іначай, неўзабаве яны сабралі немудрагелістыя пажыткі дзяўчыны і пайшлі па зарослай травой сцежцы пад какосавымі пальмамі да рэчкі, якую вы цяпер бачыце.
5	Private Williams <b>set out</b> for this assignment at about seven thirty in the morning. (McCullers, <i>Reflections in a Golden Eye</i> )	Szeregowy Williams <b>wyruszył</b> około wpół do ósmej rano.	Уільямс <b>выправіўся</b> выконваць заданне каля паловы восьмай раніцы.
6	[...] and before he could say knife they had whisked the trays and a couple of small tables into the parlour and <b>set out</b> everything afresh. (Tolkien, <i>The Hobbit, or There and Back Again</i> )	Nim Bilbo zdążył bodaj kichnąć, chwycili tacę i kilka małych stolików, zanieśli do sali i <b>zmienili</b> wszystkie <b>nakrycia</b> .	[...] і не паспеў Більба слова вымавіць, як яны падхапілі ўсе падносы і пару невялікіх столікаў у залю і <b>накрылі</b> зноўку.

Table 3.24. Examples of the 'set out' verb usage and their translations in the EPB corpus.

These several examples illustrate three out of four meanings of the verb *set out*, and even in such a small sample we can see five different translations into Polish and five into Belarusian, and based on them a student may learn not only possible translation choices but also features of phrasal verbs when translated into the two languages.

First of all, even though Polish and Belarusian usually require just one word translation, in some cases verbs with prepositions are used, such as *zabrać się do* and *узяцца за*. Both of these are reflexive verbs and the prepositions clearly indicate that the subject is carrying out an action directed at an object rather than the subject itself. This type of translation is used regularly in the case of the third meaning of the verb, in which *set out* is followed by the infinitive with *to*. Secondly, in the case of some translation choices it might be necessary to use a verb and a complementing POS in order to convey the full context, as in *opuścili chatę* ([they] left [the] hut). In this particular case it was equally correct to use the verb *wyruszyć*, as in the next example. However the translator decided to underline the aspect of finishing a certain stage in life, rather than starting a new one, and the choice of the verb imposed using also the noun, thus making the translation a two-word construction. In the example four in Belarusian the verb is omitted entirely.

The phrasal verb can be investigated also via n-grams analysis (discussed in Chapter 1.4). One of the verbs that are very productive in terms of phrasals is *to take*. There are 16 common phrasal verbs with *take*, and querying the EPB corpus for bigrams containing *take*, appearing in at least 10 samples and of minimum frequency of 10 returns 30 results, 7 of them being phrasal verbs: *take off*, *take out*, *take to*, *take up*, *take on*, *take in* and *take over*. Among those examples one might be especially interesting, that is *take in* which, according to Merriam-Webster Dictionary, has as many as 8 meanings. Repeating the procedure applied in the case of *set out*, students may familiarise themselves with possible translation solutions and analyse them in particular textual context.

Additionally, *take in* has an idiomatic usage and it appears in such a context in the EPB corpus in the following example:

It's a lot to **take in**. (Brown, *The Da Vinci Code*)

Pewnie niełatwo ci to wszystko **ogarnąć**. (Brown, *Kod Da Vinci*)

Табе столькі трэба **асэнсаваць**. (Brown, *Kod da Vinčy*)

While the Belarusian translator uses literal translation of the idiomatic meaning of *take in*, the Polish translator incorporates the word *ogarnąć*. The first three meanings of that word defined in the Dictionary of the Polish Language (Wydawnictwo Naukowe PWN, 2020) are strictly connected with physical abilities and with spacial properties (*to encompass*, *to embrace*), and only the fourth meaning has to do with feelings and emotions. Those two translations are examples of two different strategies in translating a verb used idiomatically.

To sum up, this section began by describing the overall benefits of using parallel corpora in the translator training and argued that extracting and analysing phraseology is particularly interesting in this context. In order to demonstrate this type of corpus use, several examples of translation solutions and strategies have been investigated. First, an idiomatic phrase *neck of the woods*. Second, a rare phraseologism, namely *donkey's years*. Third, the phraseological equivalence in the case of the phrase *bad luck*. And finally, two instances of phrasal verbs, that is *set out* and *take in*.

All of the examples discussed in this section and their translations were accompanied by an analysis conducted from the point of view of a translator-to-be. What these case studies revealed is that the EPB corpus is a tool that can be effectively used in translator training for a wide range of tasks, as the corpus not only provides ready-to-use solutions, but also allows for inferring and understanding the underlying translation technique.

All in all, Chapter 3 focused on practical uses of the EPB corpus in translation studies, specifically in theoretical, descriptive and applied studies. Chapter 3.1 considered three translation universals (TU), namely explicitation, simplification and levelling out. Using methodologies applied in existing theoretical research the topic of TU was investigated within language pairs not previously analysed in this respect. The results obtained for the EPB texts were supplemented with data from open parallel corpora in order to allow comparison with other Slavonic languages paired with English and other Germanic languages. The analysis showed that what was frequently described as a translation universal is in fact dependent on the language properties and might be specific only to a particular language pair, rather than the whole body of translated literature.

Next, Chapter 3.2 presented an overall stylometric analysis of the corpus as an example of use of the EPB corpus use in descriptive translation studies. Using dedicated software, dendrograms, or visualisations of general stylistic features of the EPB corpus, were created. Thanks to them it was possible to categorise the texts and select one group for further detailed analysis. The Belarusian translation of Christie's *The Kidnapped Prime Minister* was identified as an outlier and scrutinised adopting classical corpus linguistics methods, that is keywords and key clusters analysis. This identified features distinguishing that particular text from other short stories by the same author.

In Chapter 3.3. another branch of translation studies was exemplified by the use of the EPB corpus in analysing phraseology for training purposes. Querying for and analysing some phraseologisms as well as phrasal verbs demonstrated some of the many ways the parallel corpus, in particular EPB, can be exploited for familiarising students with real-life examples of translation solutions and extrapolating the translation techniques from them.

## Chapter 4: Applications of the EPB corpus (B): Discourse analysis

This chapter embarks on an analysis of discourse around terms for men and women within the EPB corpus. Investigation of the use of each word leads to establishing a way of approaching these concepts in each language, that is original English and translational Polish and Belarusian. What follows is a comparison of these features with discourses already identified and presented by researchers exploring the topic in original Polish and Belarusian texts.

Discourse, as mentioned in Section 1.3, is treated here broadly and can be described in terms of ideology. By analysing the systemic choices of, for example, near-synonymous lexical items a researcher can investigate the relationship between form and function in a given text. In other words, it is stipulated and it has been proven that the meaning of a text is expressed by specific lexical and grammatical choices. It is these choices that need to be identified and translated appropriately. “Discourse analysis in its various forms is a powerful tool for uncovering the processes and explaining the motivation behind the author’s and the translator’s choices” (Munday & Zhang, 2015, p. 333), therefore it has particular applications for translator training and translation analysis. As such DA is a fitting complement for the previous chapter.

One “omni-relevant category in most social practices” (Lazar, 2005, p. 3) and therefore one of the most common issues discussed within discourse analysis is gender. Language is a tool for constructing, and later perpetuating or challenging gender roles and power structures within wider social and cultural context. It concerns not only the source language, but also translations. As researchers point out, “societies interact through translation” (Castro, 2013, p. 6), therefore it is justified to ask to what degree certain translations legitimise or undermine the current state of gender in a given language.

Thus this section attempts to investigate the whole of the EPB corpus to give a general view of gender in the collected English texts, as well as their Polish and Belarusian counterparts. This analysis will start from a short overview from a grammatical point of view, due to the fact that grammatical gender is an important aspect in both Polish and Belarusian, and then the attention will be paid to purely discursive analysis of “men” and “women” in the corpus.

### 4.1. Gender and grammar

The topic of grammatical gender in Polish and the problem of translating it from and into that language has attracted much attention from researchers over the years. Koniuszaniec and Błaszowska (2003) give an extensive overview of the complexity of this issue, discussing the categories of gender in Polish and how masculine is a norm in language, not only on the discursive, but also the more basic, grammatical level. The authors indicate strategies for avoiding sexist language and notice the ongoing discussion around feminatives in the Polish language. To this day using feminine forms, even legitimised by long-existing lexicographic evidence, is controversial, and only recently (2019) the Polish Language Council issued an official statement supporting female names of professions (such as *pilotka* – female pilot) and titles (such as *doktorka* – female doctor).

When the topic of gender in a language is so problematic and broadly discussed, it can be no different when it comes to the translations from and into that language. What seems to be particularly challenging is translating into Polish texts where a character's sex or gender is not explicitly revealed, either for intended stylistic effect or due to the language's grammatical restrictions. In such cases the translators must seek for techniques that may cause the target text to sound slightly unnatural, such as using sentences without a subject (Hejwowski, 2011) or using the 3<sup>rd</sup> person plural rather than the impersonal form (Szymańska, 2008). Some techniques might even interfere with the meaning of the source text, for example changing the verb tense into present or future (Hejwowski, 2011) or changing from direct speech to free indirect speech and using only two out of seven declensions, namely Nominative and Vocative (Szymańska, 2008).

The discussion about gender features in the language also concerns Belarusian, as much as Polish. Due to grammatical similarities between the two it can be easily concluded that in Belarusian too the gender is nearly always explicit. Moreover, Belarusian, just like Polish, favours the masculine forms, and struggles with feminisation processes. Feminine forms which are not widespread and prevalent, are accused of "bad sounding" (Baranova, 2017, p. 12). Nevertheless researchers regard feminisation as a necessary step in the development of the language and as a logical answer to the ongoing changes in the society and culture (Baranova, 2017).

Having defined how important the grammatical aspect of gender is in Polish and Belarusian, this section will now move on to examine a few features that might indicate the preference for male or female forms. One such feature is neutral collective words, for example *brotherhood* and *sisterhood*. Naturally, the first word has a broader meaning than the second, owing to the fact that *sisterhood* only refers to the relations between sisters or, in broader context, women, whereas *brotherhood* can refer to the relations between people in general or between abstract concepts (such as nations). The results of that semantic characteristics is that *brotherhood* appears in the BNC eight times more frequently (278 vs. 35 occurrences) than *sisterhood*. Even though the difference in usage contexts in Polish and Belarusian is similar to English, the disproportion between the two words is far greater: 56 times more frequent in case of Polish (*braterstwo* – 455 vs. *siostrzeństwo* – 8) and 92 times more frequent in case of Belarusian (*братэрства* – 833 vs. *сястрынства* – 9).

Another category of neutral collective words are terms describing pairs of people, such as *aunt and uncle*, as in Rowling's *Harry Potter and the Half-Blood Prince*:

Nothing much, I've just been stuck at my aunt and uncle's, haven't I?

A few days before I came to fetch you from your aunt and uncle's, in fact.

The Polish translator applied the word *wujostwo*, a collective noun created from the word *wuj* (*uncle*). Belarusian lacks this particular collective noun and therefore the translator uses the words *сваячкі* (*relatives* in diminutive form serving as irony in this case) and *дзядзька з цёткай* (*uncle with aunt*). The latter solution has been used in two other cases of translating *aunt and uncle* and it is worth noting that the original order was changed so the male gender appears first, and additionally the possessive pronoun *your* has been translated in the plural form (which is common for all genders). Similarly in Polish translations of remaining occurrences of *aunt and uncle* the order favours the masculine form, even though two separate nouns rather than one collective were

used. The possessive pronoun, however, has been translated separately for each word, so, for example, the phrase “at your aunt and uncle’s house” becomes “w domu twojego wuja i twojej ciotki” (*at the house of your [masculine] uncle and your [feminine] aunt*).

Perhaps the most prominent example of such a problem are collective polite forms of address in Polish, particularly *Państwo* (*Mr and Mrs*, collective). This form appears 188 times in the Polish subcorpus of EPB and the table below presents what it might stand for:

English original	Polish translation	Belarusian translation	Polish feminine verb form
This room is <u>yours</u> as long as <u>you</u> care to use it. (Brown, <i>Da Vinci Code</i> )	To pomieszczenie jest do <u>państwa</u> dyspozycji tak długo, jak będą <u>państwo</u> <u>mieli</u> ochotę z niego korzystać.	Пакой <u>ваш</u> . Можаце карыстацца ім, колькі <u>вам</u> будзе заўгодна.	miały
[...] the Darlings had become acquainted with her in Kensington Gardens [...] (Barrie, <i>Peter Pan and Wendy</i> )	[...] <u>państwo</u> Darling <u>poznali się</u> z nią w parku Kensington [...].	<u>Дарлінгі</u> пазнаёміліся з ёй у Кенсігнтанскім садзе [...].	poznały się
<u>Mr and Mrs Spenlow</u> preferred the small back sitting-room. (Christie, <i>The Tape-Measure Murder</i> )	<u>Państwo Spenlow</u> <u>woleli</u> mały salonik na tyłach domu.	[...] ў прэднім пакоі <u>містэр і місіс Спэнлаў</u> бываюць рэдка, <u>яны</u> больш любяць гасціны пакой.	wolały

Table 4.25. Instances of forms of address in the EPB corpus.

The most popular way of addressing a pair of people of different sexes/genders in English is simply via personal pronoun *you* (plural), which is the same as Belarusian *вы* followed by the plural form of the verb that is common for both grammatical genders. In Polish, however, there is distinction between feminine and masculine forms of verbs in plural number, and *Państwo* is always followed by the masculine. The collective noun *Państwo* is a must in the cases where the English personal pronoun appears. In Polish using the same pronoun when addressing a pair of unacquainted adults would be regarded disrespectful. Similarly, even though using the plural form of the surname (e.g. Darling → Darlingowie) is correct and acceptable, it is regarded as more polite to use the surname preceded by the word *Państwo*.

This form of address seems to be inevitable and it is used even in translating two separated male and female forms. *Mr and Mrs* (which appears in the English subcorpus 38 times) is mostly translated with the collective *Państwo*. The exact Polish analogue, *pan i pani*, appears only 4 times in the EPB corpus, while the Belarusian version, *містэр і місіс* (which is simply transliteration from English, not Belarusian native forms), is used 26 times.

Polish *Państwo* may be translated into Belarusian as *снадапчмва* (*Mr and Mrs*, collective). This is however highly formalised and used almost exclusively in the phrase *шаноўнае снадапчмва* (*dear/honourable/respected ladies and gentlemen*, plural). Out of 17 occurrences of the word *снадапчмва*, most (12) appear in Tolkien’s texts and serve as a way of addressing a group of people (*my lords, people, all*). In some of these cases, when the group of people addressed consists

exclusively of men, the Belarusian translation is more inclusive than English, as it refers both to females and males.

To conclude, even though Polish and Belarusian deliver means for producing translations that favour female forms (like in English *aunt and uncle*) or promotes more equal attitudes (for example, changing the order of female and male forms in some of the phrases or using both forms instead of a collective noun), translators do not take that opportunity and explicitly favour masculine forms in the language. This is more prominent in the Polish language, as collectives are always followed by a masculine form of a plural verb, whereas in Belarusian the plural form of the verb is common for both masculine and feminine grammatical genders.

#### 4.2. Gender discourse in English

So far this chapter has focused on a purely grammatical approach towards gender in English, Polish and Belarusian. The remaining part will be devoted to analysing the discourse itself using some of the corpus methods mentioned in Section 1.5. Firstly, the frequency and dispersion of chosen key terms will be investigated, and secondly collocations and their concordances will be analysed. Results obtained for the English part of the EPB corpus will be compared with outcomes of the analysis of remaining data in order to ascertain the similarity of discourse patterns in original English language and in translational Polish and Belarusian, and ultimately to assess to what degree the existing discourse around gender in English is reinforced or undermined in translations.

Frequency lists, often disregarded as too simple or too generalising, are, in fact, a good starting point for analysing the corpus data, under the condition of being used sensitively. Frequency indicates the direction for further investigation and together with dispersion “can help to give the user a sociological profile of a given word or phrase enabling greater understanding of its use in particular contexts” (P. Baker, 2006, p. 47). Even when applying such a seemingly simple and straightforward method as frequency counts, the researcher is immediately faced with choices determining our results to a great degree.

The first such decision is choosing search terms. As Baker (P. Baker, 2014) notes, when looking at gender bias we should ask “what words are used to refer to gender identity” (2014, p. 78), and these are not always self-evident. The most obvious, such as *man* and *woman* or *boy* and *girl* are a good start, they can however be used in multiple contexts, for example exclamative (*oh, boy!*) or refer to a much broader category (*men* meaning the whole of humankind). Another issue is looking for both singular and plural forms, terms of address (*Mr/Mrs*), pronouns (*he/she*) or even proper nouns. Some (P. Baker, 2014) claim that the researcher should identify every gendered word in order to present a full picture of gender in a given corpus. On the other end of the spectrum there are also genderless terms, such as *person* or *human being*.

Taking all those considerations into account, the line of investigation is as follows. Firstly, the frequency list containing feminine and masculine nouns, as well as personal pronouns *she* and *he*, is created. Concerning the pronouns only the basic forms are taken into account. This is due to the fact that *she* and *he* indicate the female/male not only as an object (which is indicated by the most frequent bi-gram *she was* and *he was*), but also the subject in discourse (the second most frequent

bi-gram for *she* being *she said* and third most frequent for *he* – *he said*). In the frequency list the synonyms for *woman* and *man* listed in the Merriam-Webster Thesaurus are taken into account. Next, the dispersion of two most numerous items in both columns (feminine and masculine) is analysed.

Merriam-Webster Thesaurus lists only a few synonyms for the word *woman*, that is *female*, *lady*, *skirt* (slang), *gal*, *gill*, *girl*, *girlfriend*, *inamorata*, *lady*, *ladylove* and *old lady*. Querying the EPB corpus with KonText reveals that, in the EPB corpus, the words *skirt* and *gill* are not used as synonyms of *woman*, *inamorata* does not appear at all, and *old lady* is not taken into account as it is not a single word. The fourth, dialectal meaning of *woman* was not taken into account.

Concerning *man*, the thesaurus provides more synonyms, that is *bastard*, *beau*, *bloke*, *boy*, *boyfriend*, *buck*, *cat*, *chap*, *chappie*, *dude*, *fella*, *fellow*, *galoot*, *gent*, *gentleman*, *guy*, *hombre*, *jack*, *joe*, *joker*, *lad*, *male*, *old man*, *swain*. Out of these alternatives *buck* and *cat* do not appear in the corpus as synonyms for *man*, *Jack* and *Joe* are used as proper names, *dude* and *chappie* do not occur, and *hombre* is used exclusively in Hemingway's *For Whom the Bell Tolls* which is set in Spain thus the usage of the Spanish word. *Old man* is not taken into account for it is two rather than one word and *swain* does not occur in the corpus as a *man* synonym. Third meaning of the word *man* included in the dictionary is much broader (human) and it was not considered in the analysis, because it refers to both females and males and does not reveal any differences in gender discourse.

The table 4.25 showcases all occurrences of the synonyms mentioned above. All nouns taken from the dictionary have been used as lemmas in the KonText query, therefore the number of occurrences refers to both singular and plural forms. The exception are pronouns, as discussed above:

Number of occurrences	Feminine pronouns & nouns	Masculine pronouns & nouns
56,516		he
18,855	she	
7,979		man
2,220	woman	
1,559	girl	
1,399		boy
641	lady	
413		fella
374		fellow
205		guy
201		gentleman
140	female	
120		male

Number of occurrences	Feminine pronouns & nouns	Masculine pronouns & nouns
108		chap
105		lad
82		bastard
41		bloke
26		boyfriend
14	girlfriend	
3		joker
2		beau
1	gal	
1		galoot
1	ladylove	
<b>IN TOTAL</b>	<b>23,431</b>	<b>67,575</b>

Table 4.26. *Feminine and masculine pronoun and nouns in the English subcorpus of the EPB corpus.*

One aspect is clearly visible from this query, that is the masculine prevalence, both in terms of the number of synonyms occurring in the corpus (as well as in the dictionary which was used as the basis for the query list) and in terms of the number of occurrences of each synonym. The personal pronoun *he* occurs over three times more frequently than *she*. Taking into account only nouns, no pronouns, again proves the disproportion of feminine (4576 occurrences) versus masculine (11061 occurrences) gender, however among masculine nouns 72% of occurrences are *man* and *men* – in this case there is no way of distinguishing the type of usage (denoting a male human or human in general) as the EPB corpus is not semantically tagged. All the same, the masculine gender prevails in terms of quantity. Considering less-frequent words, the discrepancy in the number of occurrences is almost lost in the case of *girl* and *girls* versus *boy* and *boys*, and it reverses when it comes to *lady* and *ladies* – these occur over three times more often than *gentleman* and *gentlemen*. The latter case is shown in the table from the illustration 4.15:

	lady	gentleman		lady	gentleman
Adams	1		Kerouac	1	
Barrie	2	1	Kesey	1	
Baum	1		King	7	23
Bradbury	5	1	Kipling	16	5
Brown	7	5	Lewis	12	5
Bukowski	1		Lovecraft	2	
Cheever	2		Mansfield	6	1
Chesterton	3	6	Mathews	3	
Christie	95	17	Maugham	1	
Dahl	3		McCullers	15	1
Faulkner	1	4	McCullough	35	8
Fitzgerald		1	Orwell	2	5
Gaiman	4		Rowling	62	11
Golden	7	2	Saki	4	
Golding	3		Segal	3	1
Greene	27	21	Sheckley		2
Gulik	10	1	Steinbeck	48	1
Hemingway	4		Tolkien	165	1
Jackson	3		Travers	3	1
James		6	Waugh	10	
Joyce	62	68	Woolf	4	3

Illustration 4.15. Dispersion of the words 'lady' and 'gentleman' in the EPB corpus.

A dispersion query for *lady* shows that it occurs over ten times in Christie's short stories (95 hits), in Green's novels (27), Joyce's texts (62), McCullough's *The Thorn Birds* (35), three books from the Harry Potter series (62), Steinbeck's *Grapes of Wrath* (48) and all books from the *Lord of the Rings* series (165). In the Harry Potter novels *lady* is used almost exclusively as a part of the proper name *Fat Lady*, while Tolkien applies *Lady* as a form of address towards the sparse female characters (for example Lady Galadriel and Eowyn – Lady of Rohan), in Christie's texts *lady* usually serves as a form of address, while in rest of the cases it is usually the synonym for *woman*. Similarly *ladies* are in most cases a synonym for *women*, only in 12 out of total 95 hits *ladies* is appellative (*ladies and gentleman*). On the other hand, *gentleman* occurs over ten times only in *Ulysses* and the plural form – in *Ulysses* and King's *Breathing Method* which is the story of a gentlemen's club.

These dispersion patterns might indicate that, firstly, overall there are more male characters in the texts comprising the EPB corpus, and secondly, *lady* prevails as a *woman* synonym in the older or archaised texts, while *man* is replaced with *gentleman* only in rare cases.

As a way of supplementing this investigation the dispersion of four items is also analysed: *he*, *she*, *woman/women*, *man/men* (full data is available in Appendix 4). The personal pronoun *he* appears in 111 out of 113 texts of the corpus, and in almost half of the cases (52) there are over one hundred hits, while *she* occurs in only 88 samples, 31 of them containing over one hundred occurrences. A high number of occurrences is characteristic of long epic forms, however some titles show a massive disproportion, most notably Tolkien's *Hobbit* with almost two thousand occurrences of *he* and just one of *she*. There are also some texts, among novellas and short stories, exhibiting major discrepancies, in particular two samples containing over one hundred hits for *he* but none for *she*. Interestingly, one text, Dahl's *Lamb to the Slaughter*, shows the reverse tendency – it contains over one hundred instances of the pronoun *she*, but half as many of the pronoun *he*. Observations for the

nouns *man/men* and *woman/women* follow a very similar pattern as *she* and *he*; the only difference is reduced disproportions as compared with pronouns.

To sum up, the initial analysis of frequency lists and dispersion data for the EPB corpus shows that the English literature of the 20<sup>th</sup> and 21<sup>st</sup> centuries that has been translated into Polish and Belarusian is biased towards male characters. This is however only one level of male prevalence in the English part of the corpus. Therefore the next step is to find out how females and males are described in the texts, and how they speak. In order to do that, lists of collocates of the nouns analysed in dispersion study, with the addition of *girl/girls* and *boy/boys* were created. The addition has been motivated by the desire of comparing the discourse around both genders in relation to the age (following Baker 2014).

Regarding collocations, there are a few factors to be considered. Firstly, the definition of collocate is not accurate and differs greatly depending on the criteria a researcher applies. Therefore the first step of the investigation is choosing the window span, that is to define the distance between the keyword and the co-occurring words. A wide window might include contexts that do not concern the term in question, on the other hand crossing the sentence boundary is desirable to some degree, as very often the description appears in the sentence preceding the object. A too narrow window is likely to return mostly grammatical and idiomatic results. Another issue in analysing collocates is the way a researcher orders them. Ranking by raw frequency will reveal articles, prepositions, conjunctions and pronouns at the top of the list. This is why a number of statistical tests have been designed to extract words with particular features: high frequency function words, low frequency content words or a mixture of both. Depending on the characteristics of the corpus, available software options and, most of all, phenomena of interest, the preferred statistical technique will vary.

In this study, to accomplish the goal of analysing the gender discourse, the following steps were taken. Firstly, plural and singular forms of nouns were analysed separately, to determine whether there are differences in the discourse around females and males on individual and collective levels. Secondly, the mutual information (MI) statistical measure was applied. MI focuses on the strength of association between the words and indicates, for example, idiosyncratic phrases, as it focuses on low frequency content words. In relatively small-size corpora, such as subcorpora of the EPB corpus, statistical measures being the combination of the two basic types mentioned above are expected to return lexical words with high frequency, nevertheless they place more emphasis on grammatical rather than lexical words. An example of this is the log-likelihood (LL) measure – the LL collocates list for *man* contains fewer than 30% content words in the highest-ranked hits. Put simply, MI favours in-frequent and exclusive combinations, while LL reveals frequent and non-exclusive ones. However, for comparison reasons, the LL values are present in the full collocation report available in Appendix 5.

The collocates lists were created using the KonText application via searching for lemmas co-occurring with the chosen keywords (that is *man*, *men*, *boy*, *boys*, *woman*, *women*, *girl*, *girls*). The minimum frequency in the corpus and in the given range was set at 5, and the search window of 5 to the left and 5 to the right was applied. The first 50 hits of each list were analysed via close-reading

of the concordance lines in order to exclude collocations appearing in less than three different texts or in the texts by exclusively one author (such as *Vitruvian man* occurring only in Brown's *Da Vinci Code*). Moreover, concordances were scrutinised in order to ascertain the context in which the particular collocate is being used. The result are lists of top significant occurrences which, together with frequency values, as well as statistical test results, are available in Appendix 5. They include not only adjectives but also verbs and prepositions, as they can indicate whether the person is a subject or object in the discourse.

The juxtaposed collocates lists at first sight exhibit several differences. Firstly, collocates of *man* and *men* which are dispersed in the corpus and cannot be associated only with idiosyncrasies, are much less numerous than the rest of the search terms. There are 26 such valid collocates for *man* and only 17 for *men*, while the number for *boy*, *boys*, *woman*, *women*, *girl* and *girls* varies from 33 to 49. Secondly, most of *man* and *men* collocates are adjectives and nouns, and there are hardly any verbs, prepositions or numerals, unlike in the case of the remaining words investigated in this study.

The list of *man* collocates contains 22 adjectives, 3 nouns and 1 verb, all of which can be divided into the several categories. This division reflects the observed common features among these particular results, therefore the classification is different for each list:

looks	age	state or non-visual feature	other – nouns	other – verbs
portly, fat, haired, eyed, faced, tall, taller	elderly, old, young, aged	wizened, drunken, mortal, wealthy, intelligent, valiant, married, blind, dumb, decent, crazy	destiny, uniform, stranger	brace

Table 4.27. Collocates of the word 'man' (EPB corpus).

Most of the collocates presented in the table are evaluative and in most cases (11) they focus on the inner features of the character rather than on the looks. Four can be classified as strictly positive, such as *finest*, *decent*, *valiant* or *intelligent*, and five are associated with negative assessment: *fat*, *drunken*, *dumb* and *crazy*, as well as *portly* which is a euphemism for *fat*. Some of the examples appear in various contexts, such as *faced*. It may be negative, as in *silly-faced* or *brutal-faced*, positive, as in *open-faced*, but mostly it is simply neutral, e.g. *round-faced*.

In all examples but one the noun *destiny* is a part of the phrase *man of (high) destiny* – this epithet means a person destined for greatness and was originally used to describe Napoleon, therefore it might be associated with military discourse. Similarly, the noun *uniform* denotes a military uniform in all instances but one.

There is no dominating topic in the case of *man*, only a hint of a military context and a symptomatic prevalence of descriptors referring to the inner features of a character. The collocates of *men* can be described in a similar manner, with the addition of quite a few denoting the collective nature of the word:

looks	age	non-visual feature	collectivity	other – nouns	other – verbs	other POS
tall	younger	armed, equal	group, eight, many, two, four	uniform	bid, gather, surround, create	fro, among, these

Table 4.28. Collocates of the word 'men' (EPB corpus).

The adjective *armed* as well as the noun *uniform* point to male characters repetitively occurring in a military context in the EPB corpus. The prevailing topic emerging from the list of collocates is the collectivity – it is indicated not only by several numerals, a determiner *many* and a noun *group*, but also verbs, *gather* and *surround*. The latter one also occurs in a close proximity to other collocates, as illustrated in the following concordances:

Hits: 6   i.p.m. 0: 1.53 (related to the whole EPB English full)   ARF 0: 3   Result is shuffled	
doc#100,Steinbeck	The front of the truck was <b>surrounded</b> by the armed <b>men</b> . Some of them, to make a military appearance
doc#100,Steinbeck	hissing glare of the gasoline lantern. The gathering of <b>men surrounded</b> the proprietor. Tom drove the Dodge to the
doc#62,London	not snap, nor threaten to snap. The other <b>men</b> came up, and <b>surrounded</b> her, and felt her
doc#9,Brown	play. It 's over, Fache knew. His <b>men</b> would have the truck <b>surrounded</b> within minutes. Langdon was
doc#62,London	public life, in a cage, <b>surrounded</b> by curious <b>men</b> . He was exhibited as "the Fighting Wolf,
doc#62,London	steam-boat's deck was usually <b>surrounded</b> by curious <b>men</b> . He raged and snarled at them, or lay

Illustration 4.16: Word 'man' collocating with lemma 'surround' (EPB corpus).

The collocates frequently appearing in the discourse surrounding *boy* might be classified in the same categories as in the case of *man*; the particular examples are shown in table 4.28:

looks	age	state or non-visual feature	other – nouns	other – verbs	other POS
little, small, large, pale head, shoulder, lip	old, nine	naughty, foolish, dear, poor, quiet, nice, good	boy, girl, father, family	snarl, carry, shout, nod, throw, remember, live, kill, shake, play, believe	who, my

Table 4.29. Collocates of the word 'boy' (EPB corpus).

Out of 33 top-ranked collocates 11 are verbs, 7 are adjectives referring to the inner features of a character and 7 descriptors (including adjectives and nouns) describing looks. Particular collocates suggest that *boy* is often treated with disregard, in a patronising way. He is described as *little* and *small*, he can be *naughty* (which is an adjective used almost exclusively in regard to children and only humorously towards adults), *foolish* or *dear*.

Interestingly, *boy* frequently appears as a repetitive element in a sentence or just across the border of sentences, hence the word *boy* being one of the significant collocates. Some examples showcasing that phenomenon are:

"Thanks," the boy said. The boy carried the hot can of coffee [...]" Hemingway, *The Old Man and the Sea*

"A boy. A croppy boy." Joyce, *Ulysses*

"The older man laughed and he looked at his boy, and his silent boy grinned almost in triumph." Steinbeck, *Grapes of Wrath*

Naturally, some of the examples of repetition have an important function, that is to avoid ambiguity in cases when a personal pronoun can refer to more than one noun. Nevertheless, in several instances the repetition is redundant from the semantic point of view, though it might play another role, such as emphasising an item of information, building an atmosphere or mimicking a certain type of language, for instance, conversational.

Table 4.29 contains the top significant collocates of the word *boys* in the EPB corpus:

looks	age	state or non-visual feature	other – nouns	other – verbs	other POS	collectivity
little, small	older, young	better, good	boy, girl	play, shout, work, put, sit, stand, will, get	among, together, around, other, these, those, through, though, from, than, with, who, they, one	five, two, three, all, some

Table 4.30. Collocates of the word 'boys' (EPB corpus).

In terms of adjectival and nominal collocations the word *boys* is similar to *boy*. The difference between the words lies in the presence of collocates clearly pointing to the collective nature of the plural noun, as in case of *men* versus *man*. Pronouns, particularly demonstrative, indicate *boys* as an object, someone being talked about in the text, rather than someone frequently speaking.

Collocations of female nouns differ significantly from those of male nouns. The proportion of different types of collocates and their examples support reports about gender bias in English, as showcased in table 4.30:

looks	age	state or non-visual feature	other – nouns	other – verbs	subjectivity
haired plump handsome slender naked beautiful ugly fat thin tall	aged elderly young younger sixty old thirty older middle	lovely stupid poor fine	doorway Mrs daughter street	marry screaming dress ( <i>rarely a noun</i> ) approach love	whom who

Table 4.31. Collocates of the word 'woman' (EPB corpus).

In case of *woman* there are far more adjectival collocates referring to age (9 of them), moreover, some of them come in the positive and comparative degree which suggests *woman* being frequently compared to others. The proportion of looks descriptors vs the descriptors of inner traits (10–4) is the opposite of the same proportion in case of *men* (7–11). There is more verbs among the collocates, but they function in a specific way. For example, the verb *marry* in 7 out 10 cases describes someone marrying a women, not the other way around. Similarly, the most numerous verb, *love*, is used mostly with the word *women* being the object of an action, which is visible in the examples from the corpus presented in illustration 4.17:

Hits: <b>21</b>   i.p.m. <sup>0</sup> : 5.37 (related to the whole EPB English full)   ARF <sup>0</sup> : <b>9</b>   Result is shuffled	
doc#81, Maugham	Oh, it is dreadfully bitter to look at a <b>woman</b> whom you have <b>loved</b> with all your heart and soul
doc#45, Joyce	understand. He would not believe in <b>love</b> , a <b>woman</b> 's birthright. The night of the party long ago
doc#23, Faulkner	happen to a man is to <b>love</b> something, a <b>woman</b> preferably, well, hard hard hard, then to
doc#85, McCullough	be simpler men, men who could surely <b>love</b> a <b>woman</b> before all else. Men like Luke O'Neill, for
doc#34, Hemingway	." " Then you made <b>love</b> ," the <b>woman</b> said. "Be as careful with her as you
doc#79, Mansfield	much as it is possible for me to <b>love</b> any <b>woman</b> , but, truth to tell, I have come
doc#85, McCullough	. It was the child he <b>loved</b> , not the <b>woman</b> . The woman he had hated from the moment she
doc#34, Hemingway	man is after he has had sexual intercourse with a <b>woman</b> that he does not <b>love</b> . " And you,
doc#85, McCullough	the time she had furnished her <b>love</b> for him with <b>woman</b> 's objects. Admit it, he had physically wanted
doc#100, Steinbeck	And he said, "John, there's a <b>woman</b> so great with <b>love</b> -she scares me. Makes
doc#45, Joyce	Recipe for white wine vinegar. How to win a <b>woman</b> 's <b>love</b> . For me this. Say the following
doc#84, McCullers	two decades past. The preacher may <b>love</b> a fallen <b>woman</b> . The beloved may be treacherous, greasy-headed
doc#85, McCullough	love some inanimate thing more than they could <b>love</b> a <b>woman</b> ? No, surely not all men. The difficult
doc#85, McCullough	him -I never did <b>love</b> him the way a <b>woman</b> ought to love the man she marries, and maybe
doc#31, Greene	. When there was a choice between <b>love</b> of a <b>woman</b> and hate of a man, her mind could cherish
doc#34, Hemingway	him. " Did you make <b>love</b> ?" the <b>woman</b> said. " What did she say?" " "
doc#85, McCullough	and his state of mind--not <b>love</b> of a <b>woman</b> , nor love of money, nor unwillingness to obey
doc#45, Joyce	loved for ever, they say. Ugly : no <b>woman</b> thinks she is. <b>Love</b> , lie and be handsome
doc#14, Chesterton	him already; you are separating him from the good <b>woman</b> he <b>loves</b> and who loves him. But you will
doc#30, Greene	floor, and yet - it was like leaving a <b>woman</b> one <b>loved</b> untouched, untasted, to go away and
doc#9, Brown	voices accelerated now. Louder. Faster. " The <b>woman</b> whom you behold is <b>love</b> !" The women called

Illustration 4.17. Concordances of the verb 'love' collocating with the word 'woman' in EPB corpus.

This phenomenon is confirmed by the presence of pronouns *whom* and *who* among the most significant collocates – it points to the fact that *woman* is more likely to be an object, rather than subject in the text, in other words she is being talked about rather than described as talking. In fact, the activities she performs are *screaming*, *dressing* and *approaching*.

In terms of the plural, that is *women*, the adjectival collocations are less numerous, but there are many more verbs and collocates pointing to the collective nature of the noun, as showcased in table 4.31:

looks	age	collectivity	other – verbs	other – nouns	other POS
beautiful white black	older young old	two most three many four hundred both together	scream wear bear laugh watch stand run bring work use talk become	child dress (also a verb in some instances) baby man woman street house world girl boy	among after who around soon near those some always round front such other

Table 4.32. Collocates of the word 'women' (EPB corpus).

In terms of collectivity indicators, *women* does not differ substantially from *men* and *boy*, however describing looks, comparing age and the context of children and house is present here. As in the case of *woman*, *women* are frequently being talked about and with a big dose of confidence which is visible in the concordance lines for co-occurrences with the word *always*.

The collocational pattern for *girl* and *girls* seems to follow that emerging in the case of *woman* and *women*. This is presented in table 4.32:

looks	age	state or non-visual feature	other – nouns	other – verbs	other POS
attractive pretty little beautiful big tall thin  hair shoulder arm	young age	honest clever lovely poor nice good	village boy baby girl brother name dress (rarely a verb) table	live meet learn lift love	who herself next

Table 4.33. Collocates of the word 'girl'.

Just as *woman*, *girl* collocates with descriptors of looks much more often than with the descriptors of the inner traits. Similarly to *boy*, *girl* frequently appears as a repetitive element in a sentence or just across the border of sentences:

I run me down a girl, a hoor girl, like she was a rabbit. (Steinbeck, *Grapes of Wrath*)

But now she said something he could not hear to the girl and the girl rose from the cooking fire [...] (Hemingway, *For Whom the Bell Tolls*)

The maid saw the girl doing it. The girl will pay. (Golden, *Memoirs of a Geisha*)

The function of such a technique corresponds to what has been observed earlier in the case of the repetition of the word *boy*.

Most common significant collocates of the plural form, *girls*, differ significantly in some categories of descriptors, as showcased in table 4.33:

looks	age	collectivity	other – verbs	other – nouns	other POS
little  hair	young	group together two most many few some one all	get know think want give	girl school boy woman year man	those other always these who too how like (rarely a verb) any with never of no in only there

Table 4.34. Collocates of the word 'girls'.

Unlike the singular form, the word *girls* has few adjectival collocates, only two collocates referring to the looks and non describing the inner traits of the character. Similarly to other plural forms (such as *men* and *boys*) there is a high number of quantifiers and numerals pointing to the collective discourse around *girls*. What differentiates *girl* and *girls* the most is the number of function words collocating with these words, especially pronouns, such as *those*, *these*, *who* or *there*. This feature is characteristic of description and referring to something, as opposed to reporting someone performing an activity. This is consistent with the findings on other female words in this study.

All in all there are several tendencies emerging from the English texts collected in the EPB corpus. Firstly, there is a difference between singular and plural forms of the analysed nouns. In the case of plurals there is always a strong category of collocates referring to the collective nature of the noun. Secondly, there are differences in the discourses around the two genders. In the case of females the themes of age (in case of *woman* and *women* the age is frequently being compared) and marital status are much more prominent. There are more collocates pointing to females being more often objects of the conversation rather than subjects in the text. Lastly, there is more focus on the looks and little (or none – as in the case of *girls*) attention paid to the inner traits of the character. Thus the evidence gathered in this study supports earlier findings (Sigley & Holmes, 2002; Stokoe, 1998; Taylor, 2013) concerning characteristics of *man* and *boy*, and *women* and *girl*. There is much less difference in the discourse around *woman/women* versus *girl/girls*, than in case of males, that is *boy/boys* versus *man/men*.

#### 4.3. Gender discourse in Polish translational data and in contrast with monolingual corpus.

Having gathered and analysed data concerning gender discourse in English this chapter now moves on to discussing the same phenomenon in Polish. For the purposes of conducting the synonyms and collocations study, the full lemmatised, though not aligned corpus of Polish has been uploaded in the KonText application. Due to the differences between the languages (discussed in Chapter 1.1) it is not feasible to follow all steps of the procedure described in previous pages. The frequency investigation will not consider personal pronouns as these are omitted in many cases, especially in Polish. A simple query reveals the disproportion of *he/she* usage in the three languages; its results are presented in table 4.34:

English		Polish		Belarusian	
she	he	ona	on	яна	ён
18,854	57,766	1,244	3,064	13,657	34,372

Table 4.35. Number of occurrences of female and male pronouns in the EPB corpus.

Moving on to analysing the synonyms of Polish most common words denoting females and males, let us consider the list compiled by Broniarek (2005). He distinguishes the following 17 synonyms for the word *kobieta* (woman): *kobietka*, *kobieciątko*, *kobiecinka*, *kobiecina*, *kobita*, *pani*, *niewiasta*, *dama*, *damula*, *damulka*, *babka*, *panna*, *pannisko*, *jejmość*, *facetka*, *białogłowa*, *spódniczka*, and for the purposes of this study this list will be enriched with the Polish analogue of *girl/girlfriend*, that is

*dziewczyna*. The following 12 words are identified by Broniarek as synonyms for *mężczyzna* (*man*): *pan*, *dżentelmen*, *osobnik*, *jegomość*, *facet*, *facio*, *faciu*, *chłop*, *gość*, *gościu*, *gostek*, *indywiduum*. Polish equivalents of *boy/boyfriend*, that is *chłopiec* and more colloquial *chłopak*, have been added to the list. Some of the enlisted synonyms are present in the Polish part of the EPB corpus and they have been collected in the table 4.35. The number of occurrences relates to both singular and plural, because, as in case of English, the keyword was lemma:

Number of occurrences	Feminine nouns	Masculine nouns
1866	kobieta	
1608		mężczyzna
1061		chłopiec
950	dziewczyna	
313		chłopak
298		facet
163		chłop
149	dama	
36		jegomość
34		dżentelmen
14	niewiasta	
6	kobiecina	
3	kobietka	
2	kobiecinka	
1	damula	
<b>IN TOTAL</b>	<b>3132</b>	<b>3544</b>

Table 4.36. Lemmas 'kobieta', 'mężczyzna' and their synonyms (EPB corpus).

There were no hits for *kobieciątka*, *damulka*, *jejmość*, *białogłowa*, *facetka*. Moreover, *pani*, *panna* and *spódniczka* are too confusing – *pani*, as explained earlier, serves mainly as a polite form of address, the same concerns *panna*, which in its primary meaning translates as *miss*, finally *spódniczka* literally means *skirt* (diminutive). Additionally *pannisko*, which is augmentative form of *panna* has not been taken into account.

In the case of male synonyms, EPB contains no occurrences of three colloquialisms, *facio*, *faciu* and *gostek*, and a few synonyms were too confusing to take them into account. *Pan*, similarly to *pani* is usually a form of address, *osobnik* (*individual*, *specimen*) refers to both humans and, more often, to items, *indywiduum* is grammatically neutral and can refer to both male and female, lastly *gość/gościu* literally mean *guest* and this meaning – a person visiting someone – is prevalent in the EPB.

Regarding the number of occurrences, the masculine nouns are more frequent than feminine, which corresponds with the results obtained for English. However the overall amount of the words *kobieta*, *mężczyzna* and their synonyms in Polish is smaller than in the original texts. This

disproportion can be explained by some grammatical phenomena, notably using single words denoting features of a character, for example *staruszka* instead of *old woman*, or *ślepiec* instead of *blind man*. This is also characteristic of Belarusian, as explained in the next chapter.

In English the synonyms of *man* occur three times more frequently than synonyms of *woman*, moreover there is a high disproportion of *he* vs. *she* and *man/men* vs. *woman/women* – in these cases the male alternative prevails significantly. In Polish and Belarusian however the results are much more balanced, and there are actually slightly more hits for *woman/women* than *man/men* in both languages. The dispersion patterns for Polish and Belarusian can be found in Appendix 4 – because of the difference in the number of occurrences, all texts with over 20 (rather than 100) hits were marked with dark grey colour. The four main search terms in Polish and Belarusian are almost the same as in English: long epic forms usually contain more occurrences of these and the differences lie in the titles where female nouns get more hits than male.

In the original *For Whom the Bell Tolls* the words *man/men* appear twice as frequently as *woman/women*, yet in the Polish translation the numbers are the opposite. The disproportion is even bigger in Belarusian – female nouns are four times more frequent. In the *Breathing Method* by King, nouns referring to females and males are balanced, with a slight advantage of the pronoun *she* and the nouns *woman/women*. The Polish translation contains five times more female than male nouns, and Belarusian – eight times more. Similarly, in Lewis's *Magician's Nephew* the numbers are balanced for nouns and the pronoun *he* prevails, nevertheless *woman/women* are seven times more frequent than *man/men* in the Belarusian translation and five times more frequent in the Polish version of the novel. Two more texts show significant differences, yet only in the Polish translation. In Steinbeck's *Grapes of Wrath* and Tolkien's *Return of the King* feminine nouns are respectively 26 and 22 times more frequent than male equivalents, yet it is the opposite in the case of English originals, both for pronouns and nouns.

One possible explanation of these differences, in some cases extreme discrepancies, is that the noun *man* is, in fact, used in its wider meaning much more frequently than referring strictly to male gender. Another factor which might influence the number of occurrences for *woman*, *women*, *man* and *men* might be the use of synonyms, that is translators choosing to convey many instances of a simple term such as *man* in several different ways. A query in the aligned component of the EPB corpus reveals such a dependence. The results are presented in table 4.36:

	Number of occurrences of word	
<i>Man</i> overall in English corpus	632	
<i>Man</i> translated as	<i>mężczyzna</i> (pl)	<i>мужчына</i> (by)
	147 (23%)	112 (18%)
	<i>człowiek</i> (pl)	<i>чалавек</i> (by)
	195 (31%)	160 (25%)
<i>Man</i> translated otherwise	Polish corpus	Belarusian corpus
	290 (46%)	360 (57%)

Table 4.37: 'Man' translations into Polish and Belarusian (aligned section of the EPB corpus).

In both Polish and Belarusian corpora the word *man* has been translated with the use of the closest equivalent and the most neutral word, that is *mężczyzna* or *мужчына*, in less than ¼ of the sentences. In both languages ¼ up to ⅓ cases are actually occurrences of *man* denoting much broader concept of *human* and translated accordingly as *człowiek* or *чалавек*. In about half of the cases *man* has been translated in a myriad of other ways – using synonyms of *mężczyzna* and *мужчына*, using personal and demonstrative pronouns, or simply referring to the activity, profession, trait or the name of a man. Examples of these translations can be found in table 4.37:

Polish	English	Belarusian
Naczelnik potrzęsnał głową.	The Head <b>Man</b> shook his head.	Вярхоўны Жрэц пакруціў галавою.
Ślepiec z głową opartą o ścianę leżał w jednym z nasłonecznionych kątów.	The blind <b>man</b> lay in one corner under the sun, his head propped against rock.	У куце пад сонцам, прытуліўшыся галавой да камяня, ляжаў <u>сьляпец</u> .
Ale kontrast między tym <u>wieprzem</u> a <u>młodzieńcem</u> , o którym myślał, bawił go.	But the contrast between the <b>man</b> before him and the <b>man</b> he had in mind was pleasant.	Аднак кантраст паміж створаным ім вобразам і <u>госцем</u> забаўляў яго.
Według niego oficerowie i <u>żołnierze</u> , jeśli nawet biologicznie należeli do tego samego rodzaju, w każdym razie stanowili zgoła odmienne gatunki.	To him officers and <b>men</b> might belong to the same biological genus, but they were of an altogether different species.	На яго думку, афіцэры і <u>радавыя</u> , хоць, магчыма, і належалі да аднаго біялагічнага роду, але адносіліся да розных відаў.
— Ach, ten <u>Weincheck</u> ! — odezwała się Leonora.	'That <b>man</b> !' said Leonora.	- Ах, гэты <u>Вайнчак</u> ! - усклікнула Леанора.
Czemuż <u>oni</u> nie zaczynają?	Why didn't the <b>men</b> begin?	Чаму <u>яны</u> не падыходзяць першымі, чаго яны чакаюць?

Table 4.38: Examples of 'man' translations in the EPB corpus.

Turning now to the next element of this study, that is the collocates of feminine nouns in Polish, it must be noted the procedure and settings were the same as in the case of English. First, all plural or singular occurrences of a word were searched. Next, the list of collocates is being created for words occurring within the window of five words to the left and five words to the right of the keyword, and occurring at least five times in such a combination. Finally, a list ordered by the MI score is analysed via close-reading to include only those collocates which appear in at least three different texts by three different authors.

The lists of collocates for the words *kobieta* and *kobiety* exhibit some differences compared to the list obtained for English. The table 4.38 contains the highest-ranked collocates in their basic form (singular male nominative):

looks	age	non-visual feature	other – verbs	other – nouns	other POS
tęgi (stout/big) chudy (skinny) szczupły (slim)	średni (middle) młody	nieszczęsny (unfortunate) naturalny (natural)	przedstawiać (present) znosić (endure) skinąć (nod)	typ (type) mężczyzna (man) kochanie (love)	jakiś (a)

ładny (pretty) siwy (grey-haired) piękny (beautiful) drobny (little/petite) nagi (naked) suknia (dress)	(young) stary (old) wiek (age)	miły (nice) głupi (stupid) mądry (wise) biedny (poor) niezwykły (extraordinary)	odrzec (respond) przyglądać (observe) krzyczeć (shout/scream) poznać (meet) posłuchać (listen) kochać (love) odezwać (speak) milczeć (be silent)	mąż (husband) kość (bone) dziewczyna (girl)	
--	--------------------------------------	--	---	---	--

Table 4.39. Collocates of the word 'kobieta' (EPB corpus).

Among the collocates prevail verbs, but there are also plenty of looks and inner traits descriptors. Adjectives referring to the non-visual features vary and can have negative connotations, such as *głupi* (stupid), positive connotations, as in *miły* (nice), *mądry* (wise) or *niezwykły* (extraordinary), they can be assessed differently depending on the context, and finally some of them might be deemed neutral. Many of significant adjectival collocations in categories of looks and age are the same or synonymous to those appearing in English. However, unlike in the original texts there are no comparatives when it comes to the descriptors of age. There is significantly more descriptors of the inner traits, as well as verbs and nouns collocating with the Polish *kobieta*. Regarding verbs, some of them, namely *przedstawiać* (to present), *poznać* (to meet) and *kochać* (to love), are used almost exclusively with a *woman* as an object of an action, whereas actions being performed by woman are nodding, responding, observing, screaming, listening, speaking and being silent. This is a wider selection than in case of English but most of them might be classified as describing passive behaviour.

The collocates of the plural form *kobiety* are presented in table 4.39:

looks	age	collectivity	other – verbs	other – nouns	other POS
piękny (beautiful) biały (white)	młody (young) stary (old) wiek (age)	niektóry(some) większość (most) oba (both) wszystek (all) trzy (three) dwa (two) cztery (four) wiele (many) kilka (a few)	rodzić (birth) czynić (do) ubrać (dress) nosić (carry/wear) należć (belong) kochać (love) znać (know) patrzeć (look) siedzieć (sit) chodzić (walk)	mężczyzna (man) dziewczę (girl) dziecko (child) towarzystwo (company) kobieta (woman) widok (view) chłopiec (boy) praca (work) świat (world) dom (house) życie (life) prawo (right)	wśród (among) wokół (around) obok (next to) także (too) u (at) lub (or) zawsze (always) dla (for) inny (other) bez (without) nad (over) taki (such)

Table 4.40. Collocates of the word 'kobiety' (EPB corpus).

The collocates of *women* in Polish exhibit many similarities with the English counterparts, in particular there are multiple verbs and nouns, a few indicating the domestic themes. There are many function words, in contrast to what is observed for the singular form. A prominent group of descriptors refers to the collective nature of the noun *kobiety*. There is very little collocates involving description of looks and age and none at all referring to the inner traits of a character. As regards verbs, they create a very similar pattern to those collocating with the singular noun, *kobieta*.

Many are used in sentences in which women are objects rather than subjects and they describe no extraordinary activities.

The table 4.40 shows the next component of this study, namely the collocates of the analogue of 'young female', namely the word *dziewczyna* (*girl*):

looks	age	non-visual feature	other – verbs	other – nouns	other POS
jasnowłosy (fair-haired) ładny (pretty) piękny (beautiful)  włos (hair) ramię (shoulder)	młody (young)	miły (nice) obcy (strange) wspaniały (wonderful)	mieszkać (live) podać (give) podejść (approach) grać (play) odejść (leave) opowiadać (tell) zwrócić (turn) kochać (love) unieść (raise/lift) poznać (meet/recognise) spać (sleep) siedzieć (sit) zapytać (ask) zauważyć (notice) spotkać (meet)	chłopak (boy) córką (daughter) imię (name) dziewczyna (girl) moment (moment) kobieta (woman) wzrok (sight)	ostrożnie (carefully) obok (next to) tamten (that) pewien (a) przy (next to) kiedyś (sometime)

Table 4.41. Collocates of the word 'dziewczyna' (EPB corpus).

The collocates of Polish *dziewczyna* exhibit many similarities to the English *girl*. Main differences between the two are high number of co-occurring verbs and low number of co-occurring descriptors of the inner traits of the character. Of the three descriptors of non-visual features two are overtly positive (*miły*, *wspaniały*) and one is neutral (*obcy*), however it can have negative connotations in some contexts. Among verbs collocating with *girl* only one does not occur on the parallel list for Polish, namely *learn*, where as *play* is one of the frequently collocating verbs in case of *dziewczyna*. Despite the differences in the amount and type of collocates there is no strong pattern that could be identified and would clearly distinguish *girl* from Polish *dziewczyna* in terms of discourse around those two concepts.

Table 4.41 presents the most significant collocates of the plural noun *dziewczyny*:

looks	age	collectivity	other – verbs	other – nouns	other POS	other POS – spatial
piękny (beautiful)	młody (young)	dwa (two) jeden (one)	powiedzieć (say) być (be) mieć (have)	czas (time)	który (which) z (with) jak (as/like) ten (this) co (what) a (but/and) to (it) o (about) się (self) tak (such) i (and)	za (behind) do (to) w (in) na (on)

					ja (I) nie (no) że (because) on (he)	
--	--	--	--	--	---	--

Table 4.42. Collocates of the word 'dziewczyny' (EPB corpus).

When comparing *girls* to Polish *dziewczyny*, one can see that both are most commonly collocating with function words, especially with prepositions indicating spatial position in case of the Polish version. There is very little attention paid to the looks or actions of *dziewczyny*, they usually appear as a subject of a conversation, rarely are they agents performing an activity or being the main focus of detailed description. There are also no descriptors of the non-visual features of girls.

All in all, the study shows little significant differences between Polish plural and singular nouns referring to adult and young females. Among those analysed words only one, *kobieta*, collocates frequently with words referring to the non-visual features of a character. The nouns denoting females in Polish are most frequently collocated with verbs, which suggests that *kobieta* and *dziewczyna* in Polish translations is more thoroughly scrutinised when it comes to what she does rather than how she looks like. When compared to English, there is also little collocations pointing to the discourse around the age and none referred to comparing the age.

Turning now to the masculine nouns, they attract slightly more collocates than the feminine counterparts, as can be observed in table 4.42:

looks	age	non-visual feature	other – verbs	other – nouns	other POS
krępy (squat) ciemnowłosy (dark-haired) rosły (tall) przystojny (handsome) chudy (skinny) szczupły (slim) wzrost (height) niski (short) wysoki (tall) nagi (naked) brudny (dirty) ubranie (clothes) spodnie (trousers) but (shoe) płaszcz (coat) broda (beard) włos (hair) szyja (neck) niebieski (blue) czarny (black)	dorosły (grown-up) średni (middle) młody (young) wiek (age) stary (old)	samotny (lonely) obcy (strange) prawdziwy (real)	rozpoznać (recognize) przebywać (stay) podejść (approach) obserwować (observe) siedzieć (sit) spotkać (meet)	kobieta (woman) skraj (edge) miłość (love) syn (son) spojrzenie (look) nazwisko (name)	uważnie (carefully) jakiś (a)

Table 4.43. Collocates of the word 'mężczyzna' (EPB corpus).

The list of collocates of Polish *mężczyzna*, contrary to the list obtained for English *man*, contains only a few non-visual feature descriptors, namely *samotny* (lonely), *obcy* (strange) and *prawdziwy*

(*real*), and all of these adjectives are regarded as neutral, even though they may be encountered in a negative context. What constitutes the biggest difference between the two lists is a high number of Polish collocates referring to the looks of *mężczyzna*. Many of the descriptors indicate the height or posture and describe clothing (or lack of it). Also the category of co-occurring verbs is more varied in Polish than in English. Unlike the discourse around English *man*, in the Polish part of EPB there are no hints of military themes connected to *mężczyzna*; some of the nouns might indicate rather the romantic theme (such as *kobieta*, *miłość* or *spojrzenie*). In many ways this set of collocates is much more similar to Polish *kobieta* than English *man* and may indicate a more significant difference between Polish and English discourse around males and females.

Table 4.43 contains collocates of the plural form *mężczyźni*:

looks	age	collectivity	other – verbs	other – nouns	other POS
niebieski (blue) twarz (face) ciało (body) usta (lips)	młody (young)	grupa (group)	ubrać (dress)	mundur (uniform)	gdyż (because)
<b>non-visual feature</b>		cztery (four)	zebrać (gather)	kobieta (woman)	wśród (among)
		oba (both)	nosić (carry)	ognisko (bonfire)	często (often)
prawdziwy (real)		większość (most)	rzucić (throw)	widok (sight)	ku (to)
		trzy (three)	wychodzić (leaving)	dziecko (child)	przy (next to)
		dwa (two)	rozmawiać (talk)	chłopiec (boy)	
		tłum (crowd)	siedzieć (sit)	stół (table)	
		wszystek (all)	parzyć (look)		
		wiele (many)	lubić (like)		
		jeden (one; mostly in phrase <i>one of many</i> )	czekać (wait)		
		ruszyć (take off)			
		słuchać (listen to)			
		zatrzymać (stop)			

Table 4.44. Collocates of the word '*mężczyźni*' (EPB corpus).

Similarly to the singular form *mężczyźni* collocates describing looks are more numerous than descriptors of the non-visual features, however in this case most of the collocates refer to the body and its parts, rather than the shape of it or the clothes. What is common for English and Polish is that verbs and the words pointing and the collective character of *men* and *mężczyźni* are prevailing, in case of Polish by a large number. In case of *mężczyźni* there are also more co-occurring nouns, one of them, *mundur*, might be associated with the military context, but most is connected with home and family themes.

Next, the collocates for Polish *boy* can be divided into categories shown in table 4.44:

looks	state or non-visual feature	other – verbs	other – nouns	other POS
chudy (skinny) blady (pale) mały (small) ramię (arm) twarz (face)	drogi (dear; also droga – it was incorrectly lemmatised due to homophony) biedny (poor)	nieść (carry) odrzec (respond) podać (give/offer) kochać (love) brać (take) pamiętać (remember) pokazać (show) zawołać (call) nazywać (name/call) siedzieć (sit) wejść (enter) odezwać (speak) prosić (ask/beg) dać (give) zrozumieć (understand) wziąć (take) powiedzieć (say) rzec (speak/say) spytać (ask)	dziewczynka (girl) chłopiec (boy) porządek (order) brat (brother) spojrzenie (look) matka (mother) mężczyzna (man)	łagodnie (gently) cicho (quiet) ów (this) jako (as) obok (next to) mój (my)

Table 4.45. Collocates of the word 'chłopiec' (EPB corpus).

Compared to English, Polish *chłopiec* collocates with similar adjectives referring to the looks and similar nouns. There is much less descriptors of the inner traits of the character but there are many verbs, in fact they dominate the list of words frequently co-occurring with *chłopiec*. Interestingly, many of them are used in citing what the character said – they are used mainly in past tense and positioned on the left side of the keyword, eg. *odrzekł/odezwał się/powiedział/rzekł/spytał chłopiec*. This is remarkable as none of previously discussed words exhibits such a trait. Among other parts of speech two adverbs are present, *łagodnie* and *cicho*. Both refer to the way a character spoke a particular piece of dialogue, additionally *cicho* occurs several times as part of command, such as *be quiet*.

The significant co-occurring words in the case of *chłopcy* (boys) have a different composition to the singular form, as presented in table 4.45:

looks	age	collectivity	other – verbs	other – nouns	other POS
mal (small; this is actually <i>mały</i> but incorrectly lemmatised) mały (small) ręka (arm) oko (eye)	młody (young) stary (old)	oba (both) trzy (three) kilka (a few) dwa (two) wszystek (all) wszyscy (all) wiele (many) jeden (one; usually in phrase <i>one of many</i> )	siedzieć (sit) zawołać (call)	dziewczę (girl) chłopiec (boy) matka (mother) mężczyzna (man) kobieta (woman) dom (home) chwila (moment)	wy (you; plural) inny (other) mój (my) dla (for) z (with) do (to) za (behind) przez (through) żeby (because) a (but) przed (in front of) na (on) siebie (self)
<b>state or non-visual feature</b>					
dobry (good) cały (whole)					

					ale (but) się (self) czy (if)
--	--	--	--	--	-------------------------------------

Table 4.46. Collocates of the word 'chłopcy' (EPB corpus).

There is a little difference between the collocates of English *boys* and Polish *chłopcy*. Descriptors of looks, age and state or non-visual features are similar both in number and in the content. There are many function words suggesting *chłopcy* being an object of conversation rather than the agents. More nouns than in case of English *boys* creates a slightly broader context though with no particular dominant theme. As in all other plural keywords discussed in this chapter *chłopcy* frequently co-occur with numerals and adjectives pointing at the collective nature of this noun.

In general, the male nouns in the Polish literary translations, that is *mężczyzna/mężczyźni* and *chłopiec/chłopcy* do not collocate with words that could be assigned to one theme, such as in the case of English *man*. Unlike in the original texts, the most significant collocates of *mężczyzna* are predominantly descriptors of his looks rather than the inner traits. Moreover, the adjectives that refer to the non-visual features of *mężczyzna* are less varied than in case of *man*.

The features of discourse around *mężczyzna* is much more similar to the features visible in the collocates of *kobieta* rather than *man*. The gender differences are not prominent also in the plural form, *mężczyźni*, which collocational patterns resemble those characteristic of *kobiety*, namely high number of verbs and nouns, some of which could be associated with domestic themes. The nouns referring to children, *chłopiec* and *chłopcy*, exhibit many similarities both to the English analogues and Polish female counterparts (*dziewczyna/dziewczyny*). One feature common of all Polish words analysed in this chapter is the presence of strictly abstract concepts among noun collocates, such as *time, type, life, love, moment*. This is not characteristic of the collocates obtained from the English part of the EPB corpus.

In order to fully understand the validity of the results obtained for the translational data, this analysis looks also in the original language. Considering the accuracy of replication, the corpus chosen for comparison was the one that can be easily accessed via KonText application, namely InterCorp. As mentioned in earlier chapters, InterCorp is a multilingual parallel corpus, therefore it contains both original and translational data, however it is possible to restrict the search to fiction and to the original Polish data. The size of dataset in the original Polish language (InterCorp v13) is over 3.9 million words, most of which is fiction. Within this dataset the collocation study with exactly the same parameters as in previous examples has been repeated. The results for word *kobieta* are presented in the table 4.46:

looks	age	state or non-visual feature	other – verbs	other – nouns	other POS
suknia (dress) uroda (beauty) płaszcz (coat) spodnie (trousers) piers (breast) kolano (knee)	młody (young) trzydzieści (thirty) stary (old)	mądry (wise) samotny (lonely) jedyńy (only)	rodzić (birth) ubrać (dress) wołać (call) uśmiechać (smile) płakać (cry) biec (run)	mężczyzna (man) wieś (village) kobieta (woman) świat (world) widok (sight) dziecko (child)	obok (next to) zapewne (probably) jakiś (a) tamten (that)

włos (hair) ciało (body) twarz (face) ubrany (dressed) męski (male) piękny (beautiful) gruby (fat) czarny (black)			siedzieć (sit) podnieść (lift)	stół (table) list (letter) cóрка (daughter) łóżko (bed) mąż (husband)	
--	--	--	-----------------------------------	---	--

Table 4.47. Collocates of the word 'kobieta' (InterCorp).

The dominant category of significant collocates are descriptors of looks, especially referring to body and its parts, as well as the garments. There are only three adjectives that might be associated with the overall assessment of the looks (*uroda*, *piękny* and *gruby*). The list of collocated contains many nouns among which similar can be associated with domestic theme. The adjectives describing the inner features are sparse; two of them can be assessed as neutral and one has positive connotations. These set of features is more similar to those of *woman* rather than *kobieta* in the EPB corpus.

Next, the most significant collocates of the word *kobiety* are showcased in table 4.47:

looks	age	collectivity	other – verbs	other – nouns	other POS
chustka (babushka) suknia (dress) włos (hair) ramię (arm) twarz (face) ręka (arm) kobięcy (female) nagi (naked) piękny (beautiful) biały (white)	stary (old) młody (young) wiek (age)	kilkanaście (about a dozen) kilka (a few) wszystek (all) trzy (three) oba (both)	krzyczeć (scream) płakać (cry) oglądać (watch) umieć (can) siedzieć (sit) pisać (write) patrzeć (look) pić (drink) traktować (treat) lubić (like) zwrócić (return) szukać (search)	mężczyzna (man) starzec (old man) tłum (crowd) dziecko (child) koń (horse) kobięta (woman) rozmowa (conversation)	zresztą (anyway) wśród (among) ogół (part of w ogóle; in general) bo (because) często (often) zupełnie (entirely) u (at) tyle (so many) ile (how many) przecież (but) albo (or) jakiś (a) podobny (similar)

Table 4.48. Collocates of the word 'kobięty' (InterCorp).

In case of *kobięty* the dominant category of collocates are function words, however looks descriptors and verbs are almost equally well represented. Similarly to the singular *kobięta*, the words referring to the looks are predominantly body parts, some of them refer to the clothes and only two might be classified as the general descriptors of the appearance. There are no collocates referring to the non-visual features of *kobięty* and very little referring to the age. Just as in the case of the singular form, the plural exhibits more significant similarities to the English *women*.

All in all, the analysis of the collocates of *kobięta* and *kobięty* in the original Polish texts proves the higher degree of similarity between these words and the original English *woman* and *women*. The discourse around *kobięta* and *kobięty* in the Polish translations is significantly different from the discourses present in the original texts.

Next component of this study, the word *dziewczyna* and its most significant collocates are presented in the table 4.48:

looks	age	non-visual feature	other – verbs	other – nouns	other POS
sukienka (dress) ładny (pretty) plecy (back) piękny (beautiful) ramię (arm) czerwony (red) palec (finger) włos (hair) twarz (face) biały (white) ciało (body) noga (leg) ręka (arm) oko (eye) głowa (head) wysoki (tall)	młody (young)	miły (nice)	chwycić (grasp) patrzeć (look) siedzieć (sit) czytać (read) pytać (ask) patrzeć (look) spojrzeć (look/glance) leżeć (lie) czekać (wait) wejść (enter) wyjść (leave) stać (stand) wziąć (take)	chłopiec (boy) wzrok (sight) dziewczyna (girl) koń (horse) kobieta (woman) chwila (moment)	gdyż (since/because) wobec (towards) jakiś (a) przed (in front of/before) aż (so/until) bo (because) gdy (when) nad (above)

Table 4.49. Collocates of the word 'dziewczyna' (InterCorp).

The dominant category among collocates of *dziewczyna* are descriptors of the outer appearance, represented mostly by words referring to her body and its various parts. A significant part of the collocates are also verbs and functions words. There are some nouns collocating with *dziewczyna*, however with no prevailing topic. There is just one descriptor of age (*młody*) and one of the non-visual feature with positive connotations (*miły*). No definite conclusions can be made as to the similarity between *dziewczyna* in the InterCorp and its analogues analysed so far, some features of the discourse characterised via the use of collocates correspond to *dziewczyna* in the translational texts and some to *girl* in original texts.

Table 4.49 showcases the collocates of the plural form *dziewczyny*:

age	collectivity	other – verbs	other – pronouns	other POS
młody (young)	dwa (two) wszystek (all) jeden (one; usually in the phrase <i>one of many</i> )	ić (go) widzieć (see) mówić (speak) mieć (have) być (be)	żaden (nobody) siebie (self) jakiś (a) swoj (one's own) się (self) który (which) ten (this) on (he) ja (I)	nad (above) jeszcze (still) raz (one) za (behind) po (after) a (but/and) z (with) kiedy (when) jak (how/as) bardzo (very) już (already) do (to) i (and) na (on) o (about)
		<b>other – nouns</b>		
		chłopak (boy) dziewczyna (girl) dom (house)		

				w (in) że (because) co (what) tak (such) nie (no/not)
--	--	--	--	---

Table 4.50. Collocates of the word 'dziewczyny' (InterCorp).

One look at the summary of *dziewczyny* collocates is enough to see that the function words, especially pronouns, definitely prevail in the discourse around the analysed keyword. There are several collocates referring to the collective nature of the noun, a few verbs and nouns, only one collocate denoting age and no descriptors of looks or the inner traits of the character. Such a set of collocates makes *dziewczyny* in the original Polish texts slightly more similar to *girls* in the original English texts than *dziewczyny* in the translational corpus.

All in all, both *kobieta* and *dziewczyna* (singular and plural) bear more resemblance to their analogues in the original texts than in the translational. This might suggest a systemic difference between originals and translations in the matter of gender discourse, however to confirm such assumptions it is necessary to analyse the collocations of the male counterparts. Table 4.50 presents the significant collocates of the word *mężczyzna* in the InterCorp corpus:

looks	state or non-visual feature	age	other – verbs	other – nouns	other POS
postawny (stout) szczupły (slim) rys (feature/line; as in <i>facial features</i> ) przystojny (handsome) marynarka (jacket) koszula (shirt) ubranie (clothes) czarny (black) twarz (face) zielony (green) biały (white) wysoki (tall) niski (short) ramię (shoulder) ręka (arm) głowa (head) jasny (fair) piękny (beautiful)	obcy (strange) prawdziwy (real)	dorosły (adult) młody (young) wiek (age) stary (old)	podejść (approach) przerwać (interrupt) patrzeć (look) pytać (ask) stawać (become) poznać (recognize) leżeć (lie)	mundur (uniform) zwłoki (corpse) kobieta (woman) mężczyzna (man) chłopiec (boy) żona (wife) historia (story/history)	zapewne (probably) obok (next to) zupełnie (at all) tamten (that) jakiś (a) ów (this) przecież (but) choć (however)

Table 4.51. Collocates of the word 'mężczyzna' (InterCorp).

Descriptors of looks, mainly general references to the appearance prevail among the collocates of *mężczyzna*. There are only two collocates referring to the non-visual features, one of them positive and the other one neutral or negative depending on the context. There are several descriptors of age

and a handful of verbs, functions words and nouns, among which some might point at the military or crime context (*mundur* and *zwłoki*). Taking into account the proportion of various types of collocates it might be concluded that the discourse around *mężczyzna* in the original Polish texts resembles discourse around the same noun in the Polish translations of English literature.

Table 4.51 contains the significant collocates of the word *mężczyźni*:

looks	age	collectivity	other – verbs	other – nouns	other POS
biały (white) głowa (head) oko (eye) wysoki (tall)	młody (young) stary (old)	tłum (crowd) dwa (two) większość (most) grupa (group) trzy (three) kilka (a few) wszystek (all)	pić (drink) siedzieć (sit) ciągnąć (pull) opowiadać (tell) rozmawiać (talk) lubić (like) stać (stand) czekać (wait) znać (know) zacząć (start)	mundur (uniform) kobieta (woman) papieros (cigarette) stół (table) mężczyzna (man) krok (step) samochód (car) dziecko (child) rodzina (family) ojciec (father)	wśród (among) często (often) u (at) przy (next to) przecież (but) między (between) potem (later) nawet (even) lecz (however) niż (than) pod (under) nad (above) dlaczego (why) jakiś (a) siebie (self) sam (self)

Table 4.52: Collocates of the word '*mężczyźni*' (InterCorp).

The collocates of *mężczyźni* are more varied than those of *mężczyzna*, however there are some slight disproportions between particular groups. The most common significant collocates are function words, but there are also many nouns, verbs and words referring to the collective nature of the noun. Descriptors of looks and age are least frequent and descriptors of the inner traits of the character are absent. Taking into account the proportions between particular types of collocates it might be concluded that *mężczyźni* in the original Polish texts are more similar to *mężczyźni* in the translational texts than in English originals. This conclusion corresponds to what was found in the case of the singular form of the noun.

Next, table 4.52 presents significant collocates of the word *chłopiec*:

looks	age	other – verbs	other – nouns	other POS
mały (small) oko (eye) twarz (face) ręka (hand) głowa (head)	młody (young) wiek (age) stary (old)	próbować (try) siedzieć (sit) znać (know) widzieć (see) stać (stand) mówić (speak)	chłopiec (boy) dziewczynka (little girl) oficer (officer) dziewczyna (girl) szkoła (school) mężczyzna (man) ulica (street) ojciec (father) mama (mum) matka (mother) kobieta (woman) dziecko (child)	znowu (again) przecież (but) dziś (today) wtedy (then) jako (as) jakiś (a) albo (or) gdy (when) między (between) który (which) dlaczego (why) też (also)

			chwila (moment)	bardzo (very) ale (but) żeby (to/for)
--	--	--	-----------------	---

Table 4.53: Collocates of the word 'chłopiec' (InterCorp).

The collocates of the word *chłopiec* in the original Polish texts consist mostly of function words and nouns, accompanied by some verbs and descriptors of looks and age. There are no words referring to the non-visual features of a character. Most of the co-occurring nouns denote people and one of them is particularly surprising, namely *oficer*. Below we can find the broader context of all the hits of this collocation:

Hits: 5   i.p.m.: Calculate ⓘ related to the subset defined by the selected text types   ARF: 1.36   Result is sorted					1 / 1
Line selection: simple ▾					
<input type="checkbox"/>	Gombrowicz-Kosmos	w sensie berga . Był em <b>oficerem</b> sztabu generalnego .	<b>Chłopcem</b>	służącym do mszy . Kornym i karnym akolitą i wykonawcą	
<input type="checkbox"/>	Chwin-Hanemann	, Hanemann opowiedział mi o chłopcu - tak , o	<b>chłopcu</b>	w mundurze pruskiego <b>oficera</b> - zranionym , skłóconym ze wszystkimi	
<input type="checkbox"/>	Chwin-Hanemann	pruskiego oficera , który nie chciał być <b>oficerem</b> , głos	<b>chłopca</b>	, który został opuszczony przez wszystkich : „ ... znalazł	
<input type="checkbox"/>	Chwin-Hanemann	Cóż to pan J . mówił wtedy o Kleiście ?	<b>Chłopiec</b>	w mundurze pruskiego <b>oficera</b> , o oczach , w których	
<input type="checkbox"/>	Kusniewicz-KralObSic	się potem z suchym szelestem , prosząc pyłem . Mały	<b>chłopiec</b>	cygański idzie za <b>oficerami</b> ostrożnie , trzyma palec w ustach	

Illustration 4.18: Word 'chłopiec' collocating with 'oficer' (InterCorp).

These concordance line shows that actually, in most cases, *chłopiec* is the same person as *oficer*, eg.:

“I was the officer of the general staff. I was the altar boy.”

“[...] about a boy in the uniform of the Prussian officer [...]”

“[...] the voice of [...] officer who didn’t want to be an officer, the voice of a boy who has been abandoned by everyone [...]”

In most of the cases it is a part of someone’s reflection on the past, it does not imply boys being part of an army or boys being talked about in the military context.

Finally, table 4.53 contains the collocates of the plural form, that is *chłopcy*:

looks	age	collectivity	other – verbs	other – nouns	other POS
nagi (naked) noga (leg) mały (small) ręka (arm)	młody (young) stary (old)	dwa (two) wszystek (all) jeden (one)	iść (go) widzieć (see) chodzić (walk) robić (do) stać (stand) mówić (talk)	dziewczę (girl) dziewczynka (little girl) ulica (street) chłopiec (boy)	razem (together) pod (under) gdy (when) żeby (to) siebie (self) my (we) też (also) tylko (only, just) z (with) przez (through, via) a (but, and) od (from)

					i (and) przed (in front of, before) za (behind) jaki (which) już (already) jak (as, how) nasz (our) na (on, at) czy (if) coś (something) się (self) wszystko (all)
--	--	--	--	--	---

Table 4.54: Collocates of the word 'chłopcy' (InterCorp).

In the case of *chłopcy* in the original Polish texts there is a strong preference for collocations with function words, other categories of the collocates occur in more or less the same number and there is no prevailing thematic group. Both in the case of *chłopiec* and *chłopcy* the collocational pattern is most similar when comparing English and Polish part of the EPB corpus, the discourse around these words in the original Polish texts differs from the English originals and their translations.

All in all, when analysing the collocation patterns in the EPB and InterCorp corpora, both quantitatively and qualitatively, few conclusions can be drawn as to the differences in gender discourse in the English and Polish languages. Firstly, in three out of four cases, that is *woman*, *girl* and *boy* (singular and plural) the collocational behaviour of the keyword in the translate language is distinctive, both from what can be found in the original Polish language and in the original English language. Only the translation of *man/men* can be regarded as emulating the characteristics of the original Polish language. What it means is that all Polish translations fail to convey the gender discourses around men and women, at least based on the EPB corpus material and the collocational analysis of the four lemmas referring to the two genders.

#### 4.4. Gender discourse in Belarusian translational data and in contrast with monolingual corpus.

Kłyška's dictionary (1993) mentions the following 9 synonyms for the Belarusian word *жанчына* [žančyna] (*woman*); *кабета*, *кабеціна*, *баба*, *цётка*, *маладзіца*, *маладуха*, *малодка*, *дама*, *мадам*. Among these one word cannot be taken into account. The primary meaning of *цётка* is *aunt* and without semantic annotation (which is absent from the EPB corpus) there is no easy way of telling apart the occurrences in one or the another sense. In addition to the synonyms listed by Kłyška, a Belarusian analogue of *girl/girlfriend* – *дзяўчына* [dziaŭčyna] – though not mentioned in the dictionary, is included in the search. For *мужчына* [mužčyna] (*man*) the 7 synonyms are as follows: *хлопец*, *мужык*, *дзядзька*, *дзядок*, *кавалер*, *жаніх*, *братка*. Table 4.54 contains the results of querying the Belarusian part of the EPB corpus:

Number of occurrences	Feminine nouns	Masculine nouns
1680	жанчына	
832	дзяўчына	
768		мужчына
242		хлопец
172	дама	
56		дзядок
29	баба	
28	кабета	
28		мужык
17		кавалер
11		братка
8		жаніх
5	кабеціна	
3	маладзіца	
2	маладуха	
<b>IN TOTAL</b>	<b>2751</b>	<b>1130</b>

Table 4.55. Lemmas 'жанчына', 'мужчына' and their synonyms (EPB corpus).

In the EPB data there were no hits for *малодка*. Additionally *мадам*, which is used almost exclusively as a form of address, has been excluded from the search. Regarding the number of occurrences there are two key findings. First, the overall number of occurrences is lower than in English and Polish. This is likely caused by two factors. Firstly, the grammatical features, for example usage of single words with broader meaning, such as *стара* instead of *old woman*, and secondly poor lemmatisation (that aspect will be elaborated on later in this chapter). Second finding of the synonyms query is that the number of *woman* synonyms in the Belarusian part of the EPB corpus is over twice as high as the number of *man* synonyms. That, in turn, could be partially explained by the choice of synonyms, it is likely the Kłyška's dictionary does not include a few that are common in the analysed corpus. This line of investigation will not be followed in this thesis.

Finally, the Belarusian component of the EPB corpus has been tested in the same manner as the Polish subcorpus. The settings used for collocation queries were as follows: attribute – lemma, search in the range from -5 (five to the left) to 5 (five to the right), minimum frequency in corpus – 5, minimum frequency in given range – 5, sorted by MI score. There was, however, one significant difference. Due to technical problems<sup>8</sup> it was not possible to fully tag the Belarusian component and therefore it was impossible to query it for singular and plural forms separately. As long as the main aim of this chapter is to ascertain the differences between the genders it has been concluded the analysis of Belarusian component can omit the singular versus plural characteristics. The results for the word *жанчына* (*woman*) are presented in table 4.55:

<sup>8</sup> These were problems on the side of CLARIN-PL consortium and they had not been resolved upon this thesis completion.

looks	age	state or non-visual feature	other – verbs	other – nouns	other POS
маленька (маленькі; tiny) хударлявы (scraggy) высокі (tall) сукенка (dress) худы (skinny) прыгожы (beautiful) маленькі (tiny) голы (naked) валаса (волас; hair)	маладаваць (малады; young) пажылы (elderly) маладой (малады; young) малада (малады; young) пяцідзесяць (fifty) сарак (fourty) век (age) маладзейшы (younger) дарослы (adult) малады (young) трыццаць (thirty)	мудры (wise) фізічны (physical) багаты (rich) бедны (poor) здоровы (healthy)	падыхаць (падыхаць; approach) дай (даць; give) ішла (ісці; go) кахаць (love) сядзіць (sit) засмяецца (laugh) крычаць (scream) паказваць (show)	Мужчына (мужчына; man) мужчын (мужчына; man) мужчына (man) дзець (дзіця; child) становішча (state) каханне (love) дзіця (child)	така (такі; such) нейкі (a) (ні) адной (адзін; no) двара (двор, as in у двары; outside, rarely court) <b>collectivity</b> абедва (both) многія (many)

Table 4.56. Collocates of the word 'жанчына' (EPB corpus).

The list of collocates is full of incorrectly lemmatised words. In twelve cases the correct lemma has been added in the brackets together with English translation. This accounts for some of the words being repeated. When taking into account only unique lemmas it becomes clear that the dispersion of collocates among various semantic and grammatical groups is balanced, no category dominates when it comes to the collocates of the word *woman* in Belarusian translational language. The most common collocates refer to the looks – especially the posture, to the age, with an instance of age comparison, and to the activity. Verbs collocating with *жанчына* occur predominantly in their active form, rather than passive, the one exception to that is the word *даць* (to give) which is used exclusively as a predicate following a subject other than *жанчына*.

Among the frequently collocating nouns there is no predominant thematic pattern. Among the descriptors of the non-visual features there are mostly words of positive connotations and one of them, *physical*, never refers to *woman* specifically, rather it describes other concepts as physical, for example *connection*, *pleasure*, *challenges* or *proximity*. Many of the collocations containing the word *physical* appear in sexual context.

Collocates of the word *дзяўчына* in the EPB corpus are less balanced, as presented in table 4.56:

looks	age	state or non-visual feature	other – verbs	other – nouns	other POS
сукенка (dress) высокі (tall) прыгожы (beautiful) валаса (волас; hair)	малады (young) стары (old)	разумны (clever) бедны (poor) добры (good)	ішла (ісці; go) сустрэкаць (meet) падыхаць (падыхаць; approach) засмяецца (laugh) страціць (lose) павярнуцца (turn) падабацца (like)	юнак (young man) незнаёмы (stranger) мужчын (мужчына; man) здзіўленне (surprise)	побач (next to) адной (адзін; a) адна (адзін; a) май (мой;

твара (твар; face) твар (face) голос (voice)			спытаць (ask) сядзець (sit) знайсці (find) выйсці (leave) гаварыць (talk) прайсці (cross, go) глядзець (look) спыніцца (stop)	дзяўчына (girl) жонка (wife) імя (name) спіна (back) мужчына (man) жанчына (woman)	my) такі (such) ўвесь (all) сапраўды (really)
---	--	--	--	---	---

Table 4.57. Collocates of the word 'дзяўчына' (EPB corpus).

The main focus in the discourse around *girl* in the Belarusian part of the EPB corpus is on her activities. The main part of the list of collocates comprises of verbs, most of which occur in their active forms and refer to the activities performed by *girl*. Activities denoted by these collocates are diverse, but four of them describe a type of walk.

All in all, the discourse around female characters in the Belarusian translational language is different from what can be observed in the case of their English counterparts because there is much less focus on the looks, rather the analysed keywords co-occur with verbs or with various types of words but in a more balanced way. It might suggest a more depth to the description of female characters in the Belarusian translations as opposed to English originals.

In terms of masculine keywords and their collocates in the Belarusian part of the EPB corpus, they can be classified in the following manner:

looks	age	state or non-visual feature	other – verbs	other – nouns	other POS
рост (height) высокі (tall) валаса (волас; hair) твар (face) цела (body) выгляд (appearance)	пажылы (elderly) старэйшы (older) трыццаць (thirty) малады (young)	мажны (brave) сапраўдны (real) моцны (strong) прыгожы (beautiful) адзіны (only) іншы (different) падобны (similar)	сядзіць (сядзець; sit) належыць (belong) сядзець (sit) кінуць (throw) выйсці (leave) чакаць (wait) пачуць (feel) павярнуцца (turn) павінен (should) заўважыць (notice) любіць (like, love)	жанчын (жанчына; woman) жанчына (woman) мужчына (man) хлопчык (little boy) дзяўчына (girl) <b>collectivity</b> ўсе (yсе; all)	ніводзін (no) Гэта (гэта; it) напрыклад (for example) любы (any) паміж (between) нейкі (a) побач (next to) адзін (a) другі (other) пры (next to) сярод (among) такі (such) вядома (evidently) напэўна (surely) кожны (every) які (which)

Table 4.58. Collocates of the word 'мужчына' (EPB corpus).

Table 4.57 contains the collocates among which the functions words and verbs constitute the most prominent part. There is a handful of looks and inner features descriptors, all of which are either positive or neutral, and only a few collocates referring to age or co-occurring nouns. This is the opposite of what has been found for the collocates of *man* in the English part of the EPB corpus.

Finally, table 4.58 lists the most significant collocates of the word *boy* in the Belarusian translational language:

looks	state or non-visual feature	other – verbs	other – pronouns	other POS
прыгожы (beautiful)	неблаг (неблагі; not bad) добры (good)	мог (магчы; be able) адказаць (respond) пачаць (start) сказаць (say) быць (be)	нешта (something) той (that) яго (his) гэта (it) свой (own) ты (you) ён (he) тое (it) які (which) гэты (this) яны (they) яна (she)	ўжо (ужо; already) з (with) што (what) як (how, when) так (such) а (and) за (behind) на (on, at) да (to) калі (when) і (and) яшчэ (yet) ў (y; in, at) ж (particle, untranslatable) не (no) я (I) у (in, at) толькі (only) але (but)

Table 4.59. Collocates of the word 'хлопец' (EPB corpus).

It is clear from the table that function words, especially pronouns, dominate the characteristics of the discourse around *boy* in the Belarusian part of the EPB corpus. It is not *хлопец* who is the main agent in most of the sentences where he is mentioned, rather he is being referred to when someone else is speaking.

Following the procedure from chapter 4.3. this section moves on now to the analysis of Belarusian original language. In order to obtain the most reliable results this analysis will be based on the InterCorp corpus, as in the case of Polish original language. The Belarusian part of InterCorp (version 13) contains over 1,3 million tokens, vast majority of which is fiction writing. Settings for collocation search stayed the same as those used for querying the EPB corpus.

Table 4.59 displays the results obtained for the word *жанчына* in the original Belarusian language:

looks	age	collectivity	other – verbs	other – nouns	other POS
постаць (silhouette) прыгожы (beautiful) твар (face) чорны (black) рука (arm) воча (вока; eye) галава (head) рук (рука; arm)	малады (young) стары (old) год (year)	дзва (два; two)	стаесці (стояць; stand) убачыць (see) сядзець (sit) бачыць (see) відаць (see) глядзець (look) чуць (hear) зразумець (understand) думаць (think)	дзець (дзіця; child) мужчына (man) жанчына (woman) дом (house) чалавек (man, human)	гэтой (гэты; this) адна (адзін a, one) нейкі (a) такі (such) гэты (this) другі (other) яе (her) яна (she) добра (well, right) вось (well) перад (in front of)

					які (which) ды (and) ўжо (уже; already) ў (y; in) з (with) на (at, on) той (that) але (but)
--	--	--	--	--	--

Table 4.60: Collocates of the word 'жанчына' (InterCorp).

Function words dominate among the collocates of *woman* in the original Belarusian language, followed by verbs and descriptors of the looks. There are no common collocates referring to the non-visual features or state of a character when it comes to *жанчына*. Such a set of collocates is quite different to what was found for English *woman* and its Belarusian analogue in the translational language.

The difference between the original and translational Belarusian is more overt in the case of *дзяўчына* collocations presented in table 4.60:

other – verbs	other – pronouns	other POS
думаць (think) мець (have) трэба (must) ведаць (know) сказаць (say) быць (be)	нешта (something) гэты (this) яна (she) свой (own) мы (we) той (that) я (I) тое (that) ён (he)	побача (побач; next to) вельмі (very) раптам (suddenly) больш (more) такі (such) да (to) пасля (after) увесь (all) з (with) ды (and) на (on) ў (y; in, at) а (and) па (on) так (such) і (and) не (no) калі (when) пра (about) які (which) як (how, when) яшчэ (yet) у (in, at) што (what) ж (particle, untranslatable) але (but) за (behind) ад (from) толькі (only)

Table 4.61: Collocates of the word 'дзяўчына' (InterCorp).

*Girl* in the corpus of original Belarusian language collocates almost exclusively with function words, an exception to this are a few verbs. This indicates *дзяўчына* is marginalised in the texts comprising the Belarusian part of InterCorp, and the discourse around *girl* in Belarusian is very much different to what can be found in English literature of the 20<sup>th</sup> and 21<sup>st</sup> century or its Belarusian translations.

Next, the male keywords in the Belarusian original language are discussed, starting with *мужчына*:

looks	age	state or non-visual feature	other – verbs	other – nouns	other POS
чорны (black) твар (face) воча (вока; eye)	малады (young) малада (малады; young) год (year)	здоровы (healthy) сапраўдны (real)	сядзець (sit) быць (be) сказаць (say)	жанчына (woman) мужчына (man) чалавек (man, human)	ля (next to) адзін (a) перад (before) нейкі (a) тады (then) больш (more) ці (if, whether) усё (all) як (when) ўжо (ужо; already) і (i; and) пад (under) увесь (all) яшчэ (yet) а (and) нават (even) але (but) ў (y ;in, at) ад (from) з (with) толькі (only) ж (particle, untranslatable) гэты (this) да (to) па (on) які (which) за (behind) не (no) і (and) у (in, at) што (what)
			collectivity	other – pronouns	
			два (two)	яна (she) ты (you) той (that) ён (he)	

Table 4.62: Collocates of the word ‘мужчына’ (InterCorp).

The collocates listed in table 4.61 contain a handful of looks, inner features and age descriptors, alongside a few verbs and nouns, but the dominant category of co-occurring lemmas are function words. This is a recurrent feature of all keywords queried in the Belarusian InterCorp, as evident from table 4.62:

verbs	pronouns	other POS
быць (be)	гэты (this) ты (you) свой (own) той (that) ён (he) яна (she) я (I) гэта (it)	добра (well) і (i; and) ужо (already) за (behind) ад (from) так (such) ці (if, whether) па (on) а (and) з (with) як (as, how) ж (particle, untranslatable) і (and) але (but) які (which) у (in, at) не (no) на (on, at) да (to) што (what) ў (y; in, at)

Table 4.63: Collocates of the word 'хлопец' (InterCorp).

Except one verb the word *хлопец* collocates exclusively with function words in the original Belarusian language. Similarly to its female counterpart, *дзяўчына*, there is almost no diversity to the discourse around this concept in the dataset.

To sum up, all three datasets compared in this chapter, that is original English texts, Belarusian translations and Belarusian original texts, differ substantially. The collocational patterns of the four keywords in the original Belarusian language is very different to those which can be observed in the English language data. The Belarusian translations, based on the collocational analysis, neither retain the gender discourses of the English original texts nor they emulate the discourses typical for the original Belarusian language. This is common for Belarusian and Polish translations and might be indicative of translational language distinctiveness on the discourse level.

All in all, Chapter 4 comprised the analysis of gender discourse exemplified by the frequencies of some masculine and feminine words in the corpus, as well the collocates of the words *woman*, *women*, *man*, *men*, *girl*, *girls*, *boy* and *boys*. The investigation showed that the gender differences are more prominent in the English originals than the Polish and Belarusian translations.

This collocational analysis of the discourse around certain concepts raises the question of the reasons behind the Polish and Belarusian translations failing to retain the collocates of the investigated words and the clear-cut differences between genders, even though the collocational patterns obtained for the original texts in those languages bear quite a few similarities to the original English texts. There are at least two explanations of this phenomenon. It is either a conscious and consistent choice of certain solutions by the translator, or a choice that is unintentional and

motivated rather by the availability of certain equivalents in the target language. In other words, a word frequently co-occurring with another word in English can usually be translated in more than just one way. Each solution is less frequent than the original and therefore it may be not recognised as a collocate in the target text. Determining whether that is the case or not would require a more holistic look at the data, for example via analysis of semantic categories of frequently co-occurring words.

The analysis of collocates performed in Chapter 4 is by no means an exhaustive exploration of gender discourses in English, Polish and Belarusian. Rather, it is and should be treated as a starting point for a broader discussion and deeper investigation of all its aspects. The queries can be supplemented with a full (or near full) range of words referring to various genders. The keywords might be analysed in a more complex way than only via their collocational patterns. Finally, the analysed texts could be split into groups based on the sociolinguistic features to yield even more detailed results, as mentioned in Chapter 3.2.

## Chapter 5: Looking ahead – the planned direction of the EPB corpus development

Discussing the possibilities of the EPB corpus development cannot be approached in isolation from the decisions regarding corpus maintenance that have been already made, as they largely predefine what choices are available in the future.

Firstly, a number of issues were tackled at the level of corpus design. For the reasons discussed in detail in Section 2.1, the focus of the EPB corpus is on 20<sup>th</sup> and 21<sup>st</sup> century prose in three languages, namely English, Polish and Belarusian. For the purpose of restricting the amount of data and keeping the high level of quality of the texts (as they are carefully preprocessed) the governing principle is to exclude texts existing in only two of the three above-mentioned languages. Additionally, as this corpus is designed primarily for translation studies purposes, one of the languages has to be the source language, that is texts existing in English, Polish and Belarusian are not included if they are all translations from any fourth language.

For the purposes of exemplifying the practical uses of the EPB corpus the initial portion of 113 titles has been collected and processed to allow linguistic analysis. The database includes short stories, novellas, novels, and, in some cases, fragments of novels. This concerns some texts in Belarusian, as it often happens that excerpts of translations are published in renowned sources, such as *PrajdziSviet* journal. The purpose of that is to attract a potential publisher or to encourage supporters in the process of collecting financial means via crowdfunding platforms.

Therefore the first step of the EPB corpus further development is keeping the database up to date and successfully adding new titles emerging in the publishing markets of interest, as well as adding information about published full editions of novels which are already included in the corpus, however partially. This concerns also translations from Polish and from Belarusian into the other two languages, because the ultimate aim is to change the EPB corpus into a multidirectional corpus. A database containing translations not only from English into Slavic languages, but also the other way around, as well as translations between two Slavic languages from different groups would allow to investigate interesting phenomena, such as methods of translating language with strongly embedded gender characteristics or ways of approaching false friends, especially in closely related languages, such as Polish and Belarusian. Additionally, a closer look at the metadata might reveal mechanisms behind choosing the minority literature for export, especially to the Anglophone book market.

After the initial completion of text collection in mid-2019, a handful of additions were already made and the database is expected to be constantly growing. What is more, the list of corpus primary sources, as well as scheduled additions, is made publicly available<sup>9</sup> in order to let the users report titles that might have been omitted by the author.

Another issue discussed in the design of the corpus are legal conditions of publishing resources, such as the EPB corpus. Due to the current state of the EU regulations only access for the purposes of research and education can be granted. As already mentioned in Chapter 2.1, the

---

9 The project's website: <https://epbcorpus.wordpress.com/>

ultimate aim of the project is to make the corpus available to the general public (at least partially), however to do that, access to the copyright is needed, and publishers are reluctant to give permission to use the data, even in restricted contexts with no financial gain. Attempts have been made to contact some Polish publishing houses, however the response has been deeply unsatisfactory<sup>10</sup>. This is a common issue and it is a known fact that usually only big national corpora with a strong institutional support contain copyright protected works; other resources are either restricted to a very small research team or comprise only public domain literature, that is titles from the 19<sup>th</sup> century or older.

In these circumstances the first choice for the EPB corpus circulation is an infrastructure that supports institutional login, which would in turn serve as the first level of restricting the access to the data. Currently, the author is cooperating with the Polish branch of CLARIN (discussed in Section 1.4). The institutional login in CLARIN-PL is secured by a third party, namely identity federation Pionier which is dedicated for the academic community. So far (by the end of 2019) 25 universities and research institutions have joined Pionier (Polish Identity Federation, 2014), among them colleges from 15 out of 16 administrative regions in Poland, and 7 out of the 10 biggest universities in the country (Statistics Poland, 2019). These seven universities alone amount to over 200 thousand students, that is almost 17% of all Polish students.

CLARIN offers access to the data to students and researchers, and not only in Poland but also in Europe and, partially, other continents. To date, twenty one European countries have become full members of the infrastructure, four more (three European and one African) participate as observers, and the Carnegie Mellon University from the USA is a third party member. In summary, collaboration with CLARIN enables restricting access to the data to the academic community, while still authorising a high number of users due to its wide-spread presence, mainly (but not exclusively) in Europe.

Belarus has not joined CLARIN yet, however it is increasingly getting more involved in its various operations. In October 2019 researchers from Belarus (among other countries) participated in a 3-day PARTHENOS (Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies) workshop (CLARIN ERIC, 2020). Additionally, at the end of 2019 the Speech Synthesis and Recognition Laboratory of the National Academy of Sciences of Belarus started providing metadata to the Virtual Language Observatory (Gorgaini, 2020). These activities prove that some research institutions in Belarus are interested in co-operating with CLARIN and their efforts might lead to Belarus gaining an official status in the consortium. It is however unlikely to happen within the next couple of years, as it would require setting up the national consortium first and then involving the government (typically the Ministry of Education or similar structure) in order to pay the membership fee.

In the case of Belarus the problem with sharing digital resources is complex, not merely due to its absence from CLARIN structures. The difficulties start at the structural level, as in Belarus there are no popular online platforms for sharing any kind of materials, such as Moodle in Poland or Blackboard in the UK. It is actually one of the hindrances in developing distance education in

---

10 State of affairs in June 2020. Attempts to obtain copyrights or licences will continue until the end of 2020.

Belarus (Kryvoi, 2017). Lack of a common structure for sharing linguistic materials, a structure that would be designed with Belarusian language users in mind, is a major issue in the process of reaching the maximum number of users, because it is a sign of broader problem with such technologies in Belarus. As indicated in the report on distance education in Belarus (Kryvoi, 2017), there is almost no qualified personnel for managing a system of distance education, moreover the concept is not well understood at the higher education institutions and consequently it is seriously underestimated.

Currently there are only a few options for making the corpus available to researchers based in Belarus. The solution that will be initially implemented is simply granting access based on individual requests – which will be necessary also in the case of scientists working in institutions without access to CLARIN infrastructure. It also seems reasonable to organise workshops for researchers and educators who do not have experience using corpora. However for now, such activities remain in the sphere of plans.

The access to the data should not only be restricted for academic use, but also protected from misuse. This is achieved by the lack of access to the raw corpus. To ensure such a restriction the EPB corpus will be accessible via a specifically designed interface that will allow querying the data and conducting some statistical tests<sup>11</sup>. Currently, that is since the first half of the year 2020, the EPB can be accessed via the KonText tool (Machálek, 2014). At the moment (as of June 2020) 30 short stories, novellas and novels by 27 authors are uploaded, together they comprise a corpus of 0.7 million tokens. Currently adding a subcorpus containing samples from all EPB texts is not available, however the programme can extract a randomised sample from the search results. As mentioned above, the website of the project will be used as a medium for sharing the information about data processing and addition progress.

KonText has been implemented for sharing InterCorp (mentioned in Section 2.1), as this application allows for bilingual and trilingual queries.

---

11 The interface is planned to be gradually developed and to offer more advanced tools with time.

Search in the corpus

Corpus: (English) The Hitchhiker's Guide to the Galaxy

Query Type: Basic

Query:

Specify context

Specify query according to the meta-information

Aligned corpora

Add corpus (Polish) Autostopem przez Galaktykę

Search Clear All

Illustration 5.19. KonText main window.

The default window, *Search in the corpus* (New query option in *Query* tab in the menu at the top of the window) has three main elements: *Corpus*, *Query type* and *Query*. *Query* is a window for entering the search term. *Query type* allows for specifying whether the search term is a specific word form, lemma, phrase, character or a query in a specific query language. The *Corpus* button shows available corpora; only after choosing a corpus which is part of a larger parallel corpus a new option appears. The *Add a corpus* button displays a list of corpora aligned with the user's first choice. Choosing the parallel corpus results in adding the second three-element window designated to the other corpus. This operation can be then repeated to produce query windows for two or three aligned corpora. Additionally, the user can narrow down the search by filtering selected word forms (option *Specify context*) and by choosing texts with specific properties (option *Specify query according to the meta-information*).

So far, the last option contains only a handful of metadata included in the application used for aligning, namely the title and the publication date. The range of choices regarding metadata will be developed in the near future, to include date of birth and the gender of the author and the translators. Before querying the corpus, the user can create a subcorpus (tab *Subcorpora* at the top of the window) according to these metadata, e.g. by choosing only the texts published in the 1970s or texts written by women. Apart from specifying a subcorpus using the attribute list, the user can create a custom 'within' condition, but this requires familiarity with the structure of the database in which the corpus is stored.

Typing the query is obligatory for the first corpus. Such a search returns segments containing the search term (be it a character, word or phrase) in the first selected corpus and the aligned segments from the parallel corpus (or corpora). The search can be narrowed down by typing the query term

for the parallel corpus (or corpora) as well. In that case only the segments containing the query in both (or more) languages will be displayed. Presentation of the query results can be modified in various ways (tab *View options*) but by default the output is displayed as in the following figure:

Illustration 5.20. Concordance in the KonText application.

Highlighted words are the ones that were introduced in the query window. The most probable equivalents for individual words are available in another tool developed within the CLARIN infrastructure, Treq (Vavřín, 2015), which is not implemented in the KonText interface. The green strip at the top shows some basic statistics: overall number of hits, number of hits per million (in relation to the whole corpus) and the average reduced frequency (ARF) which combines the simple word count and document count, that is number of occurrences in each document. Clicking the highlighted word or the segment (when no word is highlighted) opens a pop-up window displaying a 150-word fragment of text surrounding the search term (or the segment). By default the results are shuffled, which means that even if a searched term appears few times in one text the excerpts from this particular text will not be displayed one after another but among fragments of other texts containing the query. The search results can be enhanced by means of filtering (*Filter* tab). KonText allows positive filtering, that is selecting part of the result, and negative filtering, namely removing some concordances according to words (characters, phrases or any other criteria specified with the use of the query language) occurring in a certain proximity of the main search term (this proximity is also defined by the user).

Another option available in KonText is creating word list (*Query* tab). KonText allows for reducing the word list according to specific criteria expressed by means of a regular expression pattern. The user can specify the minimum frequency of a word, as well as add a white- and blacklist. The resulting wordlist can be ordered according to simple word count, document count or ARF. Details of how the ARF is calculated are available in the Czech National Corpus manual (Pojarová, 2016). The wordlist can be saved (*Save* tab at the top of the window) in either CSV, XML or TXT format; in the case of these options only the first 50 lines are recorded. In order to save more lines or the full result the user needs to choose *Custom* in the *Save* tab.

KonText offers complex sorting options for the concordance lines (*Concordance* tab), in the case of a high number of hits it can create a random sample from the results (*Concordance* tab), additionally it includes the function of creating a frequency distribution according to a number of criteria (*Frequency* tab) and the collocation search (*Collocations* tab). A full description of possibilities given by the KonText application is available in the user manual (*Help* tab).

The previous sections show how complex and advanced the KonText tool is and what possibilities it gives when it comes to analysing the data. The one thing that might improve this application is implementing a system for marking equivalents. In the Paralela Corpus (discussed in Section 1.4), also developed and maintained within the CLARIN infrastructure, it is not only the search term that is highlighted, but also its translation determined by an algorithm based on the Dice coefficient (Dice, 1945; Pezik, 2016). As mentioned earlier, this function is available in the form of a separate tool, however it might be possible in the future to merge it with KonText.

An important element of all KonText functions is the query language, namely Corpus Query Language (CQL). CQL (used e.g. in SketchEngine), like all query languages, builds on regular expressions, expanding and adapting them for particular needs. The specificity of CQL is that, apart from creating queries relating to grammatical annotation, it also enables specifying structural attributes, such as sentence or document id. It makes querying the corpus much more versatile, provided the user is familiar with the underlying structure of the database.

In order to facilitate parallel search in the EPB corpus the data has to be aligned, therefore alignment, rather than the grammatical annotation, was the first step in preparing the text for publishing with the interface. As mentioned in Section 2.4, the corpus has been aligned with the use of the Mantel application. Due to the differences in the graphic design of Mantel and KonText some solutions utilised in the alignment process cannot be implemented in the EPB interface. In particular, the crosslink relation between segments, cannot be transferred to KonText, therefore segments connected in this way have to be merged. The initial 1.5 million-word aligned subcorpus of the EPB is gradually being enlarged and it should become at least twice as big by the end of the 2020.

In the same time another issue is being resolved, namely grammatical annotation, more specifically tokenisation, lemmatisation and POS tagging. As stated in Section 2.3, the Universal Dependencies annotation scheme has been chosen for this task. UD, supporting a number of typologically different languages, provides multi-level annotation for corpora. One layer of this annotation scheme is universal POS tags marking the core part-of-speech categories which can be identified across nearly all languages thus enabling comparison even between very distinct language groups. The possibility of tagging whole corpora makes quantitative data analysis possible and therefore enriches the qualitative analysis which is severely restricted in the case of big datasets.

Despite its obvious advantages, unifying the tagsets is criticised for several reasons. Particularly, in the case of UD the main drawback is describing similar phenomena in different ways. An example of such an issue is the Polish word *dziewiąty* (*ninth*) which can be annotated either as a numeral or as an adjective and there are no clear guidelines regarding the differentiation between the two. Another problem is the way of treating text items – either as syntactic or as orthographical words.

This is especially difficult considering the use of hyphens in English and can be exemplified by contrasting items, such as *can't* and *I'm* with *mum's* and *dad's*. Yet another issue, specific only to Polish, is that treebanks for this language have been prepared with specific properties of Polish in mind and because of that some verbs are split into two or three parts, the first always being the core and the following defining the past tense and the conditional mood. In summary, there is always a loss when converting from language-specific tags. However, the EPB corpus has been designed with translation studies in mind, rather than detailed grammatical investigation and therefore the annotation scheme which provides a common ground for all the languages included in the project seems to be the best choice. Moreover, UD annotation retains the language-specific tags in another layer alongside the universal tagset, therefore it is always possible to query each language in search for a specific phenomenon and using much more detailed tags.

In the long term perspective, the EPB has a chance of becoming a part of the National Corpus of Polish (NCP), as long as the research community and funding institutions will take action and focus on prolonging the national corpus project (Ogrodniczuk et al., 2019). Researchers have already been signalling the need to develop the NCP, not merely in terms of timeliness, that is adding new material to reflect the current use of the language, but also with regard to parallel subcorpora which are already part of the biggest and most respected national corpora worldwide. This, however, remains a subject for the unforeseeable future.

## Conclusion

This thesis posed several important questions concerning the Belarusian language and answered them by tackling various tasks discussed in particular chapters. Chapter 1 provided an introduction to the topic, in particular it discussed the specificity of the three languages of the project (Belarusian, Polish and English), moreover it gave an overview of corpus linguistics as a research discipline and as a methodology, with special attention to that field in particular languages.

The first problem considered in this thesis concentrated on the data for the project: how and where to find and collect texts for a corpus of a minority language, such as Belarusian, how to process the data most efficiently, that is digitise the texts (when needed) and then enrich them with additional linguistic data, and finally how to deposit the corpus so it would be safe, easily accessible and also available in the far future? This issue was tackled in Chapter 2 which described corpus compilation in all its stages: design, text collection, encoding and alignment, as well as storing and sharing the data. Chapter 2 not only gave an overview of existing solutions but also discussed in details which of them can be implemented in the case involving three languages, that is English, Polish and Belarusian.

The second topic of this thesis was the practical use of the corpus and the degree to which it is possible to use the collected data in some of the areas indicated in Section 1.3. From various research fields translation studies were chosen as the focal point of Section 3 and discourse analysis of Chapter 4. In particular, Section 3.1 tackled the topic of theoretical translation studies and considered three translation universals (TU), namely explicitation, simplification and levelling out. The analysis proved that indicators traditionally used for identifying the aforementioned TU do not show the same phenomena when calculated for the data gathered in the EPB corpus as well as in the OPUS corpus. Consequently, the simplification, explicitation and levelling out cannot be confirmed in the translations between English and Slavic languages, in particular Polish and Belarusian.

Next, Section 3.2 addressed descriptive translation studies and presented an overall stylistic analysis of the corpus. After the initial investigation of the corpus as a whole some translation outliers have been identified and one of them was scrutinised in order to determine the possible reasons of its dissimilarity to the other texts by the same author. The investigation shows that the Belarusian translation of *The Kidnapped Prime Minister* short story highlights some unusual features of Agatha Christie's writing, that is the use of masculine verbs and putting male characters in the position of the object of discourse, use of negation and first-person discourse.

Finally, Section 3.3 exemplified applied translation studies by the use of the EPB corpus in analysing phraseology for training purposes. This chapter focused on phraseology, in particular on several examples of idioms, rare phraseologisms and phrasal verbs. Examples of such language items were analysed from the point of view of a beginner translator in order to show the potential of the EPB corpus is providing immediately available solutions as well as patterns for dealing with certain type of translation problems.

Chapter 4 comprised the analysis of gender discourse exemplified by the frequencies of some masculine and feminine words in the corpus, as well the collocates of the words *woman*, *women*,

*man, men, girl, girls, boy and boys*. The investigated elements exhibited more differences in the English original texts than in the Polish and Belarusian translations thus showing the disparity between these languages on the discursive and lexical levels. Analysis of the same keywords in corpora of original Polish and Belarusian languages revealed the non-translational texts bearing more similarities to non-translation English.

All in all, Chapters 3 and 4 presented challenging results in several linguistic fields and by doing that pointed out the possibilities of future research. These include not only deeper analysis of the collected data in the domain of translation studies and discourse analysis but also in other areas discussed in Section 1.3, such as lexicography or pragmatics. With the prospect of future research in mind the project described in this thesis remains open-ended and Chapter 5 indicates how and in what direction it can further develop, as well as what steps have been taken in order to secure the future accessibility to the EPB corpus.

The value of creating resources for minoritised languages cannot be underestimated, on the contrary, it should be also considered in a wider context of global equality, language rights and minority access and participation in digital culture. A recent report to the European Parliament on language equality in the digital age (Evans, 2018) boldly states that

owing to a lack of adequate policies in Europe, there is currently a widening technology gap between well-resourced languages and less-resourced languages [...]. European lesser-used languages are at a significant disadvantage on account of an acute lack of tools, resources and research funding, which is inhibiting and narrowing the scope of the work done by researchers who, even if equipped with the necessary technological skills, are unable to derive the full benefit of language technologies.

It is not *fewer* tools of resources, but *an acute lack* of such that lies in the heart of this project and similar endeavours, such as the Digital Language Diversity Project (Soria et al., 2016). In the digital age English is getting stronger, threatening minority languages more than ever, but “technology could help us bring minority languages to a wider audience. If we work out how to play the game right, we could use it to help bolster linguistic diversity rather than damage it” (Evas, 2014, para. 3). The key is to gather more high-quality material – the more data the better the language processing algorithms become.

The main problem the minority languages users are facing right now is digital exclusion. As indicated in Chapter 1, in the case of Belarusian the non-professional sources, such as Wikipedia, are well developed and have the constant support of volunteer contributors. However, institutional support and the existence of datasets which can be used in professional settings are essential for empowering a language. A carefully prepared corpus, such as an EPB, is part of that agenda.

The issue of digital accessibility, especially for minorities, has never been more burning than in 2020. The Covid-19 pandemic, lockdowns affecting nearly all countries in the world and the massive shift towards online activity in almost all spheres of our lives only stress the necessity of supporting and speeding up the digitisation of less-resourced languages. The EPB corpus is serving precisely that purpose and hopefully it will contribute to preventing the digital exclusion of Belarusian language users.

## Glossary

- accuracy - “a basic score for evaluating automatic annotation tools such as parsers or part-of-speech taggers. It is equal to the number of tokens correctly tagged, divided by the total number of tokens. [...] usually expressed as a percentage” (Baker, Hardie, & McEnery, 2006, p. 7)
- alignment – a process of adding information about “which parts of a text in language A correspond to the equivalent corresponding text in language B” (P. Baker et al., 2006, p. 9)
- ambiguity – a special case in annotating where either two tags can be assigned to an item
- annotation/encoding/markup/tagging – “the process of encoding additional information to corpus data” (P. Baker et al., 2006, p. 13)
- authorship identification – “the field of text analysis which attempts to ascertain whether a given text was written by a particular author or not, usually by automatic and/or statistical methods” (P. Baker et al., 2006, p. 17)
- chi-square – “a test for determining the significance of any numeric difference observed in data” (P. Baker et al., 2006, p. 31)
- cluster – either a group of words in sequence, or a group of linguistically similar texts
- colligation – “a form of collocation which involves relationships at the grammatical rather than the lexical level” (P. Baker et al., 2006, p. 36)
- collocation – words frequently co-occurring in close proximity to each other
- concordance – “a list of all of the occurrences of a particular search term in a corpus” (P. Baker et al., 2006, p. 42) presented as KWIC (keyword in context)
- consistency – a situation in which the same linguistic phenomenon is always annotated in the same way
- content & function words – content words are those having lexical meaning, they are traditionally understood as nouns, adjectives, main verbs and adverbs, as opposed to function words, that is pronouns, prepositions, determiners, conjunctions, auxiliary and modal verbs
- corpus-based & corpus-driven – corpus-based investigations use “a corpus as a source of examples to check researcher intuition or to examine the frequency and/or plausibility of the language contained within a smaller data set” (P. Baker et al., 2006, p. 49), whereas in corpus-driven studies there are no presuppositions about the regularities which are to be found in the data
- corpus linguistics – a methodology of studying language based on “real life” examples collected in form of corpora

- dispersion – “the rate of occurrence of a word or phrase across a particular file or corpus” (P. Baker et al., 2006, p. 59)
- frequency – a number of occurrences of an item within a set of data
- gold standard – a “dataset or corpus [...] whose annotation has been checked and corrected” (P. Baker et al., 2006, p. 78)
- granularity – level of detail in the dataset
- keyword – “a word which appears in a text or corpus statistically significantly more frequently than would be expected by chance when compared to a corpus which is larger or of equal size” (P. Baker et al., 2006, p. 97), or the subject of a concordance
- lemma – a basic word form, traditionally used in dictionaries as a headword
- lemmatisation – a process of assigning a lemma to each item in a corpus
- lexical density – a percentage of lexical words in a corpus
- metadata – extra-textual information about a corpus
- morphological richness – “a reference to how many different inflectional forms the lexemes of a language have” (P. Baker et al., 2006, p. 117)
- multidimensional analysis – an approach to “the linguistic analysis of texts, genres, text types, styles or registers” (Biber, 1992, p. 332), which differentiates data five dimensions defined through the analysis of dozens of linguistic features and patterns of co-occurrence among them
- n-gram – a sequence of n words appearing in the text
- OCR – optical character recognition; process of converting an image of text into machine-readable form
- POS tagging – part-of-speech tagging; a process of assigning a part-of-speech tag to each item in a corpus
- plain text & raw corpus – “a text or corpus that does not contain any markup (whether Standard Generalised Markup Language (SGML), Extensible Markup Language (XML) or other), or any added analysis such as part-of-speech tags and contains only the actual words of the original document” (P. Baker et al., 2006, p. 131); a raw corpus consist of plain text files
- post-editing, post-processing & proof-reading – the first process is a manual correction of POS tags, whereas the second one is automatic; proof-reading, one the other hand is either manual or automatic but it concerns the text itself
- regular expression – chains of signs that allow for describing patterns
- semantic preference (semantic prosody) – the preference of a word or phrase to co-occur with a specific semantic category

- skeweness – in statistics it is a measure of asymmetry in the frequency distribution
- stylometry – the quantitative study of writing style
- tagset – “a collection of tags (or codes) that occur in an encoding or tagging scheme used to annotate corpora” (P. Baker et al., 2006, p. 155)
- token & type – token is an individual occurrence of an item in a corpus, whereas type is any item recurring in a corpus, that is number of types is a number of unique words and the number of tokens is the number of occurrences of all types
- tokenisation – “the automatic process of converting all of a text into separate tokens” (P. Baker et al., 2006, p. 160)
- type/token ratio – a proportion of types to tokens in a corpus
- vector – quantities represented as directed lines having magnitude (which corresponds with the length of the line) and direction (which corresponds with the direction of the arrowhead attached to the line)
- wordlist – “a list of all of the words that appear in a text or corpus” (P. Baker et al., 2006, p. 169)
- wordnet – a lexical database representing words in within collection of synonyms and showing also relations of hyperonymy/hyponymy

## Bibliography

### Primary sources (A) – English-Polish-Belarusian Parallel Literary Corpus

- Adams, D. (1979). *The Hitchhiker's Guide to the Galaxy*. London: Pan Books.
- Adams, D. (1992). *Autostopem przez Galaktykę* (A. Banaszak, Trans.). Poznań: CIA-Books\_Svaro (Original work published 1979).
- Adams, D. (2015). *Aŭtaspynam pa hałaktycy* (P. Kaściukievič, Trans.). Minsk: Łohvinaŭ (Original work published 1979).
- Aldridge, J. (1950). *The Hunter*. London: Bodley Head.
- Aldridge, J. (1955). *Łowca* (J. Zakrzewski, Trans.). Warsaw: Czytelnik (Original work published 1950).
- Aldridge, J. (1996). *Palaŭničy* (M. Vałoska, Trans.). Minsk: Junactva (Original work published 1950). Retrieved from [http://knihi.com/Dzejms\\_Oldrydz/Palaunicy.html](http://knihi.com/Dzejms_Oldrydz/Palaunicy.html)
- Atwood, M. (2005a). *Penelopiad*. Edinburgh: Canongate.
- Atwood, M. (2005b). *Penelopiada* (M. Konikowska, Trans.). Kraków: Znak (Original work published 2005).
- Atwood, M. (2018). *Pienielapijada* (V. Kałackaja, Trans.). Minsk: Łohvinaŭ (Original work published 2005).
- Bach, R. (1970). *Jonathan Livingston Seagull*. New York, NY: Macmillan Publishers.
- Bach, R. (1995). *Mewa* (R. Zubek, Trans.). Poznań: Dom Wydawniczy „Rebis” (Original work published 1970).
- Bach, R. (1994). *Čajka pa imieni Džonatan Livinhstan*. In I. Dabryjan (Ed.), & A. Michalčuk (Trans.), *Historyja kachannia: Apowieści*. Minsk: Mastackaja litaratura (Original work published 1970). Retrieved from [http://knihi.com/Rycard\\_Bach/Cajka\\_pa\\_imieni\\_Dzonatan\\_Livinhstan.html](http://knihi.com/Rycard_Bach/Cajka_pa_imieni_Dzonatan_Livinhstan.html)
- Barrie, J. M. (1911). *Peter Pan and Wendy*. Retrieved from <http://www.literatureproject.com/peter-pan/>
- Barrie, J. M. (1958). *Piotruś Pan. Opowiadanie o Piotrusiu i Wendy* (M. Słomczyński, Trans.). Warsaw: Nasza księgarnia (Original work published 1911).
- Barrie, J. M. (2012). *Piter Pen* (U. Lankievič, Trans.). *PrajdziSviet, 10* (Original work published 1911). Retrieved from <http://prajdzisvet.org/kit/116-piter-pen.html>
- Baum, L. F. (1900). *The Wonderful Wizard of Oz*. Chicago, IL: George M. Hill Company.

- Baum, L. F. (2012). Čaraunik Krainy Oz (S. Miadźviedzieŭ, Trans.). *Prajdzisviet*, 10 (Original work published 1900). Retrieved from <http://prajdzisvet.org/text/1209-charaunik-krainy-oz.html>
- Baum, L. F. (2017). *Czarnoksiężnik z Krainy Oz* (P. Łopatka, Trans.). Warsaw: Zielona Sowa (Original work published 1900). Retrieved from [https://ebookpoint.pl/ksiazki/czarnoksiężnik-z-krainy-oz-literatura-klasyczna-lyman-frank-baum,e\\_0m8c.htm](https://ebookpoint.pl/ksiazki/czarnoksiężnik-z-krainy-oz-literatura-klasyczna-lyman-frank-baum,e_0m8c.htm)
- Bradbury, R. (1949). The One Who Waits. *The Arkham Sampler*, (Summer). Retrieved from <http://teacherweb.com/VA/KingGeorgeHighSchool/MrsDunn/The-One-Who-Waits.pdf>
- Bradbury, R. (1953). *Fahrenheit 451*. Retrieved from <http://www.secret-satire-society.org/wp-content/uploads/2014/01/Ray-Bradbury-Fahrenheit-451.pdf>
- Bradbury, R. (1955). The Dragon. *Esquire* (August). Retrieved from <http://mrscryer.weebly.com/uploads/1/0/5/3/10533164/dragon.pdf>
- Bradbury, R. (1960). *451 stopni Fahrenheita* (A. Kaska, Trans.). Warsaw: Czytelnik (Original work published 1953).
- Bradbury, R. (1996a). Cmok (A. Kudraŭcaŭ, Trans.). *Krynica*, 17(1) (Original work published 1955).
- Bradbury, R. (1996b). Toj, chto čakaje (A. Kudraŭcaŭ, Trans.). *Krynica*, 17(1) (Original work published 1949).
- Bradbury, R. (2003). Ten kto czeka (M. S. Nowowiejski, Trans.). *Nowa Fantastyka*, 6, 33–35 (Original work published 1949).
- Bradbury, R. (2009). Smok. In A. Gren (Trans.), *Ilustrowany Człowiek i inne opowiadania*. Warsaw: Świat Książki (Original work published 1955).
- Bradbury, R. (2015). *451 hradus pa Farenhieicie* (R. Vieramiej, Trans.) (Original work published 1953). Retrieved from <http://translatedby.com/you/451-gradus-po-farengeitu-chast-1/into-be/>
- Brown, D. (2003). *The Da Vinci Code*. New York, NY: Doubleday.
- Brown, D. (2014). Kod da Vinčy (V.K., Trans.). *Arche*, 1–2(134–135), 4–380 (Original work published 2003).
- Brown, D. (2016). *Kod Leonrada da Vinci* (K. Mazurek, Trans.). Katowice: Sonia Draga (Original work published 2003).
- Bukowski, C. (1973). You and Your Beer and How Great You Are. In *South of No North* (pp. 23–27). Boston, MA: Black Sparrow Press.
- Bukowski, C. (1983). The Most Beautiful Woman in Town. In *The Most Beautiful Woman in Town*. San Francisco: City Lights Books.

- Bukowski, C. (1996). Ty, twoje piwo i to, jaki jesteś wspaniały. In L. Ludwig (Trans.), *Na południe od nigdzie*. Warsaw: Oficyna Literacka Noir sur Blanc (Original work published 1973).
- Bukowski, C. (2002). Najpiękniejsza dziewczyna w mieście. In R. Sudół (Trans.), *Najpiękniejsza dziewczyna w mieście*. Warsaw: Oficyna Literacka Noir sur Blanc (Original work published 1983).
- Bukowski, C. (2009). Samaja pryhožaja žančyna ũ horadzie (N. Hvozdziewa, Trans.). *PrajdziSviet*, 1 (Original work published 1983). Retrieved from <http://prajdzisvet.org/text/14-samaja-pryhozhaja-zhanchyna-u-horadzie.html>
- Bukowski, C. (2013). Ty, tavjo piva i tvaja krutaść (A. Javar, Trans.). *PrajdziSviet*, 11 (Original work published 1973). Retrieved from <http://prajdzisvet.org/texts/prose/2826.html>
- Carver, R. (1976). Fat. In *Will You Please Be Quiet, Please?* New York, NY: McGraw-Hill.
- Carver, R. (2002). Bambiza (A.D., Trans.). *Naša Niva*, 48 (Original work published 1976).
- Carver, R. (2013). Tłuścioch (M. Tabaczyński, Trans.). *Fabularie*, 2 (Original work published 1976).
- Cheever, J. (1947, May 17). The Enormous Radio. *The New Yorker*. Retrieved from [http://english307formsofmodernshortstory.web.unc.edu/files/2014/03/enormous\\_radio-by-John-Cheever.pdf](http://english307formsofmodernshortstory.web.unc.edu/files/2014/03/enormous_radio-by-John-Cheever.pdf)
- Cheever, J. (1969). Nowe radio. In Z. Sroczyńska (Trans.), *Włamywacz z Shady Hill*. Warsaw: Książka i Wiedza (Original work published in 1947).
- Cheever, J. (2001). Vializnaje radyjo (A. Kudraŭcaŭ, Trans.). *Krynica*, 9(69), 90–101 (Original work published 1947).
- Chesterton, G. K. (1911, May 20). The Flying Stars. *The Saturday Evening Post*. Retrieved from <http://famous-and-forgotten-fiction.com/writings/chesterton-the-flying-stars.html>
- Chesterton, G. K. (2008). Latające gwiazdy. In T. J. Dehnel (Trans.), *Przygody Księdza Browna*. Warsaw: Fronda (Original work published 1911).
- Chesterton, G. K. (2012). Latučyja zorki (A. F. Bryl, Trans.). *PrajdziSviet*, 9 (Original work published 1911). Retrieved from <http://prajdzisvet.org/text/1109-liatuchyja-zorki.html>
- Christie, A. (1924). The Kidnapped Prime Minister. In *Poirot Investigates*. London: The Bodley Head.
- Christie, A. (1925). The Veiled Lady. In *Poirot Investigates*. New York, NY: Dodd, Mead & Co.
- Christie, A. (1932). The Four Suspects. In *The Thirteen Problems*. London: Collins Crime Club.
- Christie, A. (1937). Dead Man's Mirror. In *Murder in the Mews*. London: Collins Crime Club.
- Christie, A. (1942). The Case of the Retired Jeweller (The Tape-Measure Murder). *Strand Magazine*, (614).

- Christie, A. (1950). The Case of Perfect Maid. In *Three Blind Mice and Other Stories*. New York, NY: Dodd, Mead & Co.
- Christie, A. (1988). Niabožčykava lustra. In P. Marcinovič (Trans.), *Zamiežny detektyŭ*. Minsk: Mastackaja litaratura (Original work published 1937).
- Christie, A. (1991a). Biezdakornaja pakajoŭka. In V. Čudaŭ & M. Kotaŭ (Trans.), *Zamiežny detektyŭ* (pp. 159–170). Minsk: Mastackaja litaratura (Original work published 1942).
- Christie, A. (1991b). Čaćviora padazronych. In V. Čudaŭ & M. Kotaŭ (Trans.), *Zamiežny detektyŭ* (pp. 145–159). Minsk: Mastackaja litaratura (Original work published 1932).
- Christie, A. (1991c). Vykradannje prem'ier ministra. In V. Čudaŭ & M. Kotaŭ (Trans.), *Zamiežny detektyŭ* (pp. 170–186). Minsk: Mastackaja litaratura (Original work published 1924).
- Christie, A. (1991d). Zabojsťva ŭ Łabiernet-Katedży. In V. Čudaŭ & M. Kotaŭ (Trans.), *Zamiežny detektyŭ* (pp. 133–145). Minsk: Mastackaja litaratura (Original work published 1942).
- Christie, A. (1995). Czterech podejrzaných. In M. Weiss (Trans.), *Trzyńasćie zagadek; Hotel 'Bertram'* (pp. 89–100). Warsaw: Świat Książki (Original work published 1932).
- Christie, A. (1996a). Idealna służąca. In A. Bihl (Trans.), *Śmiertelna klątwa i inne opowiadania*. Wrocław: Wydawnictwo Dolnośląskie (Original work published 1942).
- Christie, A. (1996b). Narzędzie zbrodni. In A. Bihl (Trans.), *Śmiertelna klątwa i inne opowiadania*. Wrocław: Wydawnictwo Dolnośląskie (Original work published 1942).
- Christie, A. (1998). Dama pad čornaj vuallu (M. Kandrusievič, Trans.). *Krynica*, 3(3), 15–17 (Original work published 1925).
- Christie, A. (2013). Dama z woalką. In A. Rojkowska & A. Milcarz (Trans.), *Wczesne sprawy Poirota*. Wrocław: Wydawnictwo Dolnośląskie (Original published in 1925).
- Christie, A. (2014). Lustro nieboszczyka. In J. Zaus (Trans.), *Morderstwo w zaułku*. Wrocław: Wydawnictwo Dolnośląskie (Original work published 1937).
- Christie, A. (2015). Porwanie premiera. In B. Kaliszewicz (Trans.), *Poirot prowadzi śledztwo*. Wrocław: Wydawnictwo Dolnośląskie (Original work published 1924).
- Dahl, R. (1952, May 17). Skin. *The New Yorker*, 31–32.
- Dahl, R. (1953). Lamb to the Slaughter. *Harper's Magazine*, (September). Retrieved from [http://www.depa.univ-paris8.fr/IMG/pdf/lamb\\_to\\_the\\_slaughter\\_by\\_roald\\_dahl-2.pdf](http://www.depa.univ-paris8.fr/IMG/pdf/lamb_to_the_slaughter_by_roald_dahl-2.pdf)
- Dahl, R. (1984a). Jagnię na rzeź. In A. Demkowska (Trans.), *Niespodzianki*. Warsaw: Książka i Wiedza (Original work published in 1952).
- Dahl, R. (1984b). Skóra. In A. Demkowska (Trans.), *Niespodzianki*. Warsaw: Książka i Wiedza (Original work published in 1952).

- Dahl, R. (1989). Jahnja na bojni (U. Sierpikaŭ, Trans.). *Krynica*, 10(22), 28–31 (Original work published 1984).
- Dahl, R. (n.d.). *Skura* (T. Papova, Trans.). Retrieved from [http://knihi.com/Roald\\_Dal/Skura.html](http://knihi.com/Roald_Dal/Skura.html)
- Faulkner, W. (1938). An Odor of Verbena. In *The Unvanquished*. New York, NY: Random House.
- Faulkner, W. (1991). Zapach werbeny. In E. Życieńska (Trans.), *Niepokonane*. Wrocław: Zakład Narodowy im. Ossolińskich (Original work published 1938).
- Faulkner, W. (1995). Pach vierbieny (V. Niebyšyniec, Trans.). *Krynica*, 8(13) (Original work published 1938).
- Fitzgerald, F. S. (1925). *Great Gatsby*. Retrieved from [https://ebooks.adelaide.edu.au/f/fitzgerald/f\\_scott/gatsby/index.html](https://ebooks.adelaide.edu.au/f/fitzgerald/f_scott/gatsby/index.html)
- Fitzgerald, F. S. (2013). *Wielki Gatsby* (J. Dehnel, Trans.). Kraków: Znak (Original work published 1925).
- Fitzgerald, F. S. (n.d.). Vialiki Getsby (S. Matyrka, Trans.). Retrieved from Biełaruski PEN Centr website: <https://pen-centre.by/matyrka.html>
- Fowles, J. (1974). Poor Koko. In *The Ebony Tower*. London: Jonathan Cape Ltd.
- Fowles, J. (2001). Niaboha Kako (V. Sudlakova, Trans.). *Krynica*, 7(67), 146–182 (Original work published 1974).
- Fowles, J. (2002). *Niedobry Koko* (T. Bieroń, Trans.). Poznań: Zysk i Spółka (Original work published 1974).
- Gaiman, N. (2002). *Coraline*. New York, NY: HarperCollins.
- Gaiman, N. (2003). *Koralina* (P. Braiter, Trans.). Warsaw: MAG (Original work published 2002).
- Gaiman, N. (2016). *Karalina* (D. Vaškevič, Trans.). Minsk: Zmicier Kołas (Original work published 2002).
- Golden A. (1997). *Memoirs of a Geisha*. New York, NY: Alfred A. Knopf.
- Golden A. (2001). *Wyznania gejszy* (W. Nowakowski, Trans.). Warsaw: Albatros (Original work published 1997).
- Golden A. (2015). Miemuary hiejšy (V.K., Trans.). *Arche*, 7-8 (140-141), 4–390 (Original work published 1997).
- Golding, W. (1971). The Scorpion God. In *The Scorpion God. Three Short Novels*. London: Faber and Faber.
- Golding, W. (1988). Bóg Skorpion. In L. Stafiej & M. Golewska-Stafiej (Trans.), *Bóg Skorpion*. Kraków: Wydawnictwo Literackie (Original published 1971).

- Golding, W. (2003). Boh Skarpijon (V.K., Trans.). *Dziejasłouŭ, 6* (Original work published 1988). Retrieved from <http://dziejaslou.by/old/www.dziejaslou.by/inter/dzeja/dzeja.nsf/htmlpage/hold602ec.html?OpenDocument>
- Greene, G. (1932). *Stamboul Train*. Portsmouth, NH: Heinemann.
- Greene, G. (1955). *Loser Takes All*. London: Heinemann.
- Greene, G. (1957). *Stawka o żonę* (A. Ślósarczyk, Trans.). Warsaw: Instytut Wydawniczy Pax (Original work published 1955).
- Greene, G. (1966). The invisible Japanese gentlemen. *The Spectator*, 7176, 9–10.
- Greene, G. (1993a). Chto prajhraje, biare ūsio. In V. Niebyšyniec (Trans.), *Stambulski ekspres*. Minsk: Mastackaja litaratura (Original work published 1955).
- Greene, G. (1993b). Stambulski ekspres. In V. Niebyšyniec (Trans.), *Stambulski ekspres*. Minsk: Mastackaja litaratura (Original work published 1932).
- Greene, G. (1995). Niewidzialni japońscy dżentelmeni. In E. Krasieńska (Trans.), *Pożycz nam męża, Poopy i inne opowiadania* (pp. 116–120). Warsaw: Prószyński i S-ka (Original work published 1966).
- Greene, G. (1998). Japonskija dżentlmieny-nievidzimki (V. Drozd, Trans.). *Krynica*, 39(2), 83–86 (Original work published 1966).
- Greene, G. (2015). *Pociąg do Stambułu* (P. Kuś, Trans.). Warsaw: Albatros (Original work published 1932).
- Hemingway, E. (1924). Indian Camp. *The Transatlantic Review*. Retrieved from [https://liternet.bg/publish24/e\\_hemingway/killers.htm](https://liternet.bg/publish24/e_hemingway/killers.htm)
- Hemingway, E. (1940). *For Whom the Bell Tolls*. New York, NY: Charles Scribner's Sons.
- Hemingway, E. (1952). *The Old Man and the Sea*. New York, NY: Charles Scribner's Sons.
- Hemingway, E. (1956). *Stary człowiek i morze* (B. Zieliński, Trans.). Warsaw: Państwowy Instytut Wydawniczy (Original work published 1952).
- Hemingway, E. (1986). *Komu bije dzwon* (B. Zieliński, Trans.). Warsaw: Czytelnik (Original work published 1940).
- Hemingway, E. (1988). Obóz indiański. In B. Zieliński (Trans.), *49 opowiadań*. Warsaw: Państwowy Instytut Wydawniczy (Original work published 1924).
- Hemingway, E. (1991). *Pa kim zvonit' zvon* (V. Niebyšyniec, Trans.). Minsk: Mastackaja litaratura (Original work published 1940).
- Hemingway, E. (1994). Indziejski pasiołak (A. Astašonak, Trans.). *Krynica*, 3(9), 97–98 (Original work published 1924)

- Hemingway, E. (1996). *Stary čalaviek i mora* (P. Marcinovič, Trans.). Minsk: Junactva (Original work published 1952).
- Henry, O. (1905). The Gift of the Magi. *The New York Sunday World*.
- Henry, O. (2009). Dary mudracou (S. Miadźviedzieŭ, Trans.). *PrajdziSviet*, 3 (Original work published 1905). Retrieved from <http://prajdzisvet.org/text/312-dary-mudratsou.html>
- Henry, O. (2016). Upominek świąteczny. In E. Kowalska (Trans.), *Bożonarodzeniowe opowieści* (pp. 229–237). Poznań: Zysk i Spółka (Original work published 1905).
- Hughes, T. (1960). The Rain Horse. *The London Magazine*, (February).
- Hughes, T. (1982). Deszczowy koń. In T. Truszkowska (Trans.), *Deszczowy koń*. Kraków: Wydawnictwo Literackie (Original work published 1960).
- Hughes, T. (1998). Koń i dożdż (V. Drozd, Trans.). *Krynica*, 39(2), 86–91 (Original work published 1960).
- Huxley, A. (1924). Hubert and Minnie. In *Little Mexican*. Retrieved from <https://biblioklept.org/2014/03/10/hubert-and-minnie-aldous-huxley/>
- Huxley, A. (1993). Hubert i Minnie. In J. Olędzka (Trans.), *Uśmiech Giocondy*. Warsaw: Folium (Original work published 1924).
- Huxley, A. (2001). Hiubiert i Minni (V.K., Trans.). *Krynica*, 8(68), 76–87 (Original work published 1924).
- Ishiguro, K. (2005a). *Never Let Me Go*. London: Faber and Faber.
- Ishiguro, K. (2005b). *Nie opuszczaj mnie* (A. Szulc, Trans.). Warsaw: Albatros (Original published 2005).
- Ishiguro, K. (2012). Nie adpuskaj mianie (V. Sudlankova, Trans.). *Litaratura i mastactva*, 22(4668), 15 (Original work published 2005).
- Jackson, S. (1944). Trial by Combat. *The New Yorker*. Retrieved from [https://loa-shared.s3.amazonaws.com/static/pdf/Jackson\\_Trial\\_by\\_Combat.pdf](https://loa-shared.s3.amazonaws.com/static/pdf/Jackson_Trial_by_Combat.pdf)
- Jackson, S. (1948, June 26). The Lottery. *The New Yorker*. Retrieved from <http://fullreads.com/literature/the-lottery/>
- Jackson, S. (1976a). Loteria. In M. Michałowska (Trans.), *Loteria*. Warsaw: Państwowy Instytut Wydawniczy (Original work published 1948).
- Jackson, S. (1976b). Próba sił. In M. Michałowska (Trans.), *Loteria*. Warsaw: Państwowy Instytut Wydawniczy (Original work published 1944).
- Jackson, S. (2016a). Łatareja (D. Šostak, Trans.). *Maładość*, 5(750), 93–99 (Original work published 1948).

- Jackson, S. (2016b). Pravierka bojem (D. Šostak, Trans.). *Maładość*, 5(750), 99–102 (Original work published 1944).
- James, M. R. (1928). *Wailing Well*. Retrieved from <http://www.thin-ghost.org/items/show/164>
- James, M. R. (2009). Studnia stohnaŭ (V. Burlak, Trans.). *PrajdziSviet*, 3 (Original work published 1928). Retrieved from <http://prajdzisvet.org/text/300-studnia-Ctohnau.html>
- James, M. R. (2012). Jęcząca studnia. In M. Dżdża (Trans.), *Opowieści o duchach* (pp. 186–198). Toruń: Wydawnictwo ‘C&T’ (Original work published 1928).
- Joyce, J. (1914). A Painful Case. In *Dubliners*. London: Grant Richards Ltd.
- Joyce, J. (1922). *Ulysses*. Paris: Sylvia Beach.
- Joyce, J. (1969). *Ulysses* (M. Słomczyński, Trans.). Warsaw: Państwowy Instytut Wydawniczy (Original work published 1922).
- Joyce, J. (1989). Prykraje zdarenie (I. Babkoŭ, Trans.). *Dalahlady: Zamieŭnaja litaratura* (Original work published 1914).
- Joyce, J. (1991). Przypadek godny ubolewania. In K. Wojciechowska (Trans.), *Dublińczycy*. Warsaw: Oskar (Original work published 1914).
- Joyce, J. (1993). *Ulis* (S. Šupa, Trans.). (Original work published 1922). Retrieved from [http://kamunikat.org/usie\\_knihi.html?pubid=5029](http://kamunikat.org/usie_knihi.html?pubid=5029)
- Kerouac, J. (1957). *On the road*. New York, NY: Viking Press.
- Kerouac, J. (1993). *W drodze* (A. Kołyszko, Trans.). Warsaw: Państwowy Instytut Wydawniczy (Original work published 1957).
- Kerouac, J. (2013). Na darozie (M. Łapo, Trans.). *PrajdziSviet*, 11 (Original work published 1957). Retrieved from <http://prajdzisvet.org/text/1216-na-darozie.html>
- Kesey, K. (1962). *One Flew Over the Cuckoo’s Nest*. New York, NY: Viking Press & Signet Books.
- Kesey, K. (2009). *Lot nad kukulczym gniazdem* (T. Mirkowicz, Trans.). Warsaw: Wydawnictwo Albatros (Original work published 1962).
- Kesey, K. (2017). *Palot nad hniazdom ziaziuli* (A. Znatkievič, Trans.). Minsk: A. M. Januškievič (Original work published 1962).
- King, S. (1982). *The Breathing Method*. New York, NY: Viking Press.
- King, S. (1996). Miedtad pravilnaha dychannia (Aleś Kudraŭcaŭ, Trans.). *Krynica*, 23(8), 77–110 (Original work published 1982).
- King, S. (2014). Metoda oddychania. In Z. Królicki (Trans.), *Cztery pory roku*. Warsaw: Albatros (Original published 1982).
- Kipling, R. (1894). Rikki Tikki Tavi. In *The Jungle Book*. Retrieved from <https://www.gutenberg.org/ebooks/236>

- Kipling, R. (1902a). How the Camel Got His Hump. In *Just So Stories*. Retrieved from <https://www.gutenberg.org/ebooks/2782>
- Kipling, R. (1902b). How the First Letter Was Written. In *Just So Stories*. Retrieved from <https://www.gutenberg.org/ebooks/2787>
- Kipling, R. (1902c). How the Rhinoceros Got His Skin. In *Just So Stories*. Retrieved from <https://www.gutenberg.org/ebooks/2783>
- Kipling, R. (1902d). How the Whale Got His Throat. In *Just So Stories*. Retrieved from <https://www.gutenberg.org/ebooks/2781>
- Kipling, R. (1902e). The Beginning of the Armadillos. In *Just So Stories*. Retrieved from <https://www.gutenberg.org/ebooks/2786>
- Kipling, R. (1902f). The Butterfly That Stamped. In *Just So Stories*. Retrieved from <https://www.gutenberg.org/ebooks/2789>
- Kipling, R. (1902g). The Cat That Walked by Himself. In *Just So Stories*. Retrieved from <https://www.gutenberg.org/ebooks/2788>
- Kipling, R. (1902h). The Elephant's Child. In *Just So Stories*. Retrieved from <https://www.gutenberg.org/ebooks/2784>
- Kipling, R. (1902i). The Sing-Song of Old Man Kangaroo. In *Just So Stories*. Retrieved from <https://www.gutenberg.org/ebooks/2785>
- Kipling, R. (1923). Riki Tiki Tavi. In F. Mirandola (Trans.), *Księga dżungli* (Original work published 1894). Retrieved from <https://wolnelektury.pl/katalog/lektura/ksiega-dzungli.html>
- Kipling, R. (1939a). Adkul u kita takaja hłotka. In M. Bahun (Trans.), *Kazki* (Original work published 1902). Retrieved from [http://knihi.com/Dzozef\\_Redzjard\\_Kiplinh/Kazki.html#chapter1](http://knihi.com/Dzozef_Redzjard_Kiplinh/Kazki.html#chapter1)
- Kipling, R. (1939b). Adkul u nasaroha skura. In M. Bahun (Trans.), *Kazki* (Original work published 1902). Retrieved from [http://knihi.com/Dzozef\\_Redzjard\\_Kiplinh/Kazki.html#chapter3](http://knihi.com/Dzozef_Redzjard_Kiplinh/Kazki.html#chapter3)
- Kipling, R. (1939c). Adkul uzialisia branianosy. In M. Bahun (Trans.), *Kazki* (Original work published 1902). Retrieved from [http://knihi.com/Dzozef\\_Redzjard\\_Kiplinh/Kazki.html#chapter5](http://knihi.com/Dzozef_Redzjard_Kiplinh/Kazki.html#chapter5)
- Kipling, R. (1939d). Čamu ů viarbłuda horb. In M. Bahun (Trans.), *Kazki* (Original work published 1902). Retrieved from [http://knihi.com/Dzozef\\_Redzjard\\_Kiplinh/Kazki.html#chapter2](http://knihi.com/Dzozef_Redzjard_Kiplinh/Kazki.html#chapter2)

- Kipling, R. (1939e). Jak było napisana pierwsza pisma. In M. Bahun (Trans.), *Kazki* (Original work published 1902). Retrieved from [http://knihi.com/Dzozef\\_Redzjard\\_Kiplinh/Kazki.html#chapter6](http://knihi.com/Dzozef_Redzjard_Kiplinh/Kazki.html#chapter6)
- Kipling, R. (1939f). Kazka pra staroha kienhuru. In M. Bahun (Trans.), *Kazki* (Original work published 1902). Retrieved from [http://knihi.com/Dzozef\\_Redzjard\\_Kiplinh/Kazki.html#chapter10](http://knihi.com/Dzozef_Redzjard_Kiplinh/Kazki.html#chapter10)
- Kipling, R. (1939g). Kot, jaki hulaŭ sam saboju. In M. Bahun (Trans.), *Kazki* (Original work published 1902). Retrieved from [http://knihi.com/Dzozef\\_Redzjard\\_Kiplinh/Kazki.html#chapter7](http://knihi.com/Dzozef_Redzjard_Kiplinh/Kazki.html#chapter7)
- Kipling, R. (1939h). Matylok, jaki tupnuŭ nahoju. In M. Bahun (Trans.), *Kazki* (Original work published 1902). Retrieved from [http://knihi.com/Dzozef\\_Redzjard\\_Kiplinh/Kazki.html#chapter8](http://knihi.com/Dzozef_Redzjard_Kiplinh/Kazki.html#chapter8)
- Kipling, R. (1939i). Słonik. In M. Bahun (Trans.), *Kazki* (Original work published 1902). Retrieved from [http://knihi.com/Dzozef\\_Redzjard\\_Kiplinh/Kazki.html#chapter4](http://knihi.com/Dzozef_Redzjard_Kiplinh/Kazki.html#chapter4)
- Kipling, R. (1981). Ryki-Ciki-Tavi. In S. Michałczuk (Trans.), *Maŭhli: Z 'Knigi džunglaŭ'* (Original work published 1894). Retrieved from [http://knihi.com/Dzozef\\_Redzjard\\_Kiplinh/Kazki.html#chapter9](http://knihi.com/Dzozef_Redzjard_Kiplinh/Kazki.html#chapter9)
- Kipling, R. (2000a). Jak Nosorożec nabawił się swojej skóry. In M. Feldmanowa (Trans.), *Takie sobie bajeczki*. Warsaw: Prószyński i S-ka (Original work published 1902).
- Kipling, R. (2000b). Jak powstał garb Wielbłąda. In M. Feldmanowa (Trans.), *Takie sobie bajeczki*. Warsaw: Prószyński i S-ka (Original work published 1902).
- Kipling, R. (2000c). Kot, który zawsze chadzał własnymi drogami. In M. Feldmanowa (Trans.), *Takie sobie bajeczki*. Warsaw: Prószyński i S-ka (Original work published 1902).
- Kipling, R. (2000d). O motylu który tupał nogą. In M. Feldmanowa (Trans.), *Takie sobie bajeczki*. Warsaw: Prószyński i S-ka (Original work published 1902).
- Kipling, R. (2000e). O powstaniu Żółwców. In M. Feldmanowa (Trans.), *Takie sobie bajeczki*. Warsaw: Prószyński i S-ka (Original work published 1902).
- Kipling, R. (2000f). Opowieść o Starym Jegomościu Kangurze. In M. Feldmanowa (Trans.), *Takie sobie bajeczki*. Warsaw: Prószyński i S-ka (Original work published 1902).
- Kipling, R. (2000g). Słoniątko. In M. Feldmanowa (Trans.), *Takie sobie bajeczki*. Warsaw: Prószyński i S-ka (Original work published 1902).
- Kipling, R. (2000h). W jaki sposób Wieloryb nabawił się swego przełyku. In M. Feldmanowa (Trans.), *Takie sobie bajeczki*. Warsaw: Prószyński i S-ka (Original work published 1902).

- Kipling, R. (2000i). W jaki sposób został napisany pierwszy list. In M. Feldmanowa (Trans.), *Takie sobie bajeczki*. Warsaw: Prószyński i S-ka (Original work published 1902).
- Lawrence, D. H. (1926). Sun. *The New Coterie*, 4. Retrieved from <http://gutenberg.net.au/ebooks04/0400301h.html#s03>
- Lawrence, D. H. (2009). Sonca (J. Cimaŕiejeva, Trans.). *PrajdziSviet*, 2 (Original work published 1926). Retrieved from <http://prajdzisvet.org/text/155-sontsa.html>
- Lawrence, D. H. (2010). *Słońce* (J. Wronak, Trans.). Kraków: Woblink (Original work published 1926).
- Le Guin, U. (1982). Schrödinger's Cat. In *The Compass Rose*. New York, NY: Harper & Row.
- Le Guin, U. (1993). Kot Schrodingera. In A. Sylwanowicz (Trans.), *Dziewczyny Buffalo oraz inne zwierzęce obecności*. Warsaw: Alkazar (Original published 1982).
- Le Guin, U. (2001). Kot Šrodynhiera (Aleš Kudraŭcaŭ, Trans.). *Krynica*, 9(69), 109–117 (Original work published 1982).
- Lewis, C. S. (1942). *The Screwtape Letters*. London: Geoffrey Bles.
- Lewis, C. S. (1955). *The Magician's Nephew*. Retrieved from <https://archive.org/stream/TheMagiciansNephew/TheMagiciansNephew.txt>
- Lewis, C. S. (1990). *Listy starego diabła do młodego* (S. Pietraszko, Trans.). Warsaw: Instytut Wydawniczy PAX (Original work published 1942).
- Lewis, C. S. (1998). *Kroniki Narnii. Siostrzeniec czarodzieja* (A. Polkowski, Trans.). Poznań: Media Rodzina (Original work published 1955).
- Lewis, C. S. (2015). *Listy Kruciala* (F. Korzun, Trans.). Minsk: Halijafy (Original work published 1942).
- Lewis, C. S. (2016). *Chroniki Narnii. Plamiennik čaraŭnika*. (A. Kim, Trans.). Minsk: Pazityŭ-centr (Original work published 1955).
- London, J. (1915). *White Fang*. Retrieved from <http://www.gutenberg.org/files/910/910-h/910-h.htm>
- London, J. (1926). *Biały kiel* (S. Kuszelewska, Trans.). Warsaw: Towarzystwo Wydawnicze „Rój” (Original work published 1906).
- London, J. (1939). *Bieły klyk* (J. Pflaŭbaŭm, Trans.) (Original work published 1906). Retrieved from [http://knihi.com/Dzek\\_Londan/Biely\\_klyk.html#1](http://knihi.com/Dzek_Londan/Biely_klyk.html#1)
- Lovecraft, H. P. (1916). The Alchemist. *The United Amateur*, 4(16). Retrieved from <http://www.hplovecraft.com/writings/texts/fiction/a.aspx>
- Lovecraft, H. P. (1920a). Nyarlathotep. *The United Amateur*, 20(2). Retrieved from <http://www.hplovecraft.com/writings/texts/fiction/n.aspx>

- Lovecraft, H. P. (1920b). *Polaris*. *Philosopher*, 1(1). Retrieved from <https://en.wikisource.org/wiki/Polaris>
- Lovecraft, H. P. (1920c). The Cats of Ulthar. *The Tryout*, 6. Retrieved from <http://www.hplovecraft.com/writings/texts/fiction/cu.aspx>
- Lovecraft, H. P. (1922a). Herbert West—Reanimator. *Home Brew*, 1(1–6). Retrieved from <http://www.hplovecraft.com/writings/texts/fiction/hwr.aspx>
- Lovecraft, H. P. (1922b). The Music of Erich Zann. *The National Amateur*, 44(4). Retrieved from [https://en.wikisource.org/wiki/The\\_National\\_Amateur/Volume\\_44/Number\\_4/The\\_Music\\_of\\_Erich\\_Zann](https://en.wikisource.org/wiki/The_National_Amateur/Volume_44/Number_4/The_Music_of_Erich_Zann)
- Lovecraft, H. P. (1922c). The Tomb. *The Vagrant*, 14. Retrieved from [https://en.wikisource.org/wiki/The\\_Tomb](https://en.wikisource.org/wiki/The_Tomb)
- Lovecraft, H. P. (1923a). The Lurking Fear. *Home Brew*, 2; 3(6; 1–3). Retrieved from [https://en.wikisource.org/wiki/The\\_Lurking\\_Fear](https://en.wikisource.org/wiki/The_Lurking_Fear)
- Lovecraft, H. P. (1923b). What the Moon Brings. *The National Amateur*, 45(5). Retrieved from [https://en.wikisource.org/wiki/What\\_the\\_Moon\\_Brings](https://en.wikisource.org/wiki/What_the_Moon_Brings)
- Lovecraft, H. P. (1924a). The Hound. *Weird Tales*, 3(2). Retrieved from <http://www.hplovecraft.com/writings/texts/fiction/h.aspx>
- Lovecraft, H. P. (1924b). The Rats in the Walls. *Weird Tales*, 3(3). Retrieved from [https://en.wikisource.org/wiki/The\\_Rats\\_in\\_the\\_Walls](https://en.wikisource.org/wiki/The_Rats_in_the_Walls)
- Lovecraft, H. P. (1925). Festival. *Weird Tales*, 5(1). Retrieved from [https://en.wikisource.org/wiki/The\\_Festival\\_\(H.\\_P.\\_Lovecraft\)](https://en.wikisource.org/wiki/The_Festival_(H._P._Lovecraft))
- Lovecraft, H. P. (1926). The Outsider. *Weird Tales*, 7(4). Retrieved from <http://www.hplovecraft.com/writings/texts/fiction/o.aspx>
- Lovecraft, H. P. (1927). Pickman's Model. *Weird Tales*, 10(4), 505–514.
- Lovecraft, H. P. (1928). The Call of Cthulhu. *Weird Tales*, 11(2). Retrieved from <http://www.hplovecraft.com/writings/texts/fiction/cc.aspx>
- Lovecraft, H. P. (1983a). Muzyka Ericha Zanna. In R. Grzybowska (Trans.), *Zew Cthulhu*. Warsaw: Czytelnik (Original work published 1922).
- Lovecraft, H. P. (1983b). Zew Cthulhu. In R. Grzybowska (Trans.), *Zew Cthulhu*. Warsaw: Czytelnik (Original work published 1928).
- Lovecraft, H. P. (1995a). Alchemik. In R. P. Lipski (Trans.), *Reanimator*. Warsaw: Wydawnictwo S. R. (Original work published 1916).
- Lovecraft, H. P. (1995b). Reanimator. In R. P. Lipski (Trans.), *Reanimator*. Warsaw: Wydawnictwo S. R. (Original work published 1922).

- Lovecraft, H. P. (1999a). Model Pickmana. In A. Ledwożyw (Trans.), *Przypadek Charlesa Dextera Warda*. Warsaw: Wydawnictwo S. R. (Original work published 1927).
- Lovecraft, H. P. (1999b). Ogar. In R. Lipski (Trans.), *Coś na progu*. Poznań: Zysk i Spółka (Original work published 1922).
- Lovecraft, H. P. (1999c). Przybysz. In R. P. Lipski (Trans.), *Coś na progu*. Poznań: Zysk i Spółka (Original work published 1926).
- Lovecraft, H. P. (1999d). Przyczajona groza. In R. P. Lipski (Trans.), *Coś na progu*. Poznań: Zysk i Spółka (Original work published 1926).
- Lovecraft, H. P. (1999d). Szczury w murach. In A. Ledwożyw (Trans.), *Przypadek Charlesa Dextera Warda*. Warsaw: Wydawnictwo S. R. (Original work published 1924).
- Lovecraft, H. P. (2001). Čužanica (Alaksandr Kudraŭcaŭ, Trans.). *Krynica*, 9(69), 84–90 (Original work published 1926).
- Lovecraft, H. P. (2009). Klič Ktułchu (P. Donaŭ, Trans.). *PrajdziSviet*, 3 (Original work published 1928). Retrieved from <http://prajdzisvet.org/text/284-klich-ktulkhu.html>
- Lovecraft, H. P. (2010). Ałchimik (P. Donaŭ, Trans.). *PrajdziSviet*, 4 (Original work published 1916). Retrieved from <http://prajdzisvet.org/text/396-alkhimik.html>
- Lovecraft, H. P. (2014a). Hančak (U. Hurynovič, Trans.). *PrajdziSviet*, 14 (Original work published 1922). Retrieved from <http://prajdzisvet.org/texts/prose/ganchak.html>
- Lovecraft, H. P. (2014b). Katy Ułtara (U. Hurynovič, Trans.). *PrajdziSviet*, 14 (Original work published 1920). Retrieved from <http://prajdzisvet.org/texts/prose/katy-ultara.html>
- Lovecraft, H. P. (2014c). Padstupny Chaos (U. Hurynovič, Trans.). *PrajdziSviet*, 14 (Original work published 1920). Retrieved from <http://prajdzisvet.org/texts/prose/padstupny-xaos.html>
- Lovecraft, H. P. (2017a). Co sprowadza księżyc. In M. Kopacz (Trans.), *Bestia w jaskini i inne opowiadania* (pp. 207–210). Poznań: Zysk i Spółka (Original work published 1923).
- Lovecraft, H. P. (2017b). Fest. In U. Hurynovič (Trans.), *Klič Ktułchu: Apaviadanni*. Minsk: A. M. Januškievič (Original work published 1925).
- Lovecraft, H. P. (2017c). Festyn. In R. Lipski (Trans.), *Bestia w jaskini i inne opowiadania* (pp. 229–240). Poznań: Zysk i Spółka (Original work published 1925).
- Lovecraft, H. P. (2017d). Grobowiec. In R. Lipski (Trans.), *Bestia w jaskini i inne opowiadania* (pp. 23–34). Poznań: Zysk i Spółka (Original work published 1922).
- Lovecraft, H. P. (2017e). Hierbiert Uest, reanimatar. In U. Hurynovič (Trans.), *Klič Ktułchu: Apaviadanni*. Minsk: A. M. Januškievič (Original work published 1922).
- Lovecraft, H. P. (2017f). Koty Ultharu. In R. Lipski (Trans.), *Bestia w jaskini i inne opowiadania* (pp. 103–106). Poznań: Zysk i Spółka (Original work published 1920).

- Lovecraft, H. P. (2017g). Muzyka Eryka Cana. In U. Hurynovič (Trans.), *Kliĉ Ktulchu: Apaviadanni*. Minsk: A. M. Januškievič (Original work published 1922).
- Lovecraft, H. P. (2017h). Naturščyk Pikmana. In U. Hurynovič (Trans.), *Kliĉ Ktulchu: Apaviadanni*. Minsk: A. M. Januškievič (Original work published 1927).
- Lovecraft, H. P. (2017i). Nyarlathotep. In M. Kopacz (Trans.), *Bestia w jaskini i inne opowiadania* (pp. 159–162). Poznań: Zysk i Spółka (Original work published 1920).
- Lovecraft, H. P. (2017j). Pacuki u ścienach. In U. Hurynovič (Trans.), *Kliĉ Ktulchu: Apaviadanni*. Minsk: A. M. Januškievič (Original work published 1924).
- Lovecraft, H. P. (2017g). Plarys. In U. Hurynovič (Trans.), *Kliĉ Ktulchu: Apaviadanni*. Minsk: A. M. Januškievič (Original work published 1924).
- Lovecraft, H. P. (2017k). Polaris. In R. Lipski (Trans.), *Bestia w jaskini i inne opowiadania* (pp. 41–46). Poznań: Zysk i Spółka (Original work published 1920).
- Lovecraft, H. P. (2017l). Sklep. In P. Donaŭ (Trans.), *Kliĉ Ktulchu: Apaviadanni*. Minsk: A. M. Januškievič (Original work published 1922).
- Lovecraft, H. P. (2017m). Što prynosić poŭnia. In U. Hurynovič (Trans.), *Kliĉ Ktulchu: Apaviadanni*. Minsk: A. M. Januškievič (Original work published 1923).
- Lovecraft, H. P. (2017n). Stojeny Źach. In U. Hurynovič (Trans.), *Kliĉ Ktulchu: Apaviadanni*. Minsk: A. M. Januškievič (Original work published 1923).
- Mansfield, K. (1921, November 28). Her First Ball. *Weekly Westminster Gazette*. Retrieved from <http://www.katherinemansfieldsociety.org/assets/KM-Stories/HER-FIRST-BALL1921.pdf>
- Mansfield, K. (1922). The Singing Lesson. In *The Garden Party and Other Stories*. Retrieved from <http://www.katherinemansfieldsociety.org/assets/KM-Stories/THE-SINGING-LESSON1920.pdf>
- Mansfield, K. (2008a). *Jej pierwszy bal* (B. Kopelówna, Trans.) (Original work published 1921). Retrieved from [http://www.nexto.pl/ebooki/jej\\_pierwszy\\_bal\\_p21833.xml](http://www.nexto.pl/ebooki/jej_pierwszy_bal_p21833.xml)
- Mansfield, K. (2008b). *Lekcja śpiewu* (B. Kopelówna, Trans.) (Original work published 1922). Retrieved from [http://www.nexto.pl/ebooki/lekcja\\_spiewu\\_p22692.xml](http://www.nexto.pl/ebooki/lekcja_spiewu_p22692.xml)
- Mansfield, K. (2017a). Pierśy bał (M. Cišucinaja & M. Marozavaja, Trans.). *Maladość*, 10, 63–75 (Original work published 1921).
- Mansfield, K. (2017b). Urok śpievaŭ (M. Cišucinaja & M. Marozavaja, Trans.). *Maladość*, 10, 63–75 (Original work published 1922).
- Mathews, H. (1988). *Singular Pleasures*. New York, NY: Grand Street Publications.
- Mathews, H. (1998). Pryvatnyja pryjemnaści (J. Marholin, Trans.). *Arche*, 1, 142–151 (Original work published 1988).

- Mathews, H. (2008). *Osobne przyjemności*. In K. Koziół (Trans.), *Osobne przyjemności*. Wrocław: Biuro Literackie (Original published 1988).
- Maugham, W. S. (1921a). Red. In *The Trembling of a Leaf: Little Stories of the South Sea Islands*. Retrieved from <http://www.gutenberg.org/files/26854/26854-h/26854-h.htm#IV>
- Maugham, W. S. (1921b). The Fall of Edward Barnard. In *The Trembling of a Leaf: Little Stories of the South Sea Islands*. Retrieved from <http://www.gutenberg.org/files/26854/26854-h/26854-h.htm#III>
- Maugham, W. S. (1958a). Rudy. In J. Sujkowska (Trans.), *Honolulu*. Warsaw: Czytelnik (Original work published 1921).
- Maugham, W. S. (1958b). Upadek Edwarda Barnarda. In J. Sujkowska (Trans.), *Honolulu*. Warsaw: Czytelnik (Original work published 1921).
- Maugham, W. S. (1990a). Padziennic Edwarda Barnarda. In V. Wałynski (Trans.), *Apaviadanni*. Minsk: Mastackaja litaratura (Original work published 1921).
- Maugham, W. S. (1990b). Rudy. In V. Wałynski (Trans.), *Apaviadanni*. Minsk: Mastackaja litaratura (Original work published 1921).
- McCullers, C. (1941). *Reflections in a Golden Eye*. Boston, MA: Houghton Mifflin Harcourt.
- McCullers, C. (1943). The Ballad of the Sad Café. *Harper's Bazaar*, LXXVII, 72–75.
- McCullers, C. (1967). *W zwierciadle złotego oka* (M. Skroczyńska, Trans.). Warsaw: Czytelnik (Original work published 1941).
- McCullers, C. (1970). *Ballada o smutnej knajpie* (K. Jurasz-Dąbska, Trans.). Warsaw: Państwowy Instytut Wydawniczy (Original work published 1943).
- McCullers, C. (1988a). Adlustravanni ũ załatym voku. In U. Ščasny (Trans.), *Balada pra sumnaje kafe* (pp. 83–190). Minsk: Mastackaja litaratura (Original work published 1941).
- McCullers, C. (1988b). Bałada pra sumnaje kafe. In U. Ščasny (Trans.), *Balada pra sumnaje kafe* (pp. 9–82). Minsk: Mastackaja litaratura (Original work published 1943).
- McCullough, C. (1977). *The Thorn Birds*. New York, NY: Harper & Row.
- McCullough, C. (1988). *Ptuški na cierniach* (S. Dorski & V. Rabkievič, Trans.). Minsk: Mastackaja litaratura (Original work published 1977).
- McCullough, C. (2015). *Ptaki ciernistych krzewów* (M. Grabowska & I. Zych, Trans.). Warsaw: Świat Książki (Original published 1977).
- Milne, A. A. (1926). *Winnie the Pooh*. Retrieved from [https://archive.org/stream/AAMilneWinnieThePooh/A\\_A\\_Milne\\_-\\_Winnie-the-Pooh\\_djvu.txt](https://archive.org/stream/AAMilneWinnieThePooh/A_A_Milne_-_Winnie-the-Pooh_djvu.txt)
- Milne, A. A. (1938). *Kubuś puchatek* (I. Tuwim, Trans.). Warsaw: J. Przeworski (Original work published 1926).

- Milne, A. A. (2007). *Vinia-Pych* (V. Voranaŭ, Trans.). Poznań: Bieły Krumkač (Original work published 1926).
- Orwell, G. (1945). *Animal Farm*. Retrieved from [http://www.george-orwell.org/Animal\\_Farm/index.html](http://www.george-orwell.org/Animal_Farm/index.html)
- Orwell, G. (1949). *1984*. Retrieved from <http://www.george-orwell.org/1984>
- Orwell, G. (2000a). *1984* (S. Šupa, Trans.) (Original work published 1949). Retrieved from [http://knihi.com/Dzordz\\_Oruel/1984.html](http://knihi.com/Dzordz_Oruel/1984.html)
- Orwell, G. (2000b). *Fierma* (S. Šupa & A. Minkin, Trans.) (Original work published 1945). Retrieved from [http://knihi.com/Dzordz\\_Oruel/Fierma.html](http://knihi.com/Dzordz_Oruel/Fierma.html)
- Orwell, G. (2006). *Folwark zwierzęcy* (B. Zborski, Trans.). Warsaw: Wydawnictwo Literackie MUZA SA (Original work published 1945).
- Orwell, G. (2014). *Rok 1984* (T. Mirkowicz, Trans.). Warsaw: Wydawnictwo Literackie MUZA SA (Original work published 1949).
- Palahniuk, C. (1996). *Fight Club*. New York, NY: W.W. Norton.
- Palahniuk, C. (2001). *Podziemny krąg* (J. Manicki, Trans.). Warsaw: Laguna (Original work published 1996).
- Palahniuk, C. (2017). *Bajcoŭski klub* (S. Miadźviedzieŭ, Trans.). Minsk: Łohvinaŭ (Original work published 1996).
- Parsons, T. (1999). *Man and Boy*. London: HarperCollins.
- Parsons, T. (2001). *Mężczyzna i chłopiec* (A. Szulc, Trans.). Warsaw: Libros (Original work published in 1999).
- Parsons, T. (2017). *Mužčyna i chłopčyk* (V.K., Trans.). Minsk: kniharnia.by (Original work published in 1999).
- Potter, B. (1902). *The Tale of Peter Rabbit*. Retrieved from <https://archive.org/stream/thetaleofpeterra14838gut/14838.txt>
- Potter, B. (1991). *Piotruś Królik* (M. Musierowicz, Trans.). Wrocław: Wydawnictwo Siedmioróg (Original work published 1902).
- Potter, B. (2012). *Kazka pra trusika Pitera* (A. Bašarymava, Trans.). *PrajdziSviet, 10* (Original work published 1902). Retrieved from <http://prajdzisvet.org/text/1169-kazka-pra-trusika-pitera.html>
- Rowling, J. K. (1997). *Harry Potter and the Philosopher's Stone*. London: Bloomsbury.
- Rowling, J. K. (1998). *Harry Potter and the Chamber of Secrets*. London: Bloomsbury.
- Rowling, J. K. (1999). *Harry Potter and the Prisoner of Azkaban*. London: Bloomsbury.
- Rowling, J. K. (2000a). *Harry Potter and the Goblet of Fire*. London: Bloomsbury.

- Rowling, J. K. (2000b). *Harry Potter i Kamień Filozoficzny* (A. Polkowski, Trans.). Poznań: Media Rodzina (Original work published 1997).
- Rowling, J. K. (2000c). *Harry Potter i Komnata Tajemnic* (A. Polkowski, Trans.). Poznań: Media Rodzina (Original work published 1998).
- Rowling, J. K. (2001a). *Harry Potter i Czara Ognia* (A. Polkowski, Trans.). Poznań: Media Rodzina (Original work published 2000).
- Rowling, J. K. (2001b). *Harry Potter i więzień Azkabanu* (A. Polkowski, Trans.). Poznań: Media Rodzina (Original work published 1999).
- Rowling, J. K. (2005). *Harry Potter and the Half-Blood Prince*. London: Bloomsbury.
- Rowling, J. K. (2006). *Harry Potter i Księżę Półkrwi* (A. Polkowski, Trans.). Poznań: Media Rodzina (Original work published 2005).
- Rowling, J. K. (2009). *Hary Poter i Prync Paŭkroŭka* (A. Bahač, J. Hiedranovič, J. Marozava, T. Studnieva, & I. Bondar, Trans.) (Original work published 2005). Retrieved from <https://files.belpotter.by/books/by.hp.3.pdf>
- Rowling, J. K. (2013a). *Hary Poter i Filasofski kamień* (D. Muski, Trans.) (Original work published 1997). Retrieved from <https://files.belpotter.by/books/by.hp.1.pdf>
- Rowling, J. K. (2013b). *Hary Poter i patajemnaja zała* (D. Muski, Trans.) (Original work published 1998). Retrieved from <https://files.belpotter.by/books/by.hp.2.pdf>
- Rowling, J. K. (2013c). *Hary Poter i Viazień Azkabana* (D. Muski, Trans.) (Original work published 1999). Retrieved from <https://files.belpotter.by/books/by.hp.3.pdf>
- Saki. (1914). The Open Windon. In *Beasts and Super-Beasts*. Retrieved from <http://www.eastoftheweb.com/short-stories/UBooks/OpeWin.shtml>
- Saki. (1919). Tea. In *The Toys of Peace*. Retrieved from <http://www.eastoftheweb.com/short-stories/UBooks/Tea.shtml>
- Saki. (1974). Otwarte okno. In T. Malanowski (Trans.), *Tchnienie grozy* (pp. 154–158). Poznań: Wydawnictwo Poznańskie (Original work published 1914).
- Saki. (1986). Herbata. In E. Petrajtis-O'Neill (Trans.), *Małomówność lady Anny i inne opowiadania*. Warsaw: Czytelnik (Original work published 1919).
- Saki. (2009). Adčynienaje akno (A. Kazłova, Trans.). *PrajdziSviet*, 3 (Original work published 1914). Retrieved from <http://prajdzisvet.org/text/306-adchynienaje-akno.html>
- Saki. (2010). Harbata (J. Prystaŭka, Trans.). *PrajdziSviet*, 5–6 (Original work published 1919). Retrieved from <http://prajdzisvet.org/text/379-harbata.html>
- Segal, E. (1970). *Love Story*. New York, NY: Harper & Row.

- Segal, E. (1989). *Love Story czyli o miłości* (A. Przedpeńska-Trzeciakowska, Trans.). Warsaw: Iskry (Original work published 1970).
- Segal, E. (1993). *Historyja kachannia* (A. Astašonak, Trans.). *Krynica*, 8–9 & 10 (Original work published 1970).
- Sheckley, R. (1954). *The Battle*. *If*, 9, 52–56.
- Sheckley, R. (1988). *Bitwa*. In L. Jęczyk (Trans.), *Pielgrzymka na ziemię*. Warsaw: Wydawnictwo Iskry (Original work published 1954).
- Sheckley, R. (1990). *Bitva*. In S. Michałčuk (Trans.), *Zamiežnaja fantatyka*. Minsk: Mastackaja litaratura (Original work published 1954).
- Steinbeck, J. (1939). *The Grapes of Wrath*. New York, NY: The Viking Press.
- Steinbeck, J. (1993). *Hronki gniewu* (Siamion Dorski, Trans.). Minsk: Mastackaja litaratura (Original work published 1939).
- Steinbeck, J. (2012). *Grona gniewu* (A. Liebfeld, Trans.). Warsaw: Prószyński i S-ka (Original work published 1939).
- Tolkien, J. R. R. (1937). *The Hobbit, or There and Back Again*. London: Allen & Unwin.
- Tolkien, J. R. R. (1954a). *The Fellowship of the Ring*. London: Allen & Unwin.
- Tolkien, J. R. R. (1954b). *The Two Towers*. London: Allen & Unwin.
- Tolkien, J. R. R. (1955). *The Return of the King*. London: Allen & Unwin.
- Tolkien, J. R. R. (1960). *Hobbit, czyli tam i z powrotem* (M. Skibniewska, Trans.). Warsaw: Wydawnictwo Iskry (Original work published 1937).
- Tolkien, J. R. R. (1967). *Smith of Wootton Major*. London: Allen & Unwin.
- Tolkien, J. R. R. (1996a). *Drużyna pierścienia* (M. Skibniewska, Trans.). Warsaw: Muza (Original published in 1954).
- Tolkien, J. R. R. (1996b). *Dwie wieże* (M. Skibniewska, Trans.). Warsaw: Muza (Original published in 1954).
- Tolkien, J. R. R. (1996c). *Powrót króla* (M. Skibniewska, Trans.). Warsaw: Muza (Original published in 1955).
- Tolkien, J. R. R. (1997). *Kowal z Podlesia Większego* (M. Skibniewska, Trans.). Warsaw: Amber (Original work published 1967).
- Tolkien, J. R. R. (2002). *Chobit, abo vandroŭka tudy i nazad* (D. Mahiloŭcava & T. Jafimava, Trans.). Minsk: (Original work published 1937).
- Tolkien, J. R. R. (2008a). *Džvie wieży* (D. Mahiloŭcaŭ & K. Kurčankova, Trans.). Minsk: (Original work published 1954).

- Tolkien, J. R. R. (2008b). *Zviaz piarścionka* (D. Mahiloŭcaŭ & K. Kurčankova, Trans.). Minsk: (Original work published 1954).
- Tolkien, J. R. R. (2009). *Viartannie karala* (D. Mahiloŭcaŭ & K. Kurčankova, Trans.). Minsk: (Original work published 1955).
- Travers, P. L. (1934). *Mary Poppins*. London: HarperCollins.
- Travers, P. L. (2007, February 8). *Mery Poppins* (Original work published 1934). Retrieved from *Naŭyja dzietki* website: [http://dzietki.org/article/cms\\_view\\_article.php?aid=304](http://dzietki.org/article/cms_view_article.php?aid=304)
- Travers, P. L. (2008). *Mary Poppins* (I. Tuwim, Trans.). Warsaw: Wydawnictwo Jaguar (Original work published 1934).
- van Gulik, R. (1967). He Came with the Rain. In *Judge Dee at Work*. Portsmouth, NH: Heinemann.
- van Gulik, R. (2010). Przychodził z deszczem. In E. Westwalewicz-Mogilska (Trans.), *Zagadki sędziogo Di* (pp. 67–108). Warsaw: Państwowy Instytut Wydawniczy (Original work published 1967).
- van Gulik, R. (2012). Jon prychodziŭ z daždžom (H. Jankuta, Trans.). *PrajdziSviet*, 9 (Original work published 1967). Retrieved from <http://prajdzisvet.org/text/1115-jon-prykhodziu-z-dazhdzhom.html>
- Waugh, E. (1936). Mr. Loveday's Little Outing. In *Mr. Loveday's Little Outing and Other Sad Stories*. Retrieved from [https://biblio.wiki/images/7/74/The\\_complete\\_stories\\_of\\_evelyn\\_waugh.pdf](https://biblio.wiki/images/7/74/The_complete_stories_of_evelyn_waugh.pdf)
- Waugh, E. (1957). Mała przechadzka pana Lovedaya. In B. Zieliński (Trans.), *Mała przechadzka pana Lovedaya i inne smutne opowiadania* (pp. 5–15). Warsaw: Pax (Original work published 1936).
- Waugh, E. (1994). Maleńkaja prahulanka mistera Łaŭdeja (V. Łuk'janaŭ, Trans.). *Naša Niva*, 3 (Original work published 1936).
- Woolf, V. (1920). Solid Objects. *The Athenaeum*, 4720. Retrieved from <http://gutenberg.net.au/ebooks02/0200781.txt>
- Woolf, V. (1921a). A Haunted House. In *Monday or Tuesday*. Retrieved from <http://www.bartleby.com/85/1.html>
- Woolf, V. (1921b). Monday or Tuesday. In *Monday or Tuesday*. Retrieved from <http://www.bartleby.com/85/3.html>
- Woolf, V. (1921c). The String Quartet. In *Monday or Tuesday*. Retrieved from <http://www.bartleby.com/85/5.html>
- Woolf, V. (1944). The Legacy. In *A Haunted House and Other Short Stories*. Retrieved from <https://biblioklept.org/2014/05/27/the-legacy-virginia-woolf/>

- Woolf, V. (2001a). Capraŭdnyja rečy (V.K., Trans.). *Krynica*, 11–12(71). Retrieved from <http://www.bartleby.com/85/3.html>
- Woolf, V. (2001b). Paniadziełak-aŭtorak (V.K., Trans.). *Krynica*, 11–12(71). Retrieved from <http://www.bartleby.com/85/3.html>
- Woolf, V. (2001c). Strunny kvartet (V.K., Trans.). *Krynica*, 11–12(71). Retrieved from <http://www.bartleby.com/85/3.html>
- Woolf, V. (2009). Dom z pryvidami (H. Ražancova, Trans.). *PrajdziSviet*, 2. Retrieved from <http://prajdzisvet.org/kit/12-dom-z-pryvidami.html>
- Woolf, V. (2012a). Kwartet smyczkowy. In M. Heydel (Trans.), *Nawiedzony dom. Opowiadania zebrane*. Kraków: Wydawnictwo Literackie (Original work published 1921).
- Woolf, V. (2012b). Nawiedzony dom. In M. Heydel (Trans.), *Nawiedzony dom. Opowiadania zebrane*. Kraków: Wydawnictwo Literackie (Original work published 1921).
- Woolf, V. (2012c). Poniedziałek lub wtorek. In M. Heydel (Trans.), *Nawiedzony dom. Opowiadania zebrane*. Kraków: Wydawnictwo Literackie (Original work published 1921).
- Woolf, V. (2012d). Rzeczy trwałe. In M. Heydel (Trans.), *Nawiedzony dom. Opowiadania zebrane*. Kraków: Wydawnictwo Literackie (Original work published 1920).
- Woolf, V. (2012e). Spadek. In M. Heydel (Trans.), *Nawiedzony dom. Opowiadania zebrane*. Kraków: Wydawnictwo Literackie (Original work published 1944).

### **Primary sources (B) – corpus of original Belarusian literary prose**

- Adamovič, A., Bryl, J., & Kalešnik, U. (1975). *Ja z vohniennaj wioski*. Retrieved from [http://knihi.com/Ales\\_Adamovic/Ja\\_z\\_vohniennaj\\_vioski.html](http://knihi.com/Ales_Adamovic/Ja_z_vohniennaj_vioski.html)
- Asipienka, A. (1979). Mirnaja vajna. In *Vybranyja tvory: U 2 t. T. 2: Apovieści i apaviadanni* (pp. 262–273). Retrieved from [http://knihi.com/Ales\\_Asipienka/Mirnaja\\_vajna.html](http://knihi.com/Ales_Asipienka/Mirnaja_vajna.html)
- Astraŭcoŭ, A. (2005). *Sula*. Retrieved from [http://knihi.com/Ales\\_Astraucou/Sula.html](http://knihi.com/Ales_Astraucou/Sula.html)
- Bacharevič, A. (2014). *Dzieci Alindarki*. Minsk: Halijafy.
- Broŭka, P. (1957). *Kali zlivajucca reki*. Retrieved from [http://knihi.com/Piatrus\\_Brouka/Kali\\_zlivajucca\\_reki.html](http://knihi.com/Piatrus_Brouka/Kali_zlivajucca_reki.html)
- Bryl, J. (1958). Memento mori. *Poŭymia*, 12. Retrieved from [http://knihi.com/Janka\\_Bryl/Memento\\_mori.html](http://knihi.com/Janka_Bryl/Memento_mori.html)
- Bryl, J. (1961). Usmieszka. *Poŭymia*, 11. Retrieved from [http://knihi.com/Janka\\_Bryl/Usmieszka.html](http://knihi.com/Janka_Bryl/Usmieszka.html)
- Bryl, J. (1975). Chłopczyk. *Maładość*, 8. Retrieved from [http://knihi.com/Janka\\_Bryl/Chlopczyk.html](http://knihi.com/Janka_Bryl/Chlopczyk.html)

- Bykaŭ, V. (1984). *Znak biady*. Retrieved from [http://knihi.com/Vasil\\_Bykau/Znak\\_biady.html](http://knihi.com/Vasil_Bykau/Znak_biady.html)
- Bykaŭ, V. (1997). Muzyka. In *Ściana*. Retrieved from [http://knihi.com/Vasil\\_Bykau/Muzyka.html](http://knihi.com/Vasil_Bykau/Muzyka.html)
- Bykaŭ, V. (2004). Afhaniec. *Dziejaslou*, 7. Retrieved from <http://dziejaslou.by/old/www.dziejaslou.by/inter/dzeja/dzeja.nsf/htmlpage/byk702ec.html?OpenDocument>
- Fiedarenka, A. (2009). Ničyje. In *Ničyje: Apowieści, raman*. Retrieved from [http://knihi.com/Andrej\\_Fiedarenka/Nicyje.html](http://knihi.com/Andrej_Fiedarenka/Nicyje.html)
- Ipatava, V. (2002). Załataja žryca Ašvinaŭ. In *Alhierdava dzida: Ramany*. Retrieved from [http://knihi.com/Volha\\_Ipatava/Zalataja\\_zryca\\_Asvinau.html](http://knihi.com/Volha_Ipatava/Zalataja_zryca_Asvinau.html)
- Kavaloŭ, P. (1959). *Hordaść*. Retrieved from [http://knihi.com/Paviel\\_Kavalou/Hordasc.html](http://knihi.com/Paviel_Kavalou/Hordasc.html)
- Klimkovč, M. (1994). *Miaža na dalahladzie*. Retrieved from [https://knihi.com/Maksim\\_Klimkovic/Miaza\\_pa\\_dalahladzie.html#1](https://knihi.com/Maksim_Klimkovic/Miaza_pa_dalahladzie.html#1)
- Rubleŭskaja, L. (2003). Serca marmurovaha anioła. In *Serca marmurovaha anioła: Apowieści, apoviadanni*. Retrieved from [http://knihi.com/Ludmila\\_Rubleuskaja/Serca\\_marmurovaha\\_anioła.html](http://knihi.com/Ludmila_Rubleuskaja/Serca_marmurovaha_anioła.html)
- Rubleŭskaja, L. (2013). Cieni zabytaha karnavalu. In *Nočy na Palabanskich młynach: Raman, apowieść, apaviadanni*. Retrieved from [http://knihi.com/Ludmila\\_Rubleuskaja/Cieni\\_zabytaha\\_karnavalu.html](http://knihi.com/Ludmila_Rubleuskaja/Cieni_zabytaha_karnavalu.html)

## Secondary sources

- Adamou, E. (2016). *A Corpus-Driven Approach to Language Contact: Endangered Languages in a Comparative Perspective* (Digital original edition). Mouton De Gruyter.
- Aijmer, K. (1996). *Conversational Routines in English: Convention and Creativity* (1 edition). Routledge.
- Aijmer, K. (2015). Pragmatic markers. In K. Aijmer & C. Rühlemann (Eds.), *Corpus Pragmatics: A Handbook* (pp. 29–51). Cambridge University Press.
- Aijmer, K., Altenberg, B., & Johansson, M. (Eds.). (1996). *Languages in Contrast: Papers from a Symposium on Text-based Cross-Linguistic Studies, Lund, 4-5 March 1994*. Chartwell-Bratt.
- Aksamitaŭ, A. (2000). *Słownik frazeologiczny białorusko-polski [Belarusian-Polish Phraseological Dictionary]*. Slawistyczny Ośrodek Wydawniczy.
- Al-Ajmi, H. (2004). A new English-Arabic parallel text corpus for lexicographic applications. *Lexikos*, 14. <https://doi.org/10.4314/lex.v14i1.51427>

- Algún, S. (2018, December 6). Review for Tesseract and Kraken OCR for text recognition. *Medium*. <https://medium.com/datadriveninvestor/review-for-tesseract-and-kraken-ocr-for-text-recognition-2e63c2adedd0>
- Alotaibi, & M, H. (2017). *Arabic-English Parallel Corpus: A New Resource for Translation Training and Language Teaching* (SSRN Scholarly Paper ID 3053572). Social Science Research Network. <https://doi.org/10.2139/ssrn.3053572>
- American National Corpus Project. (2015). *Open American National Corpus*. <http://www.anc.org/>
- Ananiadou, S., McNaught, J., & Thompson, P. (2012). *The English Language in the Digital Age* (G. Rehm & H. Uszkoreit, Eds.). Springer. <http://www.meta-net.eu/whitepapers/e-book/english.pdf/view>
- Andiamo! (2019). *Expansion and contraction factors*. Andiamo! The Language Professionals. <https://www.andiamo.co.uk/resources/expansion-and-contraction-factors/>
- Anthony, L. (2018). *AntConc* (3.5.7) [Linux]. Waseda University. <http://www.laurenceanthony.net/software>
- Arabski, J. (Ed.). (2010). *Polsko-angielski słownik frazeologiczny*. Społeczna Wyższa Szkoła Przedsiębiorczości i Zarządzania.
- Artsiomava, V. A. (2014). *Bielaruska-anhlijski ideahrafičny sloŭnik fraziealahizmaŭ z prastoravaj siemantykaj* [*Belarusian-English Ideaographic Dictionary of Phraseology with Spatial Semantics*]. Respublikanskij institut vysshej shkoly.
- Astapenia, R. (2016, October 14). The Belarusian Language in Education: A Reluctant Revival? *BelarusDigest*. <https://belarusdigest.com/story/the-belarusian-language-in-education-a-reluctant-revival/>
- Australian National Corpus Inc. (n.d.). *AusNC - Australian National Corpus*. Retrieved 8 January 2018, from <https://www.ausnc.org.au/>
- Awal, N. M., Ho-Abdullah, I., & Zainudin, I. S. (2014). Parallel Corpus as a Tool in Teaching Translation: Translating English Phrasal Verbs into Malay. *Procedia - Social and Behavioral Sciences*, 112, 882–887. <https://doi.org/10.1016/j.sbspro.2014.01.1245>
- Aynur Abdulnasurov. (2018). *Lingualeo—English language online*. <https://lingualeo.com/chooselanguage?returnUrl=>
- Baker, M. (1993). Corpus linguistics and Translation Studies. Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology: In honour of John Sinclair* (pp. 233–252). John Benjamins Publishing Company.
- Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2), 223–243.

- Baker, M. (1996). Corpus-based Translation Studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager* (pp. 175–186). John Benjamins Publishing Company.
- Baker, M. (2000). Towards a Methodology for Investigating the Style of a Literary Translator. *Target*, 12(2), 241–266. <https://doi.org/10.1075/target.12.2.04bak>
- Baker, M. (2004). A corpus-based view of similarity and difference in translation. *Journal of Corpus Linguistics*, 9(2), 167–193.
- Baker, P. (2006). *Using Corpora in Discourse Analysis*. Continuum.
- Baker, P. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh University Press.
- Baker, P. (Ed.). (2012). *Contemporary Corpus Linguistics*. Continuum.
- Baker, P. (2014). *Using Corpora to Analyze Gender*. Bloomsbury Academic.
- Baker, P. (2015). Does Britain need any more foreign doctors? Inter-analyst consistency and corpus-assisted (critical) discourse analysis. In N. Groom, M. Charles, & J. Suganthi (Eds.), *Corpora, Grammar and Discourse. In honour of Susan Hunston* (pp. 283–300). John Benjamins Publishing Company.
- Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of Corpus Linguistics*. Edinburgh University Press.
- Baranova, M. F. (2017). K voprosu o feminizacii russkogo i belorusskogo jazykov [About feminisation process of Russian and Belarusian languages]. *Lingvodidaktika: novye tekhnologii v obuchenii russkomu iazyku kak inostrannomu* [Linguistic didactics: new technologies in learning Russian as a foreign language], 4, 12–13.
- Baroni, M. (2009). Distributions in text. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (Vol. 2, pp. 803–821). De Gruyter Mouton. <https://doi.org/10.1515/9783110213881.2.803>
- Barszczewska, N. (2004). *Bielaruskaja emihracyja—Abaronca rodnaje movy* [Belarusian emigration—Protector of the mother tongue]. Institute of Belarusian Studies, Department of Applied Linguistics and East Slavonic Studies, University of Warsaw.
- Barycka, E., Smierzchalski, T., & Wrembel, M. (1997). *Polsko-angielski słownik frazeologiczny współczesnej terminologii politycznej i ekonomicznej*. Altravox Press.
- Barzilay, R., & McKeown, K. R. (2001). Extracting Paraphrases from a Parallel Corpus. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 50–57. <https://doi.org/10.3115/1073012.1073020>

- Becher, V. (2011). *Explicitation and implicitation intranlation. A corpus-based study of English-German and German-English translations of business texts* [University of Hamburg]. <https://d-nb.info/102042673X/34>
- Belarusian Embassy in Poland. (2017). *Archiwum wydarzeń Centrum Kulturalnego Białorusi w Polsce* [Archive of events by the Cultural Centre of Belarus in Poland]. [http://poland.mfa.gov.by/pl/bilateral\\_relations/cultural/centrum/archiwum/](http://poland.mfa.gov.by/pl/bilateral_relations/cultural/centrum/archiwum/)
- Belarusian National Technical University. (n.d.). *Corpus Albaruthenicum*. Retrieved 26 November 2017, from <http://grid.bntu.by/corpus/>
- Belarusian PEN-Centre. (2018). *Беларускі ПЭН-Цэнтр* [Belarusian PEN-Centre]. <https://pen-centre.by/>
- belpotter.by. (2018). *Belpotter.by*. <http://belpotter.by/>
- Bernardini, S. (2002). Think-aloud protocols in translation research: Achievements, limits, future prospects. *Target*, 13(2), 241–263. <https://doi.org/10.1075/target.13.2.03ber>
- Biber, D. (1992). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, 26(5–6), 331–345. <https://doi.org/10.1007/BF00136979>
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.
- Biber, D., Finegan, E., Johansson, S., Conrad, D. S., & Leech, G. (1999). *Longman Grammar of Spoken and Written English*. Longman.
- Bibliowiki. (2018). [https://biblio.wiki/wiki/Main\\_Page](https://biblio.wiki/wiki/Main_Page)
- Biel, Łucja. (2010). Corpus-based studies of legal language for translation purposes: Methodological and practical potential. *Reconceptualizing LSP. Online Proceedings of the XVII European LSP Symposium 2009*. XVII European LSP Symposium 2009, Aarhus. <https://www.asb.dk/fileadmin/www.asb.dk/isek/biel.pdf>
- Біларуская Палічка. (2017). *Біларуская Палічка—Біларуская электронная бібліятэка* [Belarusian Shelf—Belarusian digital library]. <http://knihi.com/>
- Biryła, M. V. (Ed.). (1987). *Слоўнік біларускай мовы. Арафграфія. Арафепія. Акцэнтуючыя. Словзмяніенне* [Belarusian language dictionary. Orthography. Orthoepy. Accentuation. Inflection]. Біларуская Савецкая Энцыклапедыя.
- Bisiada, M. (2017). Universals of editing and translation. In S. Hansen-Schirra, O. Czulo, & S. Hofmann (Eds.), *Empirical modelling of translation and interpreting* (pp. 241–275). Language Science Press.

- Bladyniec, T. (2013a). *Dzień Białoruskiej Wikipedii na Litwie [Day of Belarusian Wikipedia in Lithuania]*. Wikimedia. [https://pl.wikimedia.org/wiki/Dzie%C5%84\\_Bia%C5%82oruskiej\\_Wikipedii\\_na\\_Litwie](https://pl.wikimedia.org/wiki/Dzie%C5%84_Bia%C5%82oruskiej_Wikipedii_na_Litwie)
- Bladyniec, T. (2013b). *Dzień Białoruskiej Wikipedii w Czechach [Day of Belarusian Wikipedia in Czech Republic]*. Wikimedia. [https://pl.wikimedia.org/wiki/Dzie%C5%84\\_Bia%C5%82oruskiej\\_Wikipedii\\_w\\_Czechach](https://pl.wikimedia.org/wiki/Dzie%C5%84_Bia%C5%82oruskiej_Wikipedii_w_Czechach)
- Bladyniec, T. (2013c). *Dzień Białoruskiej Wikipedii w Polsce 2013 [Day of Belarusian Wikipedia in Poland 2013]*. Wikimedia. [https://pl.wikimedia.org/wiki/Dzie%C5%84\\_Bia%C5%82oruskiej\\_Wikipedii\\_w\\_Polsce\\_2013](https://pl.wikimedia.org/wiki/Dzie%C5%84_Bia%C5%82oruskiej_Wikipedii_w_Polsce_2013)
- Bladyniec, T. (2015). *Dzień Białoruskiej Wikipedii w Polsce 2015 [Day of Belarusian Wikipedia in Poland 2015]*. Wikimedia. [https://pl.wikimedia.org/wiki/Dzie%C5%84\\_Bia%C5%82oruskiej\\_Wikipedii\\_w\\_Polsce\\_2015](https://pl.wikimedia.org/wiki/Dzie%C5%84_Bia%C5%82oruskiej_Wikipedii_w_Polsce_2015)
- Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-Kulka (Eds.), *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies* (pp. 17–35). Gunter Narr.
- Blum-Kulka, S., & Levenston, E. (1978). Universals of lexical simplification. *Language Learning*, 28(2), 399–415.
- BNkorpus. (2017). *Biblijny korpus [Biblical corpus]*. <http://biblija.bnkorpus.info/index.html>
- BNkorpus. (2019). *Biełaruski N-korpus [Belarusian N-corpus]*. <http://bnkorpus.info/>
- Borowska, P. (2014, April 29). *Linguistic Initiatives Conquer Belarus*. New Eastern Europe. <http://neweasterneurope.eu/2014/04/29/linguistic-initiatives-conquer-belarus/>
- Bowker, L. (1998). Using specialized monolingual native-language corpora as a translation resource: A pilot study. *Meta*, 43(4), 631–651.
- Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3), 12:1-12:22. <https://doi.org/10.1145/2382448.2382450>
- Broniarek, W. (2005). *Gdy Ci słowa zabraknie: Słownik synonimów [When you run out of words: A dictionary of synonyms]*. Haroldson Press. <https://www.synonimy.pl/szukaj/>
- Brown, P. F., Lai, J. C., & Mercer, R. L. (1991). Aligning sentences in parallel corpora. *Meeting of the Association for Computational Linguistics.*, 169–176.
- Čarniakievič, C. (2010). Pierakłady tvoraŭ zamiežnaj litaratury Ź časopisie ‘Krynica’ (1988-2003) [Foreign literary works translations in the “Krynica” journal (1998-2003)]. *ПраўдзіСвет*. <http://prajdzisvet.org/artykulyi/seminar/3156.html>

- Castro, O. (2013). Introduction: Gender, language and translation at the crossroads of disciplines. *Gender and Language*, 7(1), 5–12. <https://doi.org/10.1558/genl.v7i1.5>
- Center for Sprogteknologi. (2018). *CST's Lemmatiser*. <https://cst.dk/online/lemmatiser/uk/>
- Centre for English Corpus Linguistics (CECL). (2018). *The Louvain Corpus of Native English Essays (LOCNESS)*. Learner Corpus Association. <http://www.learnercorpusassociation.org/resources/tools/locness-corpus/>
- Cheesman, T., Flanagan, K., Thiel, S., & et al. (2012). *Translation Array Prototype 1*. <http://www.delightedbeauty.org/vvv/>
- Cheesman, T., Flanagan, K., Thiel, S., Rybicki, J., Laramée, R. S., Hope, J., & Roos, A. (2017). Multi-Retranslation corpora: Visibility, variation, value, and virtue. *Digital Scholarship in the Humanities*, 32(4), 739–760. <https://doi.org/10.1093/lc/fqw027>
- Cheng, W., Greaves, C., & Warren, M. (2005). The creation of a prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic). *ICAME Journal*, 29, 47–68.
- Cheong, H.-J. (2006). Target Text Contraction in English-into-Korean Translations: A Contradiction of Presumed Translation Universals? *Meta*, 51(2), 343–367. <https://doi.org/10.7202/013261ar>
- Chesterman, A. (2004). Hypotheses about translation universals. In G. Hansen, K. Malmkjær, & D. Gile (Eds.), *Claims, changes and challenges in translation studies: Selected contributions from the EST congress, Copenhagen 2001* (pp. 1–13). Benjamins.
- Chujo, K., Utiyama, M., & Miura, S. (2006). Using a Japanese-English Parallel Corpus for Teaching English Vocabulary to Beginning-Level Students. *English Corpus Studies*, 13, 153–172.
- Clancy, B., & McCarthy, M. (2015). Co-constructed turn-taking. In K. Aijmer & C. Rühlemann (Eds.), *Corpus Pragmatics: A Handbook* (pp. 430–453). Cambridge University Press.
- CLARIN. (2018). *Paweł Kamocki*. <https://www.clarin.eu/view-contact/858/36>
- CLARIN ERIC. (2018). *CLARIN VLO*. <https://vlo.clarin.eu/?0>
- CLARIN ERIC. (2019). *Depositing Services*. <https://www.clarin.eu/content/depositing-services>
- CLARIN ERIC. (2020, March 19). *CLARIN Workshop Sophia*. Oralhistory.Eu. <https://oralhistory.eu/workshops/clarin-2019>
- CLARIN PL. (2014). *Spokes. Conversational data search*. <http://spokes.clarin-pl.eu/>
- CLARIN PL. (2015). *Clarin PL*. <https://clarin-pl.eu/pl/strona-glowna/>
- CLARIN PL. (2018). *Morpho-syntactic tagger*. CLARIN PL. <https://ws.clarin-pl.eu/tagger.shtml?en>
- CLARIN-PL project. (2014). *Paralela Web*. <http://paralela.clarin-pl.eu/>

- CLARIN-UK. (2018). *Resources*. <http://www.clarin.ac.uk/resources>
- College of Eastern Europe. (2018). *Biblioteka bialoruska [Belarusian library]*. KEW. <http://www.kew.org.pl/kategoria-produktu/biblioteka-bialoruska/>
- Convert PDF to Text Online. (2018). PDF to Text. <https://pdftotext.com/>
- CoolLib. (2018). <https://coollib.net/>
- Corness, P. J. (2014). ‘Horror Vacui’? Explication in a Polish Translation of Marie Heaney’s ‘Over Nine Waves’. *Translation Ireland*, 19(2), 7–27.
- Corpas Pastor, G., Mitkov, R., Afzal, N., & Pekar, V. (2008). Translation universals: Do they exist? A corpus-based NLP study of convergence and simplification. *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-08)*. <http://www.mt-archive.info/AMTA-2008-Corpas.pdf>
- Corpus of Canadian English (Strathy). (n.d.). Retrieved 9 January 2018, from <https://corpus.byu.edu/can/>
- Crawford, W., & Csomay, E. (2016). *Doing Corpus Linguistics*. Routledge.
- Cresswell, A. (2018). Looking up phrasal verbs in small corpora of interpreting. An attempt to draw out aspects of interpreted language. *InTRAlinea, Special Issue: New Findings in Corpus-based Interpreting Studies*. <http://www.intraline.org/specials/article/2319>
- Culpeper, J. (2009). Keyness: Words, Parts-of-Speech and Semantic Categories in the Character-Talk of Shakespeare’s ‘Romeo and Juliet’. *International Journal of Corpus Linguistics*, 14(1), 29–59.
- Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. College-Hill Press.
- Cyfroteka. (2013). <http://cyfroteka.pl>
- Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. <https://corpus.byu.edu/coca/>
- Davies, M. (2010). *The 400 million word Corpus of Historical American English (1810–2009)*. English Historical Linguistics 2010. <https://corpus.byu.edu/coha/>
- Davies, M. (2013). *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE)*. <https://corpus.byu.edu/glowbe/>
- Deutscher Wortschatz. (2018). *Leipzig Corpora Collection—Wortschatz Belarusian*. [http://corpora.uni-leipzig.de/de?corpusId=bel\\_newscrawl\\_2011](http://corpora.uni-leipzig.de/de?corpusId=bel_newscrawl_2011)
- Dewey, M., & Cogo, A. (2012). *Analysing English as a Lingua Franca: A Corpus-driven Investigation*. Continuum.

- Diani, G. (2015). Politeness. In K. Aijmer & C. Rühlemann (Eds.), *Corpus Pragmatics: A Handbook* (pp. 169–192). Cambridge University Press.
- Dice, L. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3), 297–302. <https://doi.org/10.2307/1932409>
- Domański, P. (1992). *Naukowo-techniczny słownik frazeologiczny angielsko-polski, polsko-angielski*. Wydawnictwa Szkolne i Pedagogiczne.
- Dynko, A., & Bigg, C. (2014, July 2). Shocking! Belarusian President Speaks Belarusian. *RadioFreeEurope/RadioLiberty*. <https://www.rferl.org/a/shocking-belarusian-president-speaks-belarusian-lukashenka/25443432.html>
- Dziejasłoŭ. (2018). <http://dziejaslou.by/>
- Eastern Partnership Civil Society Forum. (2013). *Virtual “Week” in Sweden in the Belarusian Wikipedia*. <http://eng.npbelarus.info/?p=20>
- Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *The R Journal*, 8(1), 107–121.
- Eder, M., Rybicki, J., Kestemont, M., & Pielstroem, S. (2019). *Package ‘stylo’ manual*. <https://cran.r-project.org/web/packages/stylo/stylo.pdf>
- Esileht [Main page]. (2017). In *Wikipedia, the free encyclopedia*. <https://et.wikipedia.org/w/index.php?title=Esileht&oldid=4625418>
- European Commission. (2018). *The EU copyright legislation*. Digital Single Market. <https://ec.europa.eu/digital-single-market/en/eu-copyright-legislation>
- Evans, J. (2018). *On language equality in the digital age* (No. 2018/2028). Committee on Culture and Education, European Parliament. [https://www.europarl.europa.eu/doceo/document/A-8-2018-0228\\_EN.html](https://www.europarl.europa.eu/doceo/document/A-8-2018-0228_EN.html)
- Evas, J. (2014). Minority languages fight for survival in the digital age. *The Conversation*. <https://theconversation.com/minority-languages-fight-for-survival-in-the-digital-age-22571>
- Even-Zohar, I. (1990). The Position of Translated Literature within the Literary Polysystem. *Poetics Today*, 11(1), 45–51. <https://doi.org/10.2307/1772668>
- Evert, S., & Proisl, T. (2015). *Explaining Delta, or: How do distance measures for authorship attribution work?* Corpus Linguistics Conference, Lancaster University. [https://www.researchgate.net/publication/280088331\\_Explaining\\_Delta\\_or\\_How\\_do\\_distance\\_measures\\_for\\_authorship\\_attribution\\_work](https://www.researchgate.net/publication/280088331_Explaining_Delta_or_How_do_distance_measures_for_authorship_attribution_work)
- Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., & Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl\_2), ii4–ii16. <https://doi.org/10.1093/lle/fqx023>

- Faded Page. (2018). <https://www.fadedpage.com/>
- Falanster. (2013). *Wiki-day knowledge expansion*. <https://falanster.by/en/news/wiki-day-knowledge-expansion>
- Farkas, A. (2018). *LF Aligner* (4.2) [Computer software]. <https://sourceforge.net/projects/aligner/>
- Fattah, M. A., Ren, F., & Kuroiwa, S. (2006). Sentence alignment using feed forward neural network. *International Journal of Neural Systems*, 16(6), 423–434.
- Federal University of Santa Catarina. (2018). *Repositório Institucional da UFSC [Repository of the Federal University of Santa Catarina]*.
- Fenrich, W., Siewicz, K., & Szprot, J. (2016). *Towards Open Research Data in Poland [Report]*. Wydawnictwa ICM. <https://depot.ceon.pl/handle/123456789/12489>
- Fischer-Starcke, B. (2010). *Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries* (1 edition). Bloomsbury Academic.
- Franco, E. P. C., Matamala, A., & Orero, P. (2013). *Voice-over Translation: An Overview*. Peter Lang GmbH, Internationaler Verlag der Wissenschaften. <http://ebookcentral.proquest.com/lib/swansea-ebooks/detail.action?docID=1565059>
- Frankenberg-Garcia, A. (2009). Are translations longer than source texts?: A corpus-based study of explicitation. In A. Beeby, P. Inés, & P. Sánchez-Gijón (Eds.), *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate* (pp. 47–58). John Benjamins Publishers. <https://benjamins.com/catalog/btl.82.05fra>
- Frawley, W. (1984). Prolegomenon to a theory of translation. In W. Frawley (Ed.), *Translation: Literary, Linguistic and Philosophical Perspectives* (pp. 159–175). Associated University Press.
- Gale, W. A., & Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Comput. Linguist.*, 19(1), 75–102.
- Garcia McAllister, P. (2015). Speech acts: A synchronic perspective. In K. Aijmer & C. Rühlemann (Eds.), *Corpus Pragmatics: A Handbook* (pp. 29–51). Cambridge University Press.
- Gavagai AB. (2015, September 30). A brief history of word embeddings (and some clarifications). *Gavagai*. <https://www.gavagai.se/blog/2015/09/30/a-brief-history-of-word-embeddings/>
- Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindquist (Eds.), *Translation Studies in Scandinavia* (pp. 88–95). CWK Gleerup.
- Genius Media Group Inc. (2018). *Genius*. Genius. <https://genius.com/>
- GitHub, Inc. (2019). *GitHub*. GitHub. <https://github.com>
- Glagoslav Publications Ltd. (2018). *Glagoslav Publications*. <http://www.glagoslav.com/>

- Glavna stranica [Main page]. (2017). In *Wikipedia, the free encyclopedia*. [https://hr.wikipedia.org/w/index.php?title=Glavna\\_stranica&oldid=4836213](https://hr.wikipedia.org/w/index.php?title=Glavna_stranica&oldid=4836213)
- Goldwater, S., & McClosky, D. (2005). Improving Statistical MT through Morphological Analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 676–683. <https://www.aclweb.org/anthology/H05-1085>
- Google. (2018). *Google Books*. <https://books.google.com/>
- Gorgaini, E. (2020, March 12). New collections and resources available in the VLO. *News | CLARIN Eric*. <https://www.clarin.eu/blog/new-collections-and-resources-available-vlo>
- Grabowski, L. (2012). *A corpus-driven study of translational and non-translational texts: The case of Nabokov's Lolita*. Wydawnictwo Uniwersytetu Opolskiego.
- Grabowski, L. (2013). Interfacing corpus linguistics and computational stylistics. *International Journal of Corpus Linguistics*, 18(2), 254–280.
- Grabowski, Ł. (2018). Stance Bundles in English to Polish Translation: A Corpus Informed Study. *Russian Journal of Linguistics*, 22(2), 404–422. <https://doi.org/10.22363/2312-9182-2018-22-2-404-422>
- Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International Corpus of Learner English V2—I6doc*. Presses universitaires de Louvain. /en/book/?GCOI=28001100825990
- Greenbaum, S., & Svartvik, J. (1990). The London Corpus of Spoken English: Description and research. In J. Svartvik (Ed.), *Lund Studies in English* 82 (Vol. 82). Lund University Press. <http://clu.uni.no/icame/manuals/LONDLUND/INDEX.HTM>
- Groom, N., Charles, M., & John, S. (Eds.). (2015). *Corpora, Grammar and Discourse: In Honour of Susan Hunston* (UK ed. edition). John Benjamins Publishing Company.
- Gumul, E. (2020). Dlaczego tłumacz mówi więcej niż autor? O eksplicytacji w przekładzie [Why do translators say more than authors do? On explicitation in translation]. *Rocznik Przekładoznawczy*, 0(15), 175–195. <https://doi.org/10.12775/RP.2020.009>
- Hałoŭnaja staronka [Main page]. (2018). In *Wikipedia, the free encyclopedia*. [https://be.wikipedia.org/w/index.php?title=%D0%93%D0%B0%D0%BB%D0%BE%D1%9E%D0%BD%D0%B0%D1%8F\\_%D1%81%D1%82%D0%B0%D1%80%D0%BE%D0%BD%D0%BA%D0%B0&oldid=3091963](https://be.wikipedia.org/w/index.php?title=%D0%93%D0%B0%D0%BB%D0%BE%D1%9E%D0%BD%D0%B0%D1%8F_%D1%81%D1%82%D0%B0%D1%80%D0%BE%D0%BD%D0%BA%D0%B0&oldid=3091963)
- Hardeniya, N., Perkins, J., Chopra, D., Joshi, N., & Mathur, I. (2016). *Natural Language Processing: Python and NLTK*. Packt Publishing.
- Hareide, L., & Hofland, K. (2012). Compiling a Norwegian-Spanish parallel corpus: Methods and challenges. In M. P. Oakes & M. Ji (Eds.), *Quantitative Methods in Corpus-Based*

- Translation Studies* (Vol. 51, pp. 75–114). John Benjamins Publishing Company.  
<https://doi.org/10.1075/scl.51.04har>
- Harris, T., & Moreno Jaen, M. (2011). *Corpus Linguistics in Language Teaching*. Peter Lang.
- Hejwowski, K. (2011). Płeć i rodzaj gramatyczny w przekładzie [Sex and Gender in Translation]. In A. Kukułka-Wojtasik (Ed.), *Translatio i literatura* (pp. 175–181). Wydawnictwa Uniwersytetu Warszawskiego. <https://pbn.nauka.gov.pl/sedno-webapp/works/234283>
- Heltai, P. (2005). Explicitation, redundancy, ellipsis and translation. In K. Károly & Á. Fóris (Eds.), *New Trends in Translation Studies* (pp. 45–74). Akadémiai Kiadó.
- Hofland, K. (2007). *Corpora list info page*. <http://korpus.uib.no/icame/corpora/welcome.html>
- Hofland, K., & Johansson, S. (1998). The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In S. Johansson & S. Oksefjell (Eds.), *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies* (pp. 87–100). Rodopi.
- Holmes, D. I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3), 111–117. <https://doi.org/10.1093/lc/13.3.111>
- Holmes, J. (1988). The name and nature of Translation Studies. In *Translated!: Papers on Literary Translation and Translation Studies* (pp. 66–80). Rodopi.
- Horsmann, T., Erbs, N., & Zesch, T. (2015). Fast or Accurate? – A Comparative Evaluation of PoS Tagging Models. *Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology*, 22–30. [https://www.researchgate.net/publication/307941144\\_Fast\\_or\\_Accurate\\_-\\_A\\_Comparative\\_Evaluation\\_of\\_PoS\\_Tagging\\_Models](https://www.researchgate.net/publication/307941144_Fast_or_Accurate_-_A_Comparative_Evaluation_of_PoS_Tagging_Models)
- Huerta, J. M. (2008). Vector based approaches to semantic similarity measures. *Advances in Natural Language Processing and Applications Research in Computing Science*, 33, 163–174.
- Human Rights Center ‘Viasna’. (2015, March 2). *Supreme Court: No evidence of neglect of Belarusian language in courts*. Viasna. <http://spring96.org/en/news/75926>
- Hunston, S., & Francis, G. (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. John Benjamins Publishing Company.
- Institute of Formal and Applied Linguistics, Charles University. (2018). *UDPipe [C++]*. Charles University. <https://github.com/ufal/udpipe>
- ITHAKA. (2018). *JSTOR*. <https://www.jstor.org/?refreqid=search%3A156d348ae12f20e58881b2cfdaf278f>
- Jantunen, J. (2001). Synonymity and lexical simplification in translations: A corpus-based approach. *Across Languages and Cultures*, 2(1), 97–112.

- Jantunen, J. (2004). Untypical patterns in translations: Issues on corpus methodology and synonymity. In A. Maurannen & P. Kujamäki (Eds.), *Translation Universals. Do they exist?* (pp. 101–126).
- Jaworska, T. (2002). *Słownik frazeologiczny angielsko-polski i polsko-angielski*. Wydawnictwa Naukowo-Techniczne.
- Ji, M., & Oakes, M. P. (2012). A Corpus study of early English translations of Cao Xueqin's Hongloumeng. In M. P. Oakes & M. Ji (Eds.), *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research* (pp. 177–208). John Benjamins Publishing Company.
- Johansson, S. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. <http://www.hd.uib.no/icame/lob/lob-dir.htm>
- Jones, C., & Waller, D. (2015). *Corpus Linguistics for Grammar: A Guide for Research*. Routledge.
- Juola, P. (2013). Rowling and “Galbraith”: An authorial analysis. *Language Log*. <https://languagelog.ldc.upenn.edu/nll/?p=5315>
- Jurkiewicz-Rohrbacher, E. (2019). *Polish verbal aspect and its Finnish statistical correlates in the light of a parallel corpus* [University of Helsinki]. <https://helda.helsinki.fi/handle/10138/300771>
- Kakietek, P. (2004). *Polsko-angielski słownik frazeologiczny*. Wyższa Szkoła Lingwistyczna.
- Kamunikat.org. (2018). *Bielaruskaja Internet-Biblijateka [Belarusian Online Library]*. Галоўная Старонка - Беларуская Інтэрнэт-Бібліятэка Kamunikat.Org. <http://kamunikat.org/index.php?>
- Karakanta, A., Dehdari, J., & van Genabith, J. (2018). Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1), 167–189. <https://doi.org/10.1007/s10590-017-9203-5>
- Karapapa, S. (2017, May 18). Research & Private Study. *CopyrightUser*. <https://www.copyrightuser.org/understand/exceptions/research-private-study/>
- Katinskaya, A., & Sharoff, S. (2015). Applying Multi-Dimensional Analysis to a Russian Webcorpus: Searching for Evidence of Genres. *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*, 65–74.
- Katkouski, U. (2005). *Swadesh List: Belarusian—Polish, Russian—Belarusian*. Pravapis.Org. [http://www.pravapis.org/art\\_swadesh\\_russian\\_polish.asp](http://www.pravapis.org/art_swadesh_russian_polish.asp)
- Katkouski, U., & Rrapo, J. (2005). *Introduction to Belarusian Alphabet*. [http://www.pravapis.org/art\\_belarusian\\_alphabet.asp](http://www.pravapis.org/art_belarusian_alphabet.asp)
- Kennedy, G. (1998). *An Introduction to Corpus Linguistics* (1 edition). Routledge.

- Kenning, M.-M. (2010). What are parallel and comparable corpora and how can we use them? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 487–500). Routledge.
- Kenny, D. (1999). *Norms and Creativity: Lexis and Translated Text*. Centre for Translation and Intercultural Studies UMIST.
- Kenny, D. (2001). *Lexis and Creativity in Translation: A Corpus-based Study*. St. Jerome Publishing.
- Kilgarriff, A. (2009). Simple maths for keywords. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of Corpus Linguistics Conference CL2009*. University of Liverpool, UK. <https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf>
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36. <https://doi.org/10.1007/s40607-014-0009-9>
- Klaudy, K. (1996). Concretization and generalization of meaning in translation. In B. Lewandowska-Tomaszyk & M. Thelen (Eds.), *Translation and Meaning. Part 3. Proceedings of the 2nd International Maastricht-Łódź Duo Colloquium on "Translation and Meaning"* (pp. 141–163). Hogeschool Maastricht.
- Klaudy, K. (2009). Explicitation. In M. Baker & G. Saldanha (Eds.), *Routledge Encyclopedia of Translation Studies* (Second, pp. 104–108). Routledge.
- Kłyška, M. (1993). *Слоўнік сінонімаў і блізказначных слоў [Dictionary of synonyms and near-synonyms]* (2nd ed.). Вышэйшая школа [Vyšejšaja škola]. <http://www.slounik.org/sinonimy/>
- Kobyliński, Ł., & Kieraś, W. (2018). Part of Speech Tagging for Polish: State of the Art and Future Perspectives. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 307–319). Springer International Publishing.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Europarl: A Parallel Corpus for Statistical Machine Translation* (pp. 79–86). AAMT.
- Koniuszaniec, G., & Blaszkowska, H. (2003). Language and gender in Polish. *Gender Across Languages*, 3, 259–285.
- Kriesel, D. (2007). *A Brief Introduction to Neural Networks*. [http://www.dkriesel.com/en/science/neural\\_networks](http://www.dkriesel.com/en/science/neural_networks)

- Kruger, H. (2018). That again: A multivariate analysis of the factors conditioning syntactic explicitness in translated English. *Across Languages and Cultures*, 1–33. <https://doi.org/10.1556/084.001>
- Kruger, H., & De Sutter, G. (2018). Alternations in contact and non-contact varieties. *Translation, Cognition & Behavior*, 1(2), 251–290.
- Kryvoi, Y. (2017). *The state of distance education in Belarus: Problems and perspectives*. Ostrogorski Centre. <http://belarusdigest.com/papers/distance-education.pdf>
- Kucera, H., & Francis, W. N. (1964). *Brown Corpus Manual*. <http://www.hit.uib.no/icame/brown/bcm.html>
- Kuebler, S., & Zinsmeister, H. (2015). *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Academic.
- Kul, A. (2015, November 21). *TOP-10 spartoŭcaŭ, jakija papularyzujuc bielaruskuju movy [Top 10 sportsmen who popularize Belarusian language]*. Радыё Свабода. <https://www.svaboda.org/a/top-10-spartoucau-jakija-papuliaryzujuc-bielaruskuju-movu/27382584.html>
- Kwintessential. (2019). *Text Contraction and Expansion in Translation*. Kwintessential. <https://www.kwintessential.co.uk/resources/expansion-retraction>
- Laboratorium językowe. (2014). *Korpus języka młodzieży*. [http://www.laboratoriumjezykowe.uw.edu.pl/?page\\_id=1669](http://www.laboratoriumjezykowe.uw.edu.pl/?page_id=1669)
- Lambert, W. E. (1975). Culture and Language as Factors in Learning and Education. In A. Wolfgang (Ed.), *Education of immigrant students* (pp. 55–83). Ontario Institute for Studies in Education. <https://eric.ed.gov/?id=ED096820>
- Lancaster University. (2014). *University Centre for Computer Corpus Research on Language*. <http://ucrel.lancs.ac.uk/>
- Language Over Internet, LLC. (2019). *List of 390 Most Useful English Phrasal Verbs with Definitions*. LOI English. <https://www.skypeenglishclasses.com/english-phrasal-verbs/>
- Language Technology Group G4.19. (2018). *SuperMatrix*. Wrocław University of Science and Technology. <http://nlp.pwr.wroc.pl/en/tools-and-resources/tools/supermatrix>
- Laviosa, S. (1996). *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*. Centre for Translation and Intercultural Studies UMIST.
- Laviosa, S. (1998). Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta*, 43(4), 557–570. <https://doi.org/10.7202/003425ar>
- Laviosa, S. (2000). TEC: A resource for studying what is ‘in’ and ‘of’ translational English. *Across Languages and Cultures*, 1(2), 159–177.

- Laviosa, S. (2002). *Corpus-based Translation Studies: Theory, Findings, Applications*. Rodopi.
- Lazar, M. (Ed.). (2005). *Feminist Critical Discourse Analysis. Studies in Gender, Power and Ideology*. Palgrave.
- Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech, & T. McEnery (Eds.), *Corpus Annotation* (pp. 1–18). Longman.
- Leech, G., & McEnery, T. (n.d.). *SPAAC Speech-Act Annotation Scheme*. Retrieved 17 January 2018, from <http://ucrel.lancs.ac.uk/SPAAC/SPAAC%20Annotation%20Scheme1.pdf>
- Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Longman.
- Lexical Computing CZ s.r.o. (n.d.). *BeTenTen – Belarusian corpus from the web*. Retrieved 2 February 2019, from <https://www.sketchengine.eu/betenten-belarusian-corpus/>
- Lexical Computing Ltd. (2015, July 8). *Statistics used in Sketch Engine (Chapter 5)*. <https://www.sketchengine.eu/documentation/statistics-used-in-sketch-engine/>
- Li, P., Sun, M., & Xue, P. (2010). Fast-Champollion: A Fast and Robust Sentence Alignment Algorithm. *COLING*.
- Linn, S. I. (2006). Trends in the translation of a minority language: The case of Dutch. In A. Pym, M. Shlesinger, & Z. Jettmarová (Eds.), *Sociocultural Aspects of Translating and Interpreting* (pp. 27–39). John Benjamins Publishers. [https://www.rug.nl/research/portal/publications/trends-in-the-translation-of-a-minority-language\(6120d012-92ae-439b-8b8e-8c7bdf931ef4\)/export.html](https://www.rug.nl/research/portal/publications/trends-in-the-translation-of-a-minority-language(6120d012-92ae-439b-8b8e-8c7bdf931ef4)/export.html)
- Linux-Intelligent-Ocr-Solution* (2.6). (2017). [Python]. <https://sourceforge.net/projects/lios/>
- Lobanov, B., Hetsevich, Y., & Hetsevich, S. (2015). Belarusian and Russian linguistic modules processing for the system NooJ as applied to text-to-speech synthesis. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference 'Dialog 2015'*. Dialog 2015, Moscow. [https://www.researchgate.net/publication/275342127\\_Belarusian\\_and\\_Russian\\_linguistic\\_modules\\_processing\\_for\\_the\\_system\\_NooJ\\_as\\_applied\\_to\\_text-to-speech\\_synthesis](https://www.researchgate.net/publication/275342127_Belarusian_and_Russian_linguistic_modules_processing_for_the_system_NooJ_as_applied_to_text-to-speech_synthesis)
- Looby, R. (2015). *Censorship, Translation and English Language Fiction in People's Poland*. Brill Rodopi. <https://brill.com/view/title/31705>
- Łukašaniec, A. A., & Rusak, V. P. (Eds.). (2012). *Słownik bielaruskaj movy [Belarusian language dictionary]*. Biełaruskaja navuka.
- Machálek, T. (2014). *KonText – application for working with language corpora*. FF UK. <http://kontext.korpus.cz>
- Mahlberg, M. (2013). *Corpus Stylistics and Dickens's Fiction*. Routledge.

- Mahlberg, M., & McIntyre, D. (2011). A Case for Corpus Stylistics: Ian Fleming's 'Casino Royale'. *English Text Construction*, 4(2), 204–227. <https://doi.org/10.1075/target.15.2.04mas>
- Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In A. F. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 171–189). Springer Berlin Heidelberg.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. <https://nlp.stanford.edu/IR-book/>
- Marcińczuk, M., Kaczmarek, A., Kocoń, J., Ptak, M., Szewczyk, M., & Oleksy, M. (2017). *Inforex —Web-based text corpus management system*. <https://inforex.clarin-pl.eu/>
- marketing.by. (2015, February 12). *Mova pačynajecca z ciabie! Novy sacyjalny rolik [The language starts with you! New social video]*. Marketing.By. <http://marketing.by/novosti-rynka/mova-pachynaetsta-z-tsyabe-novy-satsyyalny-rol-k/>
- Mastropierro, L. (2018). Key clusters as indicators of translator style. *Target*, 30(2), 240–259. <https://doi.org/10.1075/target.17040.mas>
- Masubelele, R. (2004). A corpus-based appraisal of shifts in language use and translation policies in two Zulu translations of the Book of Matthew. *Language Matters: Studies in the Languages of Southern Africa*, 35(1), 201–213.
- Maurannen, A. (2000). Strange Strings in Translated Language: A Study on Corpora. In M. Olohan (Ed.), *Intercultural Faultlines. Research Methods in Translation Studies I: Textual and Cognitive Aspects* (pp. 105–118). St. Jerome Publishing.
- Mayo, P. (1978). Byelorussian orthography: From the 1933 reform to the present day. *The Journal of Belarusian Studies*, IV(2), 25–47.
- McEnery, T., & Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
- McEnery, T., & Wilson, A. (2001). *Corpus Linguistics: An Introduction* (2nd Revised edition edition). Edinburgh University Press.
- McEnery, T., Xiao, R., & Tono, Y. (2005). *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge.
- Media Lingo. (2019). *Will the translated version be longer or shorter than the original document?* <http://www.media-lingo.com/gb/faqs/will-the-translated-version-be-longer-or-shorter-than-the-original-document>
- Merkel, M. (2001). Comparing source and target texts in a translation corpus. *Proceedings of the 13th Nordic Conference of Computational Linguistics (NODALIDA 2001)*. <https://aclanthology.info/papers/W01-1716/w01-1716>

- Merriam-Webster Inc. (2019). *Neck of the woods*.  
<https://www.merriam-webster.com/dictionary/neck+of+the+woods>
- Merriam-Webster Inc. (2020). *Set out, verb*.  
<https://www.merriam-webster.com/dictionary/neck+of+the+woods>
- Microsoft. (2016). *Microsoft Word 2016*. <https://products.office.com/en-us/word>
- Miłkowski, M. (2012). *The Polish Language in the Digital Age. Język polski w erze cyfrowej* (G. Rehm & H. Uszkoreit, Eds.). Springer.
- Mniejszość Białoruska – Szkoła Podstawowa nr 395 [Belarusian Minority—Ground school no 395]*. (n.d.). Retrieved 11 January 2018, from <http://pilecki.waw.pl/mniejszosc-bialoruska/>
- Modern Poland Foundation. (2018). *Wolne Lektury Internet library*. Wolne Lektury.  
<https://wolnelectury.pl/>
- Moisl, H. (2015). *Cluster Analysis for Corpus Linguistics*. De Gruyter, Inc.  
<http://ebookcentral.proquest.com/lib/swansea-ebooks/detail.action?docID=1897882>
- Mojeiko, V. (2015, April 21). *Soft Belarusization: A New Shift in Lukashenka's Domestic Policy?* BelarusDigest. <https://belarusdigest.com/story/soft-belarusization-a-new-shift-in-lukashenkas-domestic-policy/>
- Montreal Corpus Tools Revision. (2018). *Montreal Forced Aligner*. Montreal Corpus Tools Revision. <https://montreal-forced-aligner.readthedocs.io/en/latest/index.html>
- Moseley, C. (Ed.). (2010). *Atlas of the World's Languages in Danger* (3rd ed.). UNESCO Publishing. <http://www.unesco.org/culture/en/endangeredlanguages/atlas>
- MovaNanova. (2016, July 25). 'Mova Hurkova'—Epičny rolík ad 'Movy Nanova' z udziałem lehiendarnaha bajca ["Hurkov's Language"—Epic video from 'Langauge Anew' featured by legendary fighter]. *MovaNanova*. <http://www.movananova.by/naviny/mova-gurkova-novy-epichny-rolík-ad-movy-nanova.html>
- MovaNanova. (2017). *Mova Nanova – Belarusian language course*. MovaNanova.  
<http://www.movananova.by/about-mova-nanova-in-english.html>
- Munday, J. (1998). A computer-assisted approach to the analysis of translation shifts. *Meta*, 43(4), 542–556. <https://doi.org/10.7202/003680ar>
- Munday, J., & Zhang, M. (2015). Introduction. *Target. Special Issue: Discourse Analysis in Translation Studies*, 27(3), 325–334. <https://doi.org/10.1075/target.27.3.001int>
- Nacionalnyj korpus ruskogo jazyka. (2017). *Parallelnyj korpus (bieloruskij) [Parallel corpus (Belarusian)]*. <http://ruscorpora.ru/search-para-be.html>
- Nacyjanalnaja biblijateka Biełarusi. (2016). *Zvodny elektronny kataloh biblijatek Biełarusi [Collective digital catalogue of Belarus' libraries]*. <http://unicat.nlb.by/opac/index.html.be>

- Našyja dzietki. (2010). *Našyja dzietki [Our kids]*. <http://dzietki.org/>
- National Corpus of Polish. (2012). *The National Corpus of Polish*. <http://nkjp.pl/index.php?page=0&lang=1>
- National Statistical Committee of the Republic of Belarus. (2009). *Population classified by knowledge of the Belarusian and Russian languages by region and Minsk City*. <http://www.belstat.gov.by/en/perepis-naseleniya/perepis-naseleniya-2009-goda/main-demographic-and-social-characteristics-of-population-of-the-republic-of-belarus/population-classified-by-knowledge-of-the-belarusian-and-russian-languages-by-region-and-minsk-city/>
- Natural Language Processing group Belarus. (2017). <http://nlproc.by/?og=1>
- Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying Stylometry Techniques and Applications. *ACM Computing Surveys*, 50(6), 86:1-86:36. <https://doi.org/10.1145/3132039>
- Nelson, M. (2010). Building a written corpus: What are the basics? In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 53–65). Routledge.
- Nesi, H., & Thompson, P. (2006). *The British Academic Spoken English Corpus Manual*. <http://www.coventry.ac.uk/Global/08%20New%20Research%20Section/Current%20Projects/BASE%20corpus%20manual.pdf>
- nlproc.by. (2018). *Computational linguistics in Belarus*. GitHub. <https://github.com/nlprocby>
- NLTK Project. (2018a). *Natural Language Toolkit—NLTK 3.3 documentation*. <http://www.nltk.org/>
- NLTK Project. (2018b). *NLTK Data*. [http://www.nltk.org/nltk\\_data/](http://www.nltk.org/nltk_data/)
- Oakes, M. P. (2014). *Literary Detective Work on the Computer*. John Benjamins Publishing Company.
- OCLC Online Computer Library Center. (2017). *WorldCat.org: The World’s Largest Library Catalog*. <http://www.worldcat.org/>
- Office for National Statistics. (2017). *Population of the UK by country of birth and nationality: 2016* [Statistical bulletin]. Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/ukpopulationbycountryofbirthandnationality/2016>
- Ogrodniczuk, M. (2013, November 21). *Seminarium—Korpus Barokowy [Seminar—Baroque Corpus]*. <http://wiki.nlp.ipipan.waw.pl/korba/Seminarium>
- Ogrodniczuk, M., Górski, R. L., Łaziński, M., & Pęzik, P. (2019). From the National Corpus of Polish to the Polish Corpus Infrastructure. *Journal of Linguistics/Językovedný Casopis*, 70(2), 315–323. <https://doi.org/10.2478/jazcas-2019-0061>

- Ogrodniczuk, Maciej. (2020). *Language Tools and Resources for Polish*. Computational Linguistics in Poland. <http://clip.ipipan.waw.pl/LRT>
- O’Keeffe, A., & McCarthy, M. (Eds.). (2010). *The Routledge Handbook of Corpus Linguistics*. Routledge.
- Olohan, M. (2002). Leave it out! Using a comparable corpus to investigate aspects of explicitation in translation. *Cadernos de Tradução*, 9, 153–169.
- Olohan, M. (2001). Spelling out the optionals in translation: A corpus study. *Proceedings of the Corpus Linguistics 2001 Conference*, 13. <http://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/papers/olohan.pdf>
- Olohan, M., & Baker, M. (2000). Reporting that in translated English: Evidence for subconscious processes of explicitation. *Across Languages and Cultures*, 1(2), 141–158.
- Øverås, L. (1998). In search of the third code: An investigation of norms in literary translation. *Meta*, 43(4), 557–570. <https://doi.org/10.7202/003775ar>
- Pagrindinis puslapis [Main page]. (2018). In *Wikipedia, the free encyclopedia*. [https://lt.wikipedia.org/w/index.php?title=Pagrindinis\\_puslapis&oldid=5350010](https://lt.wikipedia.org/w/index.php?title=Pagrindinis_puslapis&oldid=5350010)
- Partington, A. (2015). Evaluative prosody. In K. Aijmer & C. Rühlemann (Eds.), *Corpus Pragmatics: A Handbook* (pp. 279–303). Cambridge University Press.
- Patton, J. M., & Can, F. (2012). Determining translation invariant characteristics of James Joyce’s Dubliners. In M. P. Oakes & M. Ji (Eds.), *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research* (pp. 209–230). John Benjamins Publishing Company.
- Perdek, M. (2012). *Lexicographic potential of corpus-derived equivalents. The case of English phrasal verbs and their Polish equivalents*. 376–388. <https://euralex.org/publications/lexicographic-potential-of-corpus-derived-equivalents-the-case-of-english-phrasal-verbs-and-their-polish-equivalents/>
- Peters, P. (1989). *Manual of Information to Accompany the Australian Corpus of English (ACE)*. <http://clu.uni.no/icame/manuals/ACE/INDEX.HTM#cont>
- Pęzik, P. (2016). Exploring phraseological equivalence with Paralela. In E. Gruszczyńska & A. Leńko-Szymańska (Eds.), *Polish-language Parallel Corpora* (Vol. 1, pp. 67–81). University of Warsaw, Faculty of Applied Linguistics, Institute of Applied Linguistics. [http://rownolegle.blog.ils.uw.edu.pl/files/2016/03/04\\_P%C4%99zik.pdf](http://rownolegle.blog.ils.uw.edu.pl/files/2016/03/04_P%C4%99zik.pdf)
- Pili Kíria [Main page]. (2018). In *Wikipedia, the free encyclopedia*. <https://el.wikipedia.org/w/index.php?title=%CE%A0%CF%8D%CE%BB%CE%B7:%CE%9A%CF%8D%CF%81%CE%B9%CE%B1&oldid=6809201>

- Pojarová, V. (2016, December 12). *ARF* (average reduced frequency). <https://wiki.korpus.cz/doku.php/en:pojmy:arf>
- Polish Identity Federation. (2014). *PIONIER.Id members*. Pionier.Net.Pl. <https://aai.pionier.net.pl/en/index.php?page=members>
- Portnov, A. (2011). Post-Soviet Ukraine and Belarus dealing with “The Great Patriotic War”. In N. Hayoz, L. Jesień, & D. Koleva (Eds.), *20 Years after the Collapse of Communism. Expectations, Achievements and Disillusions of 1989* (pp. 369–381). Peter Lang.
- Potter, S. (2020). English language. In *Britannica*. <https://www.britannica.com/topic/English-language/>
- Potts, C. (2011). *The Switchboard Dialog Act Corpus*. Computational Pragmatics. <http://comp prag.christopherpotts.net/swda.html>
- Pracownia Lingwistyki Migowej. (2019). *Korpus PJM [Corpus of Polish Sign Language]*. <https://www.plm.uw.edu.pl/projekty/korpus-pjm/>
- PrajdziSviet—Časopis pierakladnoj litaratury [PrajdziSviet—Journal of translational literature]*. (n.d.). Retrieved 18 October 2017, from <http://prajdzisvet.org/>
- Project Gutenberg*. (2018). Project Gutenberg. <http://www.gutenberg.org/>
- Project Gutenberg Australia*. (2018). <http://gutenberg.net.au/>
- Przepiórkowski, A. (2011). *Tagset*. <http://nkjp.pl/poliqarp/help/ense2.html#x3-40002.2>
- Puzynina, J. (1992). *Język wartości [Language of values]*. Wydawnictwo Naukowe PWN.
- Radziszewski, A., & Śniatowski, T. (2013). *WMBT tagger*. <http://nlp.pwr.wroc.pl/redmine/projects/wmbt/wiki>
- Ramasamy, L., Bojar, O., & Žabokrtský, Z. (2012). Morphological Processing for English-Tamil Statistical Machine Translation. *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, 113–122. <https://www.aclweb.org/anthology/W12-5611>
- Rao, P. (2019). The Role of English as a Global Language. *Research Journal Of English*, 4(1), 65–79.
- Rawlinson, K. (2013, January 30). *Polish is second most spoken language in England, as census reveals*. The Independent. <http://www.independent.co.uk/news/uk/home-news/polish-is-second-most-spoken-language-in-england-as-census-reveals-140000-residents-cannot-speak-8472447.html>
- Reppen, R., O’Keeffe, A., & McCarthy, M. (2010). Building a corpus. In *The Routledge handbook of corpus linguistics* (pp. 31–37). Routledge Handbooks Online. <https://doi.org/10.4324/9780203856949.ch3>

- Romero-Trillo, J. (Ed.). (2008). *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. De Gruyter.
- Rühlemann, C., & O'Donnell, M. B. (2015). Deixis. In K. Aijmer & C. Rühlemann (Eds.), *Corpus Pragmatics: A Handbook* (pp. 331–359). Cambridge University Press.
- Rusak, V. P. (Ed.). (2013a). *Hramatyčny sloŭnik dziejasłova [Grammatical dictionary of the verb]*. Biełaruskaja navuka.
- Rusak, V. P. (Ed.). (2013b). *Hramatyčny sloŭnik nazoŭnika [Grammatical dictionary of the noun]*. Biełaruskaja navuka.
- Rusak, V. P. (Ed.). (2013c). *Hramatyčny sloŭnik prymietnika, zajmiennika, ličebnika, prysłouŭja [Grammatical dictionary of the adjective, pronoun, numeral, adverb]*. Biełaruskaja navuka.
- Rybicki, J. (2012). The great mystery of the (almost) invisible translator. In M. P. Oakes & M. Ji (Eds.), *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research* (pp. 231–248). John Benjamins Publishing Company.
- Rybicki, J., & Eder, M. (2011). Deeper Delta across genres and languages: Do we really need the most frequent words? *Literary and Linguistic Computing*, 26(3), 315–321. <https://doi.org/10.1093/llc/fqr031>
- Saldanha, G. (2011a). Style of translation: An exploration of stylistic patterns in the translations of Margaret Jull Costa and Peter Bush. In A. Kruger, K. Wallmach, & J. Munday (Eds.), *Corpus-Based Translation Studies. Research and Applications* (pp. 237–258). Continuum.
- Saldanha, G. (2011b). Translator Style. Methodological Considerations. *The Translator*, 17(1), 25–50. <https://doi.org/10.1080/13556509.2011.10799478>
- Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer.
- Scarpa, F. (2006). Corpus-based Quality-Assessment of Specialist Translation: A Study Using Parallel and Comparable Corpora in English and Italian. In S. Šarcevic & M. Gotti (Eds.), *Insights into Specialized Translation* (pp. 154–172). Peter Lang.
- Scott, M. (2010). *WordSmith Tools Help*. WordSmith. [https://lexically.net/downloads/version5/HTML/index.html?type\\_token\\_ratio\\_proc.htm](https://lexically.net/downloads/version5/HTML/index.html?type_token_ratio_proc.htm)
- Scott, M. (2017). *WordSmith Tools* (Version 7) [Computer software]. Lexical Analysis Software. <http://lexically.net/wordsmith/>
- Scott, M. N. (1998). *Normalization ad Readers' Expectation: A Study of Literary Translation with Reference to Lispecotr's 'A Hora da Estrela'*. University of Liverpool.
- SDL Trados. (2018). *What is Translation Alignment?* SDL. <https://www.sdltrados.com/solutions/translation-alignment/>

- Shastri, S. V. (1986). *Manual of Information to Accompany the Kolhapur Corpus of Indian English, for Use with Digital Computers*. <http://clu.uni.no/icame/kolhapur/kolman.htm>
- Shearlaw, M. (2015, February 3). Belarus Bookshop Rallies Against Publishing Crackdown. *The Guardian*. <http://www.theguardian.com/world/2015/feb/03/belarus-bookshop-lohvinau-publishing-crackdown>
- Sigley, R., & Holmes, J. (2002). Girl-watching in corpora of English. *Journal of English Linguistics*, 30(2), 138–157.
- Silberztein, M. (2003). *NooJ Manual*. <http://www.nooj-association.org/media/k2/attachments/app/NooJManual.pdf>
- Silberztein, M. (2018). *NooJ: A Linguistic Development Environment* (5.0) [Computer software]. [http://nooj-association.org/index.php?option=com\\_k2&view=item&layout=item&id=13&Itemid=619](http://nooj-association.org/index.php?option=com_k2&view=item&layout=item&id=13&Itemid=619)
- Šilić, A., Chauchat, J.-H., Dalbelo Bašić, B., & Morin, A. (2007). N-Grams and Morphological Normalization in Text Classification: A Comparison on a Croatian-English Parallel Corpus. In J. Neves, M. F. Santos, & J. M. Machado (Eds.), *Progress in Artificial Intelligence* (pp. 671–682). Springer. [https://doi.org/10.1007/978-3-540-77002-2\\_56](https://doi.org/10.1007/978-3-540-77002-2_56)
- Simaki, V., Aravantinou, C., Mporas, I., Kondyli, M., & Megalooikonomou, V. (2017). Sociolinguistic Features for Author Gender Identification: From Qualitative Evidence to Quantitative Analysis. *Journal of Quantitative Linguistics*, 24(1), 65–84. <https://doi.org/10.1080/09296174.2016.1226430>
- Simons, G. F., & Fenning, C. D. (Eds.). (2017). *Ethnologue: Languages of Africa and Europe, Twentieth Edition* (20th ed.). SIL International.
- Singh, A. K., & Husain, S. (2005). Comparison, Selection and Use of Sentence Alignment Algorithms for New Language Pairs. *ParallelText@ACL*. <http://www.aclweb.org/anthology/W05-0816>
- Skolimowska, A., & Turska, M. (n.d.). *Corpus of Ioannes Dantiscus' Texts & Correspondence*. Retrieved 18 January 2018, from <http://dantiscus.ibi.uw.edu.pl/?&lang=pl&lang=eng>
- Smolicz, J. J., & Radzik, R. (2004). Belarusian as an endangered language: Can the mother tongue of an independent state be made to die? *International Journal of Educational Development*, 24(5), 511–528. [https://doi.org/10.1016/S0738-0593\(03\)00072-5](https://doi.org/10.1016/S0738-0593(03)00072-5)
- Soria, C., Russo, I., Quochi, V., Hicks, D., Gurrutxaga, A., Sarhimaa, A., & Tuomisto, M. (2016). Fostering digital representation of EU regional and minority languages: The Digital Language Diversity Project. *Proceedings of the Tenth International Conference on Language*

- Resources and Evaluation (LREC'16)*, 3256–3260. <https://www.aclweb.org/anthology/L16-1518>
- Spasiuk, A. (2015, May 12). *Refierendum-1995. Bielaruskuju movu zahnali ŭ adukacyjnaje padpolle [Referendum 1995. Belarusian language pushed into educational underground]*. Naviny.By. [http://naviny.by/rubrics/society/2015/05/12/ic\\_articles\\_116\\_188862](http://naviny.by/rubrics/society/2015/05/12/ic_articles_116_188862)
- Spasiuk, A. (2017, August 25). *Underground Belarusian*. OpenDemocracy. <https://www.opendemocracy.net/od-russia/elena-spas/underground-belarusian>
- Speech Synthesis and Recognition Laboratory. (2017). *Nooj tutorials: Па-беларуску [Nooj tutorials: In Belarusian]*. [https://www.youtube.com/playlist?list=PLtc\\_R9i0zr6QjyLk5\\_Vn\\_9F4balHK42XW](https://www.youtube.com/playlist?list=PLtc_R9i0zr6QjyLk5_Vn_9F4balHK42XW)
- Speech synthesis and recognition laboratory. (2018). *Lemmatizer*. <https://corpus.by/Lemmatizer/?lang=en>
- Speech Synthesis and Recognition Laboratory. (2018a). *Part-of-Speech Tagger*. <https://corpus.by/PartOfSpeechTagger/?lang=en>
- Speech Synthesis and Recognition Laboratory. (2018b). *Speech Synthesis and Recognition Laboratory*. <http://ssrlab.by/en/>
- Speech Synthesis and Recognition Laboratory. (2018c). *Voiced Electronic Grammatical Dictionary*. <http://www.corpus.by/VoicedElectronicGrammaticalDictionary/?lang=en>
- Speech Synthesis and Recognition Laboratory. (2019). *Computational platform for electronic text & speech processing*. Corpus.By. <http://corpus.by/index.php?lang=en>
- Statistics Poland. (2015). *Struktura narodowo-etniczna, językowa i wyznaniowa ludności Polski—Narodowy Spis Powszechny 2011 [National, ethnical, linguistics and religious structure of the Polish population—National Census 2011]*. Statistics Poland. <http://stat.gov.pl/spisy-powszechne/nsp-2011/nsp-2011-wyniki/struktura-narodowo-etniczna-jezykowa-i-wyznaniowa-ludnosci-polski-nsp-2011,22,1.html>
- Statistics Poland. (2019). *Szkoły wyższe i ich finanse w 2018 r. Higher education institutions and their finances in 2018*. Statistical Office in Gdańsk. <https://stat.gov.pl/obszary-tematyczne/edukacja/edukacja/szkoly-wyzsze-i-ich-finance-w-2018-roku,2,15.html>
- Stokoe, E. (1998). Talking about Gender: The Conversational Construction of Gender Categories in Academic Discourse. *Discourse & Society*, 9(2).
- Straka, M. (2018). *UDPipe (1.2)* [Computer software]. <http://lindat.mff.cuni.cz/services/udpipe/>
- Supernova-soft. (2014). *NOVA Text Aligner (1.1)* [Computer software]. Supernova-soft. <http://www.supernova-soft.com/wpsite/products/text-aligner/>

- Sztencel, M. (2009). Boundaries crossed: The influence of English on modern Polish. *E-Pisteme*, 2(1), 3–17.
- Szymańska, I. (2008). Translation as Rewriting. The Problem of Grammatical Gender in the Polish Translation of Jeffrey Archer's 'You'll never live to regret it'. In A. Korzeniowska & M. Grzegorzewska (Eds.), *DUET encounters* (pp. 127–158). Institute of English Studies, University of Warsaw. <https://pbn.nauka.gov.pl/sedno-webapp/works/22101>
- TakeLab. (2014). *CORAL*. TakeLab. [http://takelab.fer.hr/coral\\_s/](http://takelab.fer.hr/coral_s/)
- Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn Treebank: An Overview. In A. Abeillé (Ed.), *Treebanks: Building and Using Parsed Corpora* (pp. 5–22). Springer Netherlands. [https://doi.org/10.1007/978-94-010-0201-1\\_1](https://doi.org/10.1007/978-94-010-0201-1_1)
- Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. *Corpora*, 8(1), 81–113. <https://doi.org/10.3366/cor.2013.0035>
- Taylor, L. J., & Knowles, G. (1988). *Manual of information to accompany the SEC corpus: The Machine-Readable Corpus of Spoken English*. <http://clu.uni.no/icame/manuals/SEC/INDEX.HTM>
- TEI Consortium. (2020). *Text Encoding Initiative*. TEI: Text Encoding Initiative. <http://www.tei-c.org/>
- Teich, E. (2003). *Cross-Linguistic Variation in System and Text, A Methodology for the Investigation of Translations and Comparable Texts*. De Gruyter Mouton. <https://doi.org/10.1515/9783110896541>
- Text Encoding Initiative. (2018). *Guidelines*. <http://www.tei-c.org/guidelines/>
- The Apache Software Foundation. (2018). *Apache Tika* (1.18) [Computer software]. The Apache Software Foundation. <http://tika.apache.org/>
- The Document Foundation, Debian and Ubuntu. (2017). *Writer* (5.1.6.2) [Computer software]. <https://www.libreoffice.org/discover/writer/>
- The European Parliament and the Council of the European Union. (1996). *DIRECTIVE 96/9/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 March 1996 on the legal protection of database*. <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML>
- The GIMP Development Team. (2017). *GIMP* (2.8) [Computer software]. <https://www.gimp.org/>
- The ICE Project. (2016). *International Corpus of English (ICE)*. <http://ice-corpora.net/ice/>
- The Institute of the Polish Language at the Polish Academy of Sciences (IJP PAN). (n.d.). *KORBA*. Retrieved 17 April 2018, from [http://korba.nlp.ipipan.waw.pl/query\\_corpus/](http://korba.nlp.ipipan.waw.pl/query_corpus/)

- The Internet Archive. (2018). *The Internet Archive*. The Internet Archive: Digital Library. <https://archive.org/>
- The London Magazine*. (2018). <https://www.thelondonmagazine.org/>
- The New Yorker*. (2018). <https://www.newyorker.com/>
- The Polish Language Council. (2019). *Stanowisko Rady Języka Polskiego przy Prezydium PAN w sprawie żeńskich form nazw zawodów i tytułów [Statement concerning the female forms of professions and titles, by the Polish Language Council by the presidium of the Polish Academy of Science]*. [http://www.rjp.pan.pl/index.php?option=com\\_content&view=article&id=1861:stanowisko-rjp-w-sprawie-zenskich-form-nazw-zawodow-i-tytulow&catid=98&Itemid=58](http://www.rjp.pan.pl/index.php?option=com_content&view=article&id=1861:stanowisko-rjp-w-sprawie-zenskich-form-nazw-zawodow-i-tytulow&catid=98&Itemid=58)
- The University of Adelaide. (2018). *EBooks@Adelaide* [Text]. <http://ebooks.adelaide.edu.au/index.html>
- The University of Manchester. (n.d.). *ARCHER: A Representative Corpus of Historical English Registers*. ARCHER: A Representative Corpus of Historical English Registers. Retrieved 17 January 2018, from <http://www.projects.alc.manchester.ac.uk/archer/>
- The University of Michigan. (2007). *MICASE. The Michigan Corpus of Academic Spoken English*. MICASE. <https://quod.lib.umich.edu/m/micase/>
- Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., Lu, Y., Li, S., Wang, Y., & Wang, L. (2014). UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1837–1842. [http://www.lrec-conf.org/proceedings/lrec2014/pdf/774\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/774_Paper.pdf)
- Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*. 8th International Conference on Language Resources and Evaluation, Istanbul. [http://www.lrec-conf.org/proceedings/lrec2012/pdf/463\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf)
- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins Publishing Company.
- Toury, G. (1995). *Descriptive Translation Studies – and beyond*. John Benjamins Publishing Company.
- Toury, G. (2012). *Descriptive Translation Studies – and beyond*. John Benjamins Publishing Company.
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of HLT-NAACL 2003, 1*, 252–259. *Translated by humans*. (n.d.). Retrieved 2 November 2018, from <http://translatedby.com/>

- Tribble, C. (2010). What are concordances and how are they used? In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 167–183). Routledge.  
<https://doi.org/10.4324/9780203856949.ch13>
- Tsuchimura, N. (2016). Stylistic Analysis of Agatha Christie’s Works: Comparing with Dorothy Sayers. *Proceedings of the 6th Conference of Japanese Association for Digital Humanities*, 66–67.
- Tymoczko, M. (1998). Computerized corpora and the future of Translation Studies. *Meta*, 43(4), 652–660. <https://doi.org/10.7202/004515ar>
- Universal Dependencies contributors. (2018a). *CoNLL-U Format*.  
<http://universaldependencies.org/format.html>
- Universal Dependencies contributors. (2018b). *Universal Dependencies*. <http://www.anc.org/>
- Universal Dependencies contributors. (2018c). *Universal POS tags*.  
<http://universaldependencies.org/u/pos/index.html>
- University of Birmingham. (2017). *Resources*.  
<https://www.birmingham.ac.uk/research/activity/corpus/resources.aspx>
- University of Birmingham, & University of Nottingham. (2017). *CLiC*. <http://clic.bham.ac.uk/>
- University of Gothenburg. (2016). *ASPAC – Swedish-Belarusian corpus*. Språkbanken (the Swedish Language Bank). <https://spraakbanken.gu.se/eng/resource/aspacsvbe>
- University of Gothenburg. (2018). *Resources*. Språkbanken.  
<https://spraakbanken.gu.se/eng/resources>
- University of Oxford. (2015). *British National Corpus* [Text]. <http://www.natcorp.ox.ac.uk/>
- University of Warsaw. (2020). *Corpus Research Centre*. <https://www.ils.uw.edu.pl/institut/zaklady-i-pracownie/pracownia-badan-korpusowych/>
- University of Wolverhampton. (2018). *MA Practical Corpus Linguistics for ELT, Lexicography, and Translation*. <http://courses.wlv.ac.uk/course.asp?code=WL049P31UVD>
- Uniwersytet Łódzki. (2013). *PELCRA Learner English Corpus*. <http://pelcra.pl/plec/>
- Vasilevich, H. (2013). Managing language diversity in Belarus. In K. Selnes & T. Senyushkina (Eds.), *Identity and Collective Memory* (pp. 242–283). Norway humanist association.
- Vavřín, M. (2015). *Treq – database of translation equivalents*. FF UK. <https://treq.korpus.cz/>
- Viačorka, V. (2017, February 21). 1 da 205. Prava biełaruskamoŭnych na adukacyju u ličbach. Da Mižnarodnaha dnia matčynaj movy [1 in 205. Belarusian Language Speaker’s Rights to Education in Numbers. By the Occasion of the International Mother Language Day]. *Радыё Свабода*. <https://www.svaboda.org/a/dzien-matcynaj-movy/28322310.html>

- Viberg, Å. (2017). *Contrasts in morphology: The case of UP/DOWN and IN/OUT as bound morphemes in Swedish and their English correspondences* (pp. 32–74). Cambridge Scholars Publishing. <http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-337436>
- von Waldenfels, R. (2011). *ParaSol—A Parallel Corpus of Slavic and Other Languages*. <http://parasolcorpus.org/>
- W3Techs. (2018). *Usage Survey of Character Encodings broken down by Ranking*. [https://w3techs.com/technologies/cross/character\\_encoding/ranking](https://w3techs.com/technologies/cross/character_encoding/ranking)
- Walkowiak, T., & Kaliński, M. (2013). *WCRFT2 Tagger*. <http://ws.clarin-pl.eu/tager.shtml>
- Walkowiak, T., & Piasecki, M. (2018). *WebSty. System analizy podobieństwa tekstów [WebSty. System for the analysis of texts similarity]*. <http://ws.clarin-pl.eu/websty.shtml>
- Waszczuk, J. (2018). *Concraft-pl* (0.7) [Computer software]. <http://zil.ipipan.waw.pl/Concraft>
- Wawer, A. (2018). *Language Tools and Resources for Polish*. <http://clip.ipipan.waw.pl/LRT>
- Weisser, M. (2016). *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis*. Wiley-Blackwell.
- Wijffels, J. (2020). *Package 'udpipe'*. <https://cran.r-project.org/web/packages/udpipe/udpipe.pdf>
- Wikipedia. (n.d.). Retrieved 29 March 2018, from <https://www.wikipedia.org/>
- Wikisource, the free library. (2018). [https://en.wikisource.org/wiki/Main\\_Page](https://en.wikisource.org/wiki/Main_Page)
- Wilson, A. (2016). Belarus: From a social contract to a security contract? *The Journal of Belarusian Studies*, 8(1), 78–91.
- Winters, M. (2009). Modal Particles Explained: How Modal Particles Creep into Translations and Reveal Translators' Styles. *Target*, 21(1), 74–97. <https://doi.org/10.1075/target.21.1.04win>
- Woliński, M., & Lenart, M. (2018). *Morfeusz Polimorf* (Version 2) [Computer software]. <http://sgjp.pl/morfeusz/index.html>
- Woliński, M., Miłkowski, M., Ogrodniczuk, M., Przepiórkowski, A., & Szalkiewicz, Ł. (2012). PoliMorf: A (not so) new open morphological dictionary for Polish. *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, 860–864. [http://www.lrec-conf.org/proceedings/lrec2012/pdf/263\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/263_Paper.pdf)
- Wołk, K. (2015). Polish to English Statistical Machine Translation. *ArXiv:1510.00001 [Cs, Stat]*. <http://arxiv.org/abs/1510.00001>
- Wrocław University of Technology. (2016). *SłowoSieć*. <http://plwordnet.pwr.wroc.pl/wordnet/>
- Wu, D., & Xia, X. (1994). Learning an English-Chinese Lexicon from a Parallel Corpus. *AMTA*, 206–213. <https://www.aclweb.org/anthology/1994.amta-1.26.pdf>
- Wydawnictwo Naukowe PWN. (2020). Ogarnąć. In *SJP PWN*. PWN.
- Wyszukiwarka korpusowa Monco [Corpus browser Monco]. (2016). <http://monco.frazeo.pl/index>

- Xiao, R. (2010). How different is translated Chinese from native Chinese?: A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, 15(1), 5–35.  
<https://doi.org/10.1075/ijcl.15.1.01xia>
- Xiaoyi, M. (2006). Champollion: A robust parallel text sentence aligner. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation* (pp. 489–492). European Language Resources Association (ELRA).  
[http://www.lrec-conf.org/proceedings/lrec2006/pdf/746\\_pdf.pdf](http://www.lrec-conf.org/proceedings/lrec2006/pdf/746_pdf.pdf)
- Yekelchik, S. (2008). Out of Russia's long shadow: The making of modern Ukraine, Belarus, and Moldova. In *Europe's Last Frontier?* (pp. 9–29). Palgrave Macmillan.  
[https://doi.org/10.1007/978-1-137-10170-9\\_2](https://doi.org/10.1007/978-1-137-10170-9_2)
- Zabawa, M. (2008). English-Polish language contact and its influence on the semantics of Polish. *Studia Germanica Gedanensia*, 17, 154–164.
- Zalizniak, A. A. (1980). *Грамматический словарь русского языка: Словоизменение* [*Grammatical dictionary of the Russian language: Inflection*]. Russkij jazyk.
- Zanettin, F. (2013). Corpus Methods for Descriptive Translation Studies. *Procedia - Social and Behavioral Sciences*, 95, 20–32. <https://doi.org/10.1016/j.sbspro.2013.10.618>
- Zaprudski, S. (2007). In the grip of replacive bilingualism: The Belarusian language in contact with Russian. *International Journal of the Sociology of Language*, 183, 97–118.
- Zeldes, A. (2012, May 8). Machine Translation between Language Stages: Extracting Historical Grammar from a Parallel Diachronic Corpus of Polish. *Proceedings of the Corpus Linguistics Conference*. CL2007, Birmingham.  
[http://ucrel.lancs.ac.uk/publications/CL2007/paper/60\\_Paper.pdf](http://ucrel.lancs.ac.uk/publications/CL2007/paper/60_Paper.pdf)
- Zinn, C. (2018). *D2.5 LR Switchboard (software)*. <https://switchboard.clarin.eu/#/>
- Zubillaga, N., Sanz, Z., & Uribarri, I. (2015). Building a trilingual parallel corpus to analyse literary translations from German into Basque. In C. Fantinuoli & F. Zanettin (Eds.), *New directions in corpus-based translation studies* (pp. 71–92). Language Science Press.  
<https://doi.org/10.17169/langsci.b76.64>

## Appendix 1: Testing accuracy of Belarusian lemmatisers and POS taggers

### 1. Overview of tested tools

#### (a) SSRlab

The first tool discussed here was developed by Speech Synthesis and Recognition Laboratory in Minsk, Belarusian Academy of Science. It is a web-based application with menu available in three languages (Belarusian, Russian, English).

The interface of the Lemmatizer (Speech synthesis and recognition laboratory, 2018) consists of the window for pasting the text for analysis (there is no possibility of uploading a file), and a window for specifying the token (in this case – lemma) and category delimiter. There is a handful of lexical resources to choose from (Biryła, 1987; Rusak, 2013b; Zalizniak, 1980), some of which are marked as ‘in the process of development’ (Rusak, 2013a, 2013c).

The interface of the Part-of-Speech Tagger (Speech Synthesis and Recognition Laboratory, 2018a) looks exactly the same, apart from dictionaries which are more numerous. In addition to the resources available in the Lemmatizer tool, they include a newer version of the Belarusian dictionary (Łukašaniec & Rusak, 2012), two lists of most common words from the online speech synthesizer (from 2016), and a list of problematic words from Russian (also from 2016). The advantage of using the new lexical resources is adding a small representation of classical variant of Belarusian orthography, as it increases the chances of processing texts from all sources.

The output of the Lemmatizer and POS Tagger analysis is displayed in a new window below the blue ‘Show word list’ button. On the right side of the new window there is a second window containing unknown words – these are also marked in the main output with the tag ‘НевядомаеСлова’. The programme returns the original word with the accent marked by the symbol ‘+’ and its lemma (without marked accent) after chosen delimiter, e.g.

дзялі+ў\_дзяліць [dziali+ŭ\_dzalić] → the original verb дзяліў with the accent on the vowel i and its lemma дзяліць.

The POS Tagger uses language specific tags based on the Penn Treebank Tagset. The meaning of a particular tag can be checked via Voiced Electronic Grammatical Dictionary (Speech Synthesis and Recognition Laboratory, 2018c). Ambiguity is dealt with via assigning multiple tags in cases where more than one is deemed correct. The output of the Lemmatizer and POS Tagger analysis cannot be downloaded, instead it has to be copied and then pasted into preferred file type. The SSRlab platform does not offer a service combining the two tools, the text must be lemmatised and POS tagged separately.

#### (b) NooJ

Next tool examined here was developed by Max Silberztein (2018) and is designed for analysing multiple languages, including Belarusian. Belarusian model has been developed by Belarusian Academy of Science researchers since 2012 (Lobanov et al., 2015). It is a stand-alone application available for Windows, Mac and Linux, however non-Windows versions are written in Java and

there are slight differences in using them – these should be discussed later on in this section. Full description of the interface is available in the manual (Silberztein, 2003), and the specifics of the work with Belarusian module is fully explained in a series of YouTube tutorials (Speech Synthesis and Recognition Laboratory, 2017).

Recognition of the word forms in NooJ is based on a dictionary that is compiled out of a printed edition of Biryła dictionary (1987) and “new words from modern texts on literature, science and technology” (Lobanov et al. 2015: 3); no specification as to where these words come from, how many of them are there or what part of speech they are is given. Overall the Belarusian dictionary has over 2.1 million forms and over 137 thousand lemmas, and uses 914 language-specific tags.

When it comes to using NooJ software on non-Windows operating system the problem lies the fact that the Java version requires above-mentioned dictionaries in a different format. At the NooJ association webpage in the section ‘Linguistic resources’ there is a set of files for each language available for download, among else the dictionary in .nod format. As it was revealed during the tests of the NooJ programme, this format is not compatible with the Java version and the developers of the Belarusian module informed authors of this paper that a dictionary in .jnod is necessary to work with the programme, however none such dictionary exists as it was never needed.

The output of the lemmatiser and the POS tagger (that is ‘Lexical analysis’ in NooJ) can be exported to .xml format; it should be noted though that the right export command in the case of Belarusian is <DIC>, despite the instructions in the manual and video tutorial. The annotated data takes form of a string of tags starting with lemma followed by general grammatical category (similar to universal tags discussed in the next section). In ambiguous cases there are more than one category assigned for a single word. Words unrecognised by the software are simply left unannotated.

### (c) UDPipe

The third tool tested here was developed by the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic (Institute of Formal and Applied Linguistics, Charles University, 2018). It is a programme with the broadest array of applications, as it serves all languages which have treebanks available within the Universal Dependencies (UD) project. UD is “a framework for cross-linguistically consistent grammatical annotation and an open community effort with over 200 contributors producing more than 100 treebanks in over 70 languages” (Universal Dependencies contributors, 2018b).

The interface of the UDPipe consists of a language selection bar and the widow for input text. If multiple versions of treebank are available the user may choose from among them. The input text can be loaded from a file of plain text format. The tool performs lemmatising and POS tagging (option ‘Tag and Lemmatize’), it can additionally perform syntactic analysis (‘Parse’). The output is available in three forms: plain text with lemmas, tags and additional information separated by tabs or a table, additionally in case of parsed text programme returns graphic files with the trees. All these types of output can be downloaded. The textual data is stored in the CoNLL-U format (Universal Dependencies contributors, 2018a).

UDPipe uses two types of tagsets. Firstly, the universal tags (Universal Dependencies contributors 2018c) that can be used for all languages represented in the UD project, from Indo-European languages, through Afro-Asiatic, Sino-Tibetan, Turkic, Uralic languages, or even Basque and Swedish sign language. Secondly, there are language-specific POS tags; as in the case of the SSRlab tagger they are based on the Penn Treebank, however they contain less morphological detail than in the case of SSRlab tool.

Lemmatiser and POS tagger is trained on particular treebanks. In the case of Belarusian two versions have been tested: 2.3 with treebank of 8 thousand words from online news, and 2.4 with treebank of 13 thousand words supplemented with non-fiction, legal and fiction texts (among else, a 500-word sample of translational language). The newer version contains also data in both orthographic variants of Belarusian. It must be noted here that the UD treebanks are updated every 6 months and by the end of the 2020 the Belarusian treebank has grown significantly – from 13 thousand words in version 2.4 up to 275 thousand in the 2.7 version.

## 2. Experimental setup

### (a) Data

Translational Belarusian literature varies in terms of size (short stories, novellas and novels) and genre (e.g. fantasy, war novel, family saga). Even though the EPB corpus contains only the literature from the 20<sup>th</sup> and 21<sup>st</sup> centuries, the language, both of the narrator and the characters, differs significantly between the titles and is reflected in the translations. Therefore two types of samples were used in the test:

- i. single sentences from 68 texts of the corpus: 546 words
- ii. three 10-sentence block from three texts of various genre: 456 words

All together it constitutes 1002-word sample. This sample is relatively small, however the general feeling was that it would suffice for estimating the POS tagger accuracy while still remaining feasible in terms of limited time and workforce, as the tags were checked manually. The sample was run through each lemmatizer and POS tagger. Next, the results were saved and analysed in the spreadsheet.

### (b) Evaluation criteria

Following rules of evaluation have been applied:

- i. in the case of adjectives a singular male form was considered the right lemma (e.g. *цёмны* [ciomny] for *цёмная* [ciomnaja])
- ii. verbs were regarded as right lemmas for participial adjectives (e.g. *утварыць* [utvaryć] for *утвораны* [utvorany]) and participial adverbs (e.g. *пачакаць* [pačakać] for *пачакаўшы* [pačakaišy])
- iii. in the case of all lemmatizers punctuation marks were not taken into account

- iv. in the case of NooJ lemmatizer and POS tagger only the first interpretation of a word was taken into account
- v. in the case of the UDPipe POS tagger only the universal POS tags were evaluated
- vi. in the case of SSR lab POS tagger only the first part of the tag (the one that denominated the general part of speech) was evaluated
- vii. in the case of all POS taggers recognition of punctuation marks was taken into account and evaluated
- viii. words were counted accordingly to the number of tokens indicated in the outcome of the lemmatizer (e.g. *што-небудзь* [što-niebudž] is counted as two words in Pluma text editor but in fact it is just one word – *something*)

### 3. Results and discussion

The resulting accuracy percentages can be seen in the Table 1.

Name of the programme	Lemmatisation accuracy	POS tagging accuracy	Unknown lemmas	Unknown POS tags
UDPipe v2.3	84.33%	78.74%	na	na
UDPipe v2.4	86.93%	81.80%	na	na
SSR Lab	86.91%	87.62%	8.69%	5.91%
NooJ	73.80%	81.63%	6.12%	6.12%

These results indicate several phenomena. Firstly, the accuracy of the UDPipe v.2.4 in lemmatising and POS tagging is higher than v2.3 by 2.6% and 3.06% respectively. Adding just a small sample (less than 4% of the treebank size) of translational language to the training data significantly improves the performance of the programme. Secondly, it is the newest version of UDPipe and SSR Lab tools that are the most accurate in lemmatising and POS tagging Belarusian language, however their performance gives a lot of room for improvements, when compared to other languages. Thirdly, UDPipe proves to be very effective, despite operating on more limited lexical resources than the SSRlab tools or NooJ software.

## Appendix 2: Testing relations between the contraction factor of the translations and the external factors included in the metadata

	pl_FC	be_FC	Fc_diff
eng_date	-0,08 0,407 107	0 0,967 107	0,06 0,51 107
pol_date	-0,07 0,498 106	0,09 0,383 106	0,15 0,123 106
bel_date	0,09 0,36 106	0,06 0,533 106	-0,01 0,928 106
date_diff_pl	-0,05 0,639 107	0,07 0,455 107	0,12 0,223 107
date_diff_be	0,13 0,175 107	0,02 0,864 107	-0,09 0,337 107
eng_gender	-0,04 0,689 107	-0,13 0,186 107	-0,11 0,271 107
pol_gender	-0,09 0,332 107	-0,11 0,253 107	-0,04 0,672 107
bel_gender	-0,07 0,487 98	0,09 0,398 98	0,15 0,135 98
gender_diff_pl	-0,05 0,581 107	0 0,995 107	0,05 0,644 107
gender_diff_be	0,02 0,82 98	0,19 0,068 98	0,19 0,067 98
eng_size	0 0,968 107	0,04 0,711 107	0,04 0,661 107
eng_DOB	-0,04 0,666 106	0,07 0,49 106	0,11 0,26 106
pol_DOB	-0,09 0,47 74	0 0,974 74	0,07 0,552 74
bel_DOB	0,07 0,605 53	-0,19 0,167 53	-0,21 0,122 53
DOB_diff_pl	0,04 0,737 75	0,06 0,591 75	0,04 0,752 75
DOB_diff_be	0,2 0,154	-0,27 0,049	-0,37 0,007

### Key to abbreviations:

eng\_date: publication date of the originals

pol\_date: publication date of the Polish translational

bel\_date: publication date of the Belarusian translational

date\_diff\_pl: difference between publication dates of the original and the Polish translation

date\_diff\_be: difference between publication dates of the original and the Belarusian translation

eng\_gender: gender of the author

pol\_gender: gender of the Polish translator

bel\_gender: gender of the Belarusian translator

gender\_diff\_pl: difference between the gender of the author and the Polish translator

gender\_diff\_be: difference between the gender of the author and the Belarusian translator

eng\_size: size of the original

eng\_DOB: date of birth of the author

pol\_DOB: date of birth of the Polish translator

bel\_DOB: date of birth of the Belarusian translator

DOB\_diff\_pl: difference between the dates of birth of the author and the Polish translator

DOB\_diff\_be: difference between the dates of birth of the author and the Belarusian translator

pl\_FC: contraction factor of Polish translation

be\_FC: contraction factor of Belarusian translation

Fc\_diff: difference between contraction factors of Polish and Belarusian translations

### Appendix 3: Close reading of source text samples and their translations

Bradbury, Ray, *The Dragon*, 1955

Bradbury, Ray, *Cmok*, 1996, translated by Alaksandr Kudraŭcaŭ

Bradbury, Ray, *Smok*, 2009, translated by Anna Gren

Note: contraction factors are calculated for this particular sample.

Polish (108w, F <sub>c</sub> =72%)	English (149w)	Belarusian (118w, F <sub>c</sub> =79%)
<i>Przez ponure pustkowia, w którym były jedynie ciemność i nicość płynące z serca wrzosowisk, przetoczył się podmuch wiatru, niosąc pył z zegarów, które powoli zmieniały się w proch.</i>	<i>Across the dim country, full of night and nothingness from the heart of the moor itself, the wind sprang full of dust from clocks that used dust for telling time.</i>	<i>Праз цёмную даліну, поўную ночы і пустаты з самага сэрца пустэчы вецер узняў слуп пылу з гадзіннікаў, што выкарыстоўваюць пыл, каб паказваць час.</i>
1. clause: word-for-word 2. clause: omitted 'itself' but added 'jedynie' [only] – different elements underlined 3. clause: translated as three clauses (due to requirements of Polish syntax), changed meaning – instead of 'clocks that used dust for telling time' there are 'zegary, które powoli zmieniały się w proch' [clocks that slowly changed into dust] * all clauses omit article 'the'	1. clause: word-for-word 2. clause: word-for-word 3. clause: translated as three clauses (due to requirements of Belarusian syntax), slightly changed meaning – instead of 'wind sprang full of dust' there is 'вецер узняў слуп пылу' [wind raised column of dust]  * all clauses omit article 'the'	
<i>W sercu nowego wiatru płonęły czarne słońca i miliony liści strząśniętych z jesiennych drzew za horyzontem.</i>	<i>There were black suns burning in the heart of this new wind and a million burnt leaves shaken from some autumn tree beyond the horizon.</i>	<i>У сэрцы гэтага новага ветру палалі чорныя сонцы, а таксама мільёны асенніх лісткаў, якія наспадлі з дрэваў далёка за даляглядам.</i>
Function words omitted: 'there were', 'the' (twice), 'this' (could have been easily translated to make clear that it is the wind that was mentioned in the previous sentence), 'a', 'some'.	Function words omitted: 'there were', 'the' (twice), 'a', 'some'. 'and' translated with 'а таксама' which fits better the Belarusian stylistic requirements 'beyond the horizon' expanded to 'далёка за даляглядам' [far beyond horizon]	
<i>Wiatr topił ziemię, rozciągał kości jak biały воск, sprawiał, że mętna krew gęstniała w błotnistej komorze mózgu.</i>	<i>This wind melted landscapes, lengthened bones like white wax, made the blood roil and thicken to a muddy deposit in the brain.</i>	<i>Гэты вецер плавіў краявіды, расцягваў косткі, быццам тая былі а белага воску, і прымушаў кроў загусаць у мозгу, як ліпкі бруд.</i>
Demonstrative pronoun 'this' omitted in the first clause. Last clause simplified by omission.	Article 'the' omitted in the last clause.	
<i>Wiatr niósł tysiące umierających dusz i zagubiony czas.</i>	<i>The wind was a thousand souls dying and all time confused and in transit.</i>	<i>Вецер быў тысячамі душаў, што паміралі ў страшэнным непакоі.</i>
Articles 'the' and 'a' omitted, last clause ('and in transit') omitted.	Articles 'the' and 'a' omitted, second half of the sentence shortened to two words with meaning distant from original.	

<i>Ciemność nabrzmiewała mgłą. Była to kraina niczyja, nie istniały tam rok ani godzina. Byli tylko ci mężczyźni, sami pośród bezkształtnej pustki, szronu i burzy, otoczeni piorunami, które pojawiały się za wielką, nabrzmiałą od błyskawic płytą z zielonego szkła.</i>	<i>It was a fog inside of a mist inside of a darkness, and this place was no man's place and there was no year or hour at all, but only these men in a faceless emptiness of sudden frost, storm, and white thunder which moved behind the great falling pane of green glass that was the lightning.</i>	<i>Гэта быў туман пасярэдзіне імжы, пасярэдзіне цемры. То было месца не для людзей, яно не мела часу ўвогуле. Толькі гэтыя двое стаялі пасярэдзіне жахлівае пустаты, дзе іх заснеў нечаканы мароз, шторм і белы гром, які рухаўся па-за вялікай масай зялёнае травы, што была маланкай.</i>
All articles omitted, some subordinate clauses translated with the use of participle rather than conjunction and verb.	All articles omitted, besides no changes.	

Saki, *Tea*, 1919

Saki, *Harbata*, 2010, translated by Janina Prystaŭka

Saki, *Herbata*, 1986, translated by Elżbieta Petrajtis-O'Neill

Polish (85w, F <sub>c</sub> =87%)	English (98w)	Belarusian (81w, F <sub>c</sub> =83%)
— <i>Właśnie jestem w trakcie piknikowego posiłku</i> — <i>oznajmiła.</i>	<i>"I'm having a picnic meal," she announced.</i>	— <i>У мяне сёння нешта накіталт пікніку, — заявіла яна.</i>
Present Continuous construction translated with Present Simple (due to lack of Continuous constructions in Polish): 'właśnie jestem w trakcie' [I'm just in the middle of]. Personal pronoun 'she' omitted.	Verb omitted (due to Belarusian syntax), the time of the action marked by adverb 'сёння' [today].	
— <i>W tamtym słoiku jest kawior.</i>	<i>"There's caviare in that jar at your elbow.</i>	— <i>У сłoіку каля твайго локця — ікра.</i>
Construction 'there is' translated with a verb 'jest' [is]. Omitted additional indicator of location: 'at your elbow'.	Construction 'there is' omitted (due to Belarusian syntax) and expressed with punctuation mark '—'.	
<i>Zaczynj jeść, a ja dokroję chleba.</i>	<i>Begin on that brown bread-and-butter while I cut some more.</i>	<i>Бяры вунь той хлеб з маслам, пакуль я адрэжу яшчэ крыху.</i>
'Brown bread-and-butter' simplified to 'chleb' (bread).	No changes.	
<i>Znajdź sobie filiżankę, imbryk z herbatą jest za tobą.</i>	<i>Find yourself a cup; the teapot is behind you.</i>	<i>Выбірай кубак, імбрычак за табой.</i>
Articles omitted, however 'teapot' is translated with the use of three words ('imbryk z herbatą').	Articles omitted, verb 'find' used with relative pronoun is translated with verb not requiring the use of pronoun.	
<i>A teraz opowiedz mi, co się u ciebie dzieje.</i>	<i>Now tell me about hundreds of things."</i>	<i>А зараз расказвай мне ўсё-ўсё-ўсё!</i>
'Hundred of things' interpreted as direct question about interlocutor affairs.	Repetition used for underlining a huge number of 'things'.	

<i>Nie wspomniała nic więcej, na temat jedzenia, tylko opowiadała różne zabawne historie i stworzyła taką atmosferę, że James też zaczął zabawnie i ciekawie opowiadać.</i>	<i>She made no other allusion to food, but talked amusingly and made her visitor talk amusingly too.</i>	<i>Больша пра ежу яна не сказала ні слоўка, аднак размаўляла з жывасцю, чым прымусіла і свайго суразмоўніка ажывіцца.</i>
Both occurrences of ‘amusingly’ translated with the use of a phrase rather than individual word.		No changes.
<i>Jednocześnie z dużą zręcznością kroїła chleb, smarowała go masłem, podawała paprykę i cytrynę w plasterkach, zamiast, jak inne kobiety, podawać usprawiedliwienie, dlaczego nie ma cytryny.</i>	<i>At the same time she cut the bread-and-butter with a masterly skill and produced red pepper and sliced lemon, where so many women would merely have produced reasons and regrets for not having any.</i>	<i>Адначасова яна спрытна нарэзала яшчэ хлеба, намазала яго маслам і прынесла чырвонага перцу і нарэзанага лімону, на адсутнасць якіх паскардзіліся б у такой сітуацыі іншыя жанчыны.</i>
All articles and personal pronouns omitted. Conditional construction and present continuous translated with one word verbs.		All articles and personal pronouns omitted. Conditional construction and present continuous translated with one word verbs.

Rowling, J.K., *Harry Potter and the Chamber of Secrets*, 1998

Rowling, J.K., *Harry Potter i komnata tajemnic*, 2000, translated by Andrzej Polkowski

Rowling, J.K., *Hary Poter i patajemnaja zala*, 2013, translated by Dzianis Muski

Polish (87w, F <sub>c</sub> =78%)	English (111w)	Belarusian (87w, F <sub>c</sub> =78%)
<i>- Wróć! - rzekł Ron, kiedy patrzyli, jak gnomy znikają pod żywopłotem na drugim końcu pola.</i>	<i>“They'll be back,” said Ron as they watched the gnomes disappear into the hedge on the other side of the field.</i>	<i>- Яны хутка вярнуцца, - гледзячы на то, як гномы знікаюць недзе на другім баку поля, сказаў Рон.</i>
First utterance translated with one word – no pronoun, no auxiliary verb and no adverb needed. All articles omitted.		No auxiliary verb in the character’s utterance. All articles omitted.
<i>- Uwielbiają nasz ogród.</i>	<i>“They love it here...”</i>	<i>- Ім тут надабаеца...</i>
Personal pronoun omitted.		Personal pronoun omitted.
<i>Tata jest dla nich za miękki, uważa, że są takie śmieszne...</i>	<i>Dad's too soft with them; he thinks they're funny...”</i>	<i>Тата замягкі да іх, ён лічыць, што гномы пацешныя...</i>
No major differences.		No major differences.
<i>W tym momencie rozległ się łoskot frontowych drzwi.</i>	<i>Just then, the front door slammed.</i>	<i>Адчуўся грукат зачыняючыхся дзвярэй.</i>
The verb translated in more elaborated way ‘rozległ się łoskot’ [a slam was heard].		The verb translated in more elaborated way (similarly to Polish), but time descriptor omitted.
<i>- Wrócił! - krzyknął George.</i>	<i>“He's back!” said George.</i>	<i>- Вярнуўся! - закрычаў Джордж.</i>
No personal pronoun, single-word verb instead of phrasal verb.		No personal pronoun, single-word verb instead of phrasal verb.
<i>- Tata wrócił!</i>	<i>“Dad's home!”</i>	<i>- Тата вярнуўся.</i>
Single-word verb instead of phrasal verb.		Single-word verb instead of phrasal verb.

<i>Pobiegli w stronę domu.</i>	<i>They hurried through the garden and back into the house.</i>	<i>Яны накінулі сад і наспяшалі дадому.</i>
Part of the sentence ('through the garden') omitted.		No articles, second clause translated with single-word adverb.
<i>Pan Weasley siedział już na kuchennym krześle; zdjął okulary, a oczy miał zamknięte.</i>	<i>Mr. Weasley was slumped in a kitchen chair with his glasses off and his eyes closed.</i>	<i>Містэр Візлі сядзеў на адным з кухонных крэслаў, скінуўшы акуляры і заплючыўшы вочы.</i>
Possessive pronouns omitted.		Possessive pronouns omitted.
<i>Był to chudy mężczyzna zaczynający łysieć, ale resztką włosów, jaka mu jeszcze pozostała, była ruda jak włosy jego dzieci.</i>	<i>He was a thin man, going bald, but the little hair he had was as red as any of his children's.</i>	<i>Ён быў хударлявым і лысым, але валасы, што ў яго яшчэ засталіся былі гэтакімі ж рудымі, як і ўва ўсіх яго дзяцей.</i>
No articles.		No articles.
<i>Miał na sobie długą zieloną szatę, zakurzoną i bardzo znozoną.</i>	<i>He was wearing long green robes, which were dusty and travel-worn.</i>	<i>На ім была пыльная і зношаная доўгая зялёная мантыя.</i>
No major differences.		No major differences.

## Appendix 4: Dispersion of chosen words in the EPB corpus (results by AntConc)

name of the file	file size (in words)			woman/	
		he	she	man/men	women
Adams_The-Hitchhiker's-Guide	46960	809	58	110	5
Aldridge_Hunter	74242	2234	222	248	28
Atwood_Penelopiad	2323	13	22	4	2
Bach_Jonathan-Livingston-Seagull	8928	224			
Barrie_Peter-Pan-and-Wendy	3066	39	74	1	2
Baum_The-Wonderful-Wizard-of-Oz	5113	32	96	12	16
Bradbury_Fahrenheit-451	45827	840	197	178	53
Bradbury_The-Dragon	1021	2		14	1
Bradbury_The-one-who-waits	2047	5		9	
Brown_The-Da-Vinci-Code	139453	1866	711	346	131
Bukowski_The-Most-Beautiful-Woman-in-Town	2375	2	67	13	5
Bukowski_You,your-beer	1761	28	11	3	4
Carver_Fat	1666	66	3	11	1
Cheever_enormous-radio	4367	44	83	13	13
Chesterton_The-Flying-Stars	5375	81	13	29	1
Christie_Dead-Man-s-Mirror	23339	446	163	77	15
Christie_Four-suspects	5150	60	31	12	4
Christie_Kidnapped-prime-minister	6388	108	2	32	3
Christie_Tape_Measure-Murder	4609	63	77	12	12
Christie_The-case-of-the-perfect-maid	4389	11	79	2	5
Christie_Veiled-Lady	4066	68	26	8	1
Dahl_Lamb	3492	62	142	16	1
Dahl_Skin	5872	136	4	33	5
Faulkner_An-Odor-of-Verbena	12728	202	98	35	17
Fitzgerald_Great-Gatsby1	1825	17		10	3
Fowles_Poor-Koko	14377	252	4	28	1
Gaiman_Coraline	30830	101	989	37	11
Golden_Memoirs-of-a-Geisha	186848	1478	2129	502	227
Golding_The-Scorpion-God	12842	240	56	142	27
Greene_Invisible-japanese	1363	24	23	1	
Greene_Looser-Takes-All	24294	277	254	80	33
Greene_Stamboul-Train	73212	1726	1106	253	80
Guin_Schroedinger-s-Cat	2929	51	4	2	
Gulik_He-came-with-the-rain	11008	251	60	42	5
Hemingway_For-Whom-the-Bell-Tolls	174272	3926	783	684	301
Hemingway_Indian-camp	1454	27	12	5	9
Hemmingway_The-Old-Man-and-the-Sea	26599	1167	9	272	5
Henry_Gift-of-magi	2073	17	33	3	
Hughes_Rain-horse	3822	153		1	
Huxley_Hubert-and-Minnie	4316	66	90	9	4
Ishiguro_Never-let-me-go-7.1	2085	1	45		
Jackson_Lottery	3378	32	16	19	5
Jackson_Trial	1976	2	72	3	1
James_Wailing-Well	4189	90	1	12	8
Joyce_Painful-case	3622	145	24	4	2
Joyce_Ulysses	137049	2234	771	310	158
Kerouac_On-the-road	8220	110	17	20	4
Kesey_One-Flew-Over-the-Cukoo-s-Nest	4087	60	39	8	1
King_The-Breathing-Method	24975	151	279	50	80
Kipling_How-the-Camel-Got-His-Hump	1071	23		5	
Kipling_How-the-First-Letter-Was-Written	3314	65	29	34	

Kipling_How-the-Rhinoceros-Got-His-Skin	863	37		1	
Kipling_How-the-Whale-Got-His-Throat	1039	58		7	
Kipling_Rikki-Tikki-Tavi	5787	175	33	17	
Kipling_The-Beginning-of-the-Armadillos	2595	38	16		
Kipling_The-Butterfly-That-Stamped	3353	61	22	7	3
Kipling_The-Cat-That-Walked-by-Himself	4043	58	32	25	47
Kipling_The-Elephant's-Child	2544				
Kipling_The_Sing-Song-of-Old-Man-Kangaroo	1124	55		6	
Lawrence_Sun	8511	148	268	29	33
Lewis_The-Magician-s-Nephew	41399	587	304	31	35
Lewis_The-Screw-tape-Letters	30635	468	40	142	30
London_White-Fang	72100	2247	266	274	18
Lovecraft_Cats-of-Ulthar	1342	11	1	8	1
Lovecraft_Festival	3645	18		16	8
Lovecraft_Herbert-West_Reanimator	12007	132	1	29	3
Lovecraft_Music-of-Erich	3451	45		10	
Lovecraft_Nyarlatotep	1137	7		4	
Lovecraft_Pickman-s-model	5513	60		17	1
Lovecraft_Polaris	1516	2		9	
Lovecraft_Rats-in-the-Walls	7883	33		35	1
Lovecraft_The-Alechemist	3679	19		13	
Lovecraft_The-Call-of-Cthulu	11883	106	2	58	2
Lovecraft_The-Hound	2972	8			
Lovecraft_The-Lurking-Fear	8115	26		15	
Lovecraft_Tomb	4152	9		8	
Lovecraft_What-the-moon-brings	722	3			
Lovecraft_Outsider	2567	2		1	
Mansfield_HER-FIRST-BALL	2590	24	47	26	2
Mansfield_THE-SINGING-LESSON	2089	12	42	4	1
Mathews_Singular-pleasures	4265	75	64	32	33
Maugham_Red	8890	291	79	48	12
Maugham_The-Fall-of-Edward-Barnard	12252	370	108	41	5
McCullers_Reflections-in-a-Golden-Eye	34690	918	435	52	22
McCullers_The_Ballad_of_Sad_Cafe	25407	334	304	52	20
McCullough_The-Thorn-Birds	224796	3500	2759	576	252
Milne_Winnie-the-Pooh	22116	644	22	1	
Orwell_Animal-Farm	30063	324	52	59	
Orwell_1984	94723	1893	379	187	82
Palahniuk_Fight-club16	2676	13		4	
Potter_The-Tale-of-peter-Rabbit	954	35	7		
Rowling_HP-and-the-Chamber-of-Secrets	85560	1536	269	29	5
Rowling_HP-and-the-Goblet-of-Fire	49599	794	167	53	13
Rowling_HP-and-the-Half_Blood-Prince	170352	3586	843	119	27
Rowling_HP-and-the-Philosopher-s-Stone	77587	1757	252	40	21
Rowling_HP-and-the-Prisoner-of-Azkaban	104931	2065	356	46	9
Saki_Tea	1583	16	9	1	7
Saki_The-open-window	1211	15	10	1	
Segal_Love-Story	25050	260	372	23	4
Sheckley_Battle	1553	24		3	
Steinbeck_The-Grapes-of-Wrath	179598	2860	1518	811	180
Tolkien_Smith-of-Wootton	10178	371	34	19	1
Tolkien_The-Fellowship-of-the-Ring	178232	2930	158	183	5
Tolkien_The-Hobbit,or-There-and-Back-Again	95546	1919	1	122	2

Tolkien_The-Return-of-the-King	135385	2422	186	470	24
Tolkien_The-Two-Towers	149914	2589	104	434	8
Travers_Mary-Poppins	8740	114	144	31	2
Waugh_Mr-Loveday-s-Little-Outing	2642	56	17	6	3
Woolf_Haunted-House	680	6	4		1
Woolf_Monday-or-Tuesday	301			1	1
Woolf_Solid-Objets	2394	46	2	4	1
Woolf_The-Legacy	3033	129	95	6	5
Woolf_The-String-Quartet	1480	3	5		2

name of the file	file size (in words)	kobieta/kobiety	mężczyzna/mężczyźni
Adams_Autostopem-przez-galaktykę.txt	39367	5	10
Aldridge_Łowca.txt	65198	16	28
Atwood_Peneliada.txt	1751	2	4
Barrie_Piotruś-Pan.txt	2780	2	
Baum_Czarnoksiężnik-z-Oz.txt	3887	6	5
Bradbury_Smok.txt	810	1	13
Bradbury_451-stopni-Fahrenheita.txt	39047	57	49
Brown_Kod-da-Vinci.txt	128752	138	113
Bukowski_Najpiękniejsza-dziewczyna-w-mieście.txt	1843	2	2
Bukowski_Ty,twoje-piwo.txt	1547	1	
Carver_Tłuścioch.txt	1333	1	
Cheever_Nowe-radio.txt	3761	4	7
Chesterton_Latające-gwiazdy.txt	4843	1	2
Christie_Czterech-podejrzanych.txt	4580	3	5
Christie_Idealna-służąca.txt	3563	6	
Christie_Lustro-nieboszczyka.txt	18701	13	19
Christie_Narzędzie-zbrodni.txt	3500	7	3
Christie_Porwanie-premiera.txt	5329	3	13
Christie_Dama-z-woalką.txt	2488	5	3
Dahl_Jagnię-na-rzeź.txt	2700	1	4
Dahl_Skóra.txt	3942	8	14
Faulkner_Zapach-werbeny.txt	11386	16	14
Fitzgerald_Wielki-Gatsby1.txt	1567	2	1
Gaiman_Koralina.txt	22767	11	8
Golden_Wyznania-gejszy.txt	122550	56	71
Golding_Bóg-skorpion.txt	9958	28	15
Greene_Pociąg-do-Stambułu.txt	63026	90	120
Greene_Stawka-o-żonę.txt	21129	32	14
Gulik_Przychodził-z-deszczem.txt	8703	4	6
Hemingway_Komu-bije-dzwon.txt	138798	242	112
Hemingway_Obóz-indiański.txt	1077	8	2
Hemingway_Stary-człowiek-i-morze.txt	20306	4	5
Henry_Upominek-święteczny.txt	1792	1	
Huxley_Hubert-i-Minnie.txt	3636	4	4
Jackson_Loteria.txt	2954	4	5
Jackson_Próba-sił.txt	1750	4	1
James_Jęcząca-studnia.txt	3359	8	7
Joyce_Przypadek-godny-ubolewania.txt	3003	3	3
Joyce_Ulises.txt	121857	140	60
Kerouac_W-drodze.txt	7007	5	4
King_Metoda-oddychania.txt	19368	58	10
Kipling_Kot-który-zawsze-chadzał-własnymi-drogami.txt	3344	50	14
Kipling_O-motyłu-który-tupał-nogą.txt	2696	1	3
Kipling_Riki-Tiki-Tavi.txt	4867	1	13
Kipling_W-jaki-sposób-został-napisany-pierwszy-list.txt	3020	1	
Lawrence_Słońce.txt	7114	43	29
Lewis_Listy-starego-diabła-do-młodego.txt	28568	37	23
Lewis_Siostrzeniec-czarodzieja.txt	35658	35	6
London_Biały-kieł.txt	56465	19	14

Lovecraft_Alchemik.txt	3018		8
Lovecraft_Festyn.txt	3273	1	1
Lovecraft_Grobowiec.txt	3518		3
Lovecraft_Herbert-West_Reanimator.txt	10662	1	10
Lovecraft_Polaris.txt	1250		4
Lovecraft_Przyczajona-groza.txt	8032		6
Lovecraft_Szczury.txt	6030	1	2
Lovecraft_Zew-Cthulhu.txt	10662	2	3
Mansfield_Jej-pierwszy-bal.txt	2182	1	4
Mansfield_Lekcja-śpiewu.txt	1758	1	1
Mathews_Osobne-przyjemności.txt	3664	40	26
Maugham_Rudy.txt	6652	10	4
Maugham_UpadekEdwardaBamarda.txt	9320	8	2
McCullers_Ballada-o-smutnej-knajpie.txt	21678	16	20
McCullers_W-zwierciadle-złotego-oka.txt	27345	20	11
McCullough_Ptaki-ciemistych-krzewów.txt	167236	201	181
Orwell_Folwark-zwierzęcy.txt	23290		3
Orwell_1984.txt	80304	67	45
Palahniuk_Podziemny-krąg.txt	2220		2
Rowling_HP-i-Czara-Ognia.txt	42557	10	16
Rowling_HP-i-Kamień-Filozoficzny.txt	67573	17	7
Rowling_HP-i-Komnata-Tajemnic.txt	76851	3	6
Rowling_HP-i-Półkwi-Księżę.txt	143358	19	36
Rowling_HP-i-Więzień-Azkabanu.txt	93000	6	17
Saki_Otwarte-okno.txt	1074		2
Saki_Herbata.txt	1229	4	4
Segal_O-miłości.txt	21823	2	254
Steinbeck_Grona-gniewu.txt	158367	211	8
Tolkien_Drużyna-pierścienia.txt	149114	3	5
Tolkien_Dwie-wieże.txt	131933	9	9
Tolkien_Hobbit,czyli-tam-i-z-powrotem.txt	81145	2	5
Tolkien_Kowal-z-Podlesia-Większego.txt	8637	1	14
Tolkien_Powrót-króla.txt	113255	22	1
Travers_Mary-Poppins.txt	7434	1	
Waugh_Mała-przechadzka-pana-Lovedaya.txt	2338	2	
Woolf_Nawiedzony-dom.txt	569	1	
Woolf_Poniedziałek-lub-wtorek.txt	285	1	1
Woolf_Rzeczy-trwałe.txt	2030	1	2
Woolf_Spadek.txt	2478	4	3

name of the file	file size (in words)	жанчына/жанчыны	мужчына/мужчыны
Adams_Autaspynam-pa-halaktycy.txt	43401	3	23
Aldridge_Palauniczy.txt	61323	15	7
Atwood_Pienielapijada.txt	1867	3	3
Barrie_Peter-Pen.txt	2531	2	
Baum_Czaraunik-Krainy-Oz.txt	3706	4	1
Bradbury_Cmok.txt	813	1	3
Bradbury_451-hradus-pa-Farenheitu.txt	31885	55	9
Brown_Kod-da-Vinczy.txt	117088	133	101
Bukowski_Samaja-pryhozhaja.txt	2007	6	5
Bukowski_Ty,tvajo-piva.txt	1446	3	1
Cheever_Vializnaje-radyjo.txt	3195	15	9
Christie_Biezdakornaja-pakajouka.txt	3491	5	
Christie_Czacviora-padazronych.txt	4473	4	4
Christie_Dama-pad-vuallu.txt	1785	1	
Christie_Niabożczykava-lustra.txt	19508	16	12
Christie_Vykradannie.txt	5143	3	8
Christie_Zabojstva.txt	3762	17	2
Dahl_Jahnia-na-bojni.txt	2880		3
Dahl_Skura.txt	4257	5	9
Faulkner_Pach-vierbieny.txt	10316	12	5
Fitzgerald_Vialiki-Hetsbi1.txt	1642		3
Gaiman_Karalina.txt	26281	11	6
Golden_Miemuary-Giejszy.txt	147350	137	267
Golding_Boh_Skarpion.txt	9424	22	11
Greene_Chto-prajhraje,biare-usio.txt	22626	27	12
Greene_Japonskija.txt	1240		1
Greene_Stambulski-ekspres.txt	64559	66	83
Guin_Kot-Szrodynhiera.txt	2328	1	1
Gulik_Jon-prychodziu-z-daždžom.txt	8539	5	
Hemingway_Indziejski-pasiołak.txt	1102	9	2
Hemingway_Pa-kim-zvonić-zvon.txt	148018	237	52
Hemingway_Stary-czaławiek-i-mora.txt	21821	4	5
Henry_Dary-mudarcou.txt	1750	1	
Huxley_Hiubiert-i-Mini.txt	3484	1	3
Jackson_Łatareja.txt	2470	7	4
Jackson_Pravierka-bojem.txt	1541	6	1
James_Studnia-stohnau.txt	3123	7	7
Joyce_Prykraje-zdarennie.txt	2791	5	3
Joyce_Ulis.txt	120170	114	89
Kerouac_Na-darozie.txt	6922	4	4
King_Mietad-pravilnaha-dychannia.txt	19670	79	9
Kipling_Jak-było-napisana-pierszaje-piśmo.txt	2824	1	1
Kipling_Kot-jaki-hulau-sam-saboku.txt	3328	51	21
Kipling_Matylok-jaki-tupau-nahoj.txt	2835	4	
Kipling_Ryki-ciki-tavi.txt	5258		1
Lawrence_Sonca.txt	7043	24	22
Lewis_Listy-staroha-diabła.txt	26253	25	19
Lewis_Plamiennik-czaraunika.txt	32519	35	5
London_Bieły-kłyk.txt	58242	22	9
Lovecraft_Alchimik.txt	2876		1
Lovecraft_Hierbiert-Uest.txt	10785	3	

Lovecraft_Klicz_Ktulhu.txt	10169	2	
Lovecraft_Pacuki.txt	6748	2	
Lovecraft_Palarys.txt	1270		1
Lovecraft_Stojeny-żach.txt	7886		2
Mansfield_Pierszy-bal.txt	2146	3	10
Mansfield_Urok-śpievau.txt	1756	1	2
Mathews_Pryvatnyja-pryjemności.txt	3543	38	26
Maugham_Padziennie-Edwarda-Barnarda.txt	9084	7	3
Maugham_Rudy.txt	6515	9	1
McCullers_Adliustravanni-u-zalatym-voku.txt	28153	22	11
McCullers_Balada-pra-sumnaje-kafe.txt	20271	20	20
McCullough_Ptuski-na-ciemiach.txt	202507	243	187
Orwell_1984.txt	83050	83	13
Rowling_HP-i-Filasofski-Kamień.txt	67444	25	5
Rowling_HP-i-Kielich-ahniu.txt	43751	32	15
Rowling_HP-i-patajemnaja-zała.txt	77380	9	4
Rowling_HP-i-Prync_Paukrouka.txt	132422	18	20
Rowling_HP-i-Viazień-Azkabana.txt	95739	10	21
Saki_Adczynienaje-akno.txt	968		2
Saki_Harbata.txt	1268	6	1
Segal_Historyja-kachannia.txt	23548	4	6
Sheckley_Bitva.txt	1382		1
Steinbeck_Hronki-hnievu.txt	152755	197	161
Tolkien_Chobit.txt	76103	2	
Tolkien_Dżwie-vieży.txt	119867	4	1
Tolkien_Viartannie-karala.txt	104411	14	6
Travers_Mery-Popins.txt	7403	3	
Woolf_Dom-z-pryvidami.txt	541	2	
Woolf_Paniadzielak.txt	280		1
Woolf_Sapraudnyja-reczy.txt	1919	1	
Woolf_Spadak.txt	2497	2	3
Woolf_Strunny.txt	1247		1

## Appendix 5. Highest ranked collocates of chosen words from the EPB corpus.

<i>man</i>	Freq	MI	LL	<i>men</i>	Freq	MI	LL
portly	5	8,284	49,9	armed	15	7,643	131,351
elderly	22	7,962	207,369	group	27	7,172	217,867
wizened	6	7,962	56,537	uniform	6	6,67	44,043
faced	17	7,445	145,968	bid	5	6,533	35,723
drunken	7	7,225	57,704	fro	5	6,496	35,461
old	781	7,222	6555,001	equal	5	6,34	34,361
wealthy	7	7,107	56,433	younger	11	6,258	74,339
young	237	7,087	1914,124	eight	13	6,075	84,53
haired	14	7,033	111,313	many	110	6,057	716,426
fat	53	6,874	409,16	among	34	6,038	219,565
valiant	5	6,825	38,185	two	183	6,034	1190,869
married	10	6,477	71,247	tall	22	5,968	139,85
mortal	11	6,452	77,976	gather	17	5,886	106,074
blind	33	6,422	232,603	four	45	5,881	280,923
dumb	6	6,404	42,109	surround	6	5,863	37,225
intelligent	11	6,392	77,013	create	5	5,845	30,897
eyed	15	6,377	104,709	these	78	5,84	483,485
destiny	5	6,346	34,664				
taller	6	6,325	41,417				
stranger	27	6,269	184,289				
decent	8	6,207	53,868				
tall	49	6,193	329,32				
brace	6	6,155	39,946				
crazy	13	6,068	84,94				
aged	6	6,066	39,181				
uniform	7	5,962	44,674				

<i>boy</i>	Freq	MI	LL	<i>boys</i>	Freq	MI	LL
naughty	10	10,066	122,835	older	7	8,027	64,251
foolish	7	7,848	62,701	boy	12	6,559	85,862
dear	28	7,197	225,594	play	7	6,538	49,784
snarl	6	6,779	44,705	among	5	6,199	33,204
nine	6	6,097	39,027	five	7	6,195	46,486
boy	24	6,028	154,247	girl	9	5,971	57,05
pale	8	5,789	48,664	shout	5	5,909	31,221
girl	20	5,592	116,49	two	22	5,905	138,266
little	54	5,503	309,65	together	6	5,7	35,77
poor	9	5,495	51,134	young	6	5,641	35,28
old	45	5,431	253,295	around	10	5,413	55,819
small	17	5,38	94,051	other	19	5,404	106,28
quiet	6	5,317	32,616	little	17	5,366	94,116
father	11	5,23	58,552	three	7	5,235	37,326
carry	9	5,114	46,482	small	5	5,145	26,027
shout	8	5,057	40,686	these	6	5,067	30,614
who	52	5,053	266,308	better	5	5,054	25,412
nod	5	5,033	25,257	those	5	4,831	23,917
throw	7	4,9	34,112	work	6	4,781	28,312
head	31	4,89	151,359	put	7	4,757	32,822
shoulder	8	4,86	38,562	sit	7	4,65	31,822
remember	12	4,86	57,886	one	20	4,365	84,016
nice	5	4,838	23,942	through	7	4,354	29,069
live	10	4,731	46,494	though	5	4,299	20,38
my	68	4,72	318,797	good	7	4,29	28,479
kill	8	4,715	37,01	all	25	4,275	102,394
lip	5	4,686	22,922	from	22	4,272	89,833
shake	7	4,616	31,453	than	7	4,175	27,424
good	25	4,596	112,046	stand	6	4,143	23,241
family	5	4,545	21,979	will	18	4,115	69,588
play	5	4,521	21,825	with	34	4,06	130,438
large	6	4,493	25,968	some	7	4,059	26,36
believe	7	4,453	29,935	who	9	4,053	33,866
				they	53	3,908	195,433
				get	13	3,872	45,977
				too	5	3,815	17,219

<i>woman</i>	Freq	MI	LL	<i>women</i>	Freq	MI	LL
haired	10	8,31	96,621	child	43	7,625	373,18
aged	7	8,051	64,963	dress	11	6,999	85,319
plump	8	7,988	73,512	older	6	6,891	45,593
elderly	5	7,587	43,043	scream	7	6,651	50,87
handsome	8	7,379	66,502	baby	6	6,298	40,667
slender	5	7,256	40,683	young	16	6,142	105,249
young	78	7,246	637,239	beautiful	7	6,123	45,774
younger	11	7,09	86,942	two	46	6,055	299,147
naked	8	7,015	62,372	among	8	5,963	50,568
sixty	6	6,964	46,338	often	7	5,88	43,44
beautiful	24	6,72	177,386	most	19	5,87	117,966
old	132	6,42	929,653	man	80	5,856	502,437
ugly	5	6,384	34,546	three	20	5,836	123,279
marry	10	6,287	67,782	woman	18	5,547	103,787
whom	11	6,122	72,038	many	19	5,537	109,327
doorway	6	6,024	38,465	wear	8	5,489	45,386
lovely	7	5,998	44,627	old	30	5,464	170,091
scream	10	5,984	63,58	four	8	5,402	44,435
thirty	9	5,968	57,013	street	5	5,301	27,065
older	7	5,932	43,99	hundred	5	5,151	26,049
stupid	5	5,824	30,671	both	7	5,004	35,096
fat	7	5,716	41,907	who	32	4,97	160,105
Mrs	7	5,67	41,462	bear	6	4,957	29,696
middle	9	5,656	53,151	around	13	4,878	63,097
poor	14	5,568	81,032	laugh	7	4,804	33,221
dress	9	5,528	51,572	together	6	4,787	28,327
fine	8	5,489	45,412	soon	6	4,747	28,007
thin	9	5,424	50,292	near	5	4,698	23,002
who	99	5,418	558	house	10	4,668	45,677
daughter	5	5,365	27,521	white	7	4,606	31,369
tall	8	5,341	43,789	watch	9	4,597	40,24
street	11	5,257	58,978	those	8	4,596	35,751
approach	5	5,18	26,266	some	19	4,586	84,945
love	21	5,15	109,667	black	7	4,564	30,977
				stand	15	4,552	66,293
				run	10	4,541	43,979
				always	8	4,528	35,026
				round	6	4,524	26,222
				world	6	4,514	26,146
				front	5	4,492	21,632
				bring	7	4,478	30,174
				girl	6	4,472	25,808
				work	9	4,452	38,515
				use	7	4,419	29,622
				boy	5	4,382	20,905
				talk	7	4,338	28,878
				become	5	4,337	20,606
				such	6	4,271	24,211
				other	16	4,243	64,182



<i>mężczyzna</i>	Freq	MI	LL	<i>mężczyźni</i>	Freq	MI	LL
krępy	6	9,076	64,548	mundur	5	8,551	49,685
ciemnowłosy	5	9,061	53,671	grupa	14	8,438	137,068
rosły	7	8,992	74,427	kobieta	74	7,762	661,097
przystojny	12	8,816	124,508	ubrać	8	7,2	64,27
dorosły	13	8,417	127,273	cztery	16	6,553	114,369
chudy	14	8,242	133,531	ognisko	5	6,541	35,569
szczupły	8	8,052	74,061	niebieski	5	6,507	35,334
średni	5	7,952	45,555	oba	21	6,448	147,23
wzrost	6	7,195	48,218	zebrać	5	6,428	34,787
niski	17	7,102	134,588	młody	21	6,387	145,456
kobieta	62	6,827	469,848	większość	6	6,317	40,838
wysoki	39	6,699	287,635	trzy	25	6,312	170,736
rozpoznać	5	6,41	34,694	dwa	50	6,228	337,725
nagi	6	6,352	41,16	tłum	5	6,108	32,581
przebywać	5	6,347	34,255	nosić	6	6,085	38,918
samotny	6	6,215	40,022	rzucać	5	6,061	32,261
ubranie	6	6,189	39,807	wychodzić	5	5,876	30,994
brudny	5	6,178	33,09	rozmawiać	6	5,792	36,506
spodnie	5	6,154	32,923	siedzieć	17	5,528	97,608
broda	6	6,154	39,513	gdyż	5	5,376	27,582
obcy	7	6,023	44,837	widok	5	5,327	27,246
uważnie	5	5,891	31,11	wszystek	22	5,296	119,556
niebieski	5	5,827	30,672	patrzeć	7	5,253	37,465
młody	21	5,707	125,69	wśród	7	5,243	37,37
wiek	9	5,635	52,873	często	5	5,207	26,435
skraj	5	5,585	29,013	dziecko	10	5,194	52,764
włos	15	5,539	86,231	lubić	6	5,172	31,448
but	5	5,537	28,686	twarz	21	5,135	109,497
jakiś	53	5,529	306,017	prawdziwy	5	5,109	25,774
podejść	11	5,512	62,791	chłopiec	6	4,952	29,667
stary	42	5,51	240,956	ku	12	4,91	58,753
miłość	7	5,443	39,273	wiele	13	4,88	63,144
obserwować	6	5,429	33,543	czekać	9	4,88	43,66
czarny	25	5,42	139,912	stół	6	4,878	29,069
syn	10	5,387	55,366	ciało	6	4,802	28,451
szyja	5	5,353	27,433	ruszyć	6	4,782	28,294
siedzieć	24	5,346	131,873	jeden	28	4,77	132,549
prawdziwy	9	5,278	48,486	słuchać	5	4,722	23,172
plaszcz	5	5,273	26,885	przy	16	4,692	73,767
spotkać	9	5,263	48,309	usta	6	4,685	27,516
spojrzenie	8	5,198	42,224	zatrzymać	5	4,669	22,816
nazwisko	5	5,181	26,263				

<b>chłopiec</b>	<b>Freq</b>	<b>MI</b>	<b>LL</b>	<b>chłopcy</b>	<b>Freq</b>	<b>MI</b>	<b>LL</b>
chudy	6	7,375	49,691	mał	6	11,156	82,168
drogi	7	7,113	55,413	dziewczę	5	8,507	49,238
łagodnie	6	6,731	44,286	młody	10	6,423	69,69
blady	9	6,602	64,858	chłopiec	5	5,796	30,462
dziewczynka	6	6,576	42,995	siedzieć	9	5,717	53,986
mały	40	6,091	261,559	zawołać	6	5,678	35,606
chłopiec	15	5,95	94,694	mały	10	5,522	57,359
nieść	5	5,854	30,843	oba	5	5,484	28,338
odrzec	9	5,794	54,829	wy	9	5,427	50,435
biedny	6	5,704	35,793	trzy	6	5,359	33,011
podać	5	5,527	28,61	matka	6	5,241	32,052
porządek	5	5,515	28,527	kilka	6	5,214	31,829
kochać	9	5,486	51,035	mężczyzna	5	5,196	26,387
brać	6	5,477	33,927	dwa	11	5,15	57,585
cicho	7	5,366	38,535	kobieta	5	4,981	24,94
pamiętać	10	5,287	54,008	wszystek	8	4,944	39,576
młody	12	5,255	64,321	wszyscy	6	4,824	28,677
ów	8	5,253	42,824	inny	9	4,752	42,24
pokazać	5	5,226	26,563	stary	7	4,712	32,433
jako	9	5,211	47,677	mój	12	4,638	54,616
brat	5	5,209	26,446	wiele	5	4,608	22,437
mój	47	5,177	249,012	dom	7	4,586	31,262
ramię	13	5,001	65,231	dla	8	4,431	34,106
spojrzenie	5	4,875	24,192	chwila	8	4,308	32,799
zawołać	9	4,831	43,064	dobry	5	4,282	20,281
droga	19	4,828	91,062	z	77	4,253	328,885
matka	12	4,81	57,12	do	45	4,142	180,249
nazywać	5	4,799	23,685	jeden	8	4,07	30,296
siedzieć	12	4,701	55,364	za	15	4,022	56,198
wejść	7	4,653	31,809	przez	11	4,019	41,037
obok	6	4,613	26,933	cały	6	3,961	21,833
odezwać	5	4,589	22,281	żeby	7	3,835	24,357
prosić	9	4,55	39,684	a	25	3,771	86,325
dać	12	4,543	52,838	ręka	5	3,702	16,511
zrozumieć	5	4,529	21,877	przed	5	3,697	16,477
wziąć	8	4,483	34,546	na	55	3,686	189,358
powiedzieć	62	4,45	269,055	i	75	3,629	257,961
mężczyzna	8	4,443	34,121	oko	5	3,526	15,383
twarz	16	4,418	67,865	siebie	9	3,428	26,688
rzec	16	4,38	67,053	ale	12	3,408	35,384
spytać	7	4,364	29,117	się	65	3,377	200,099
				czy	6	3,328	16,98

<b>kobieta</b>	<b>Freq</b>	<b>MI</b>	<b>LL</b>	<b>kobiety</b>	<b>Freq</b>	<b>MI</b>	<b>LL</b>
średni	10	8,618	101,121	rodzić	6	8,374	58,119
tęgi	9	8,509	89,533	mężczyzna	75	7,996	694,945
nieszczęsny	5	7,423	41,849	dziewcę	7	7,886	62,97
młody	81	7,32	671,145	dziecko	40	7,194	323,127
typ	7	7,317	57,549	niektóry	12	7,127	95,271
chudy	8	7,101	63,325	większość	8	6,732	59,068
szcuple	5	7,04	39,137	czynić	6	6,62	43,349
nagi	11	6,893	83,869	ubrać	5	6,521	35,434
naturalny	6	6,742	44,453	towarzystwo	5	6,497	35,268
mężczyzna	63	6,731	468,677	kobieta	24	6,138	158,116
suknia	6	6,547	42,816	młody	16	5,995	102,051
wiek	19	6,379	131,333	piękny	13	5,963	82,292
przedstawiać	5	6,317	34,078	oba	15	5,963	94,99
ładny	7	6,206	46,643	wszystek	32	5,837	198,01
znosić	5	6,157	32,964	nosić	5	5,822	30,619
siwy	5	6,142	32,86	wśród	9	5,606	52,512
miły	7	5,974	44,388	widok	6	5,59	34,852
skinać	6	5,913	37,539	trzy	15	5,575	87,033
kochanie	5	5,832	30,72	wokół	6	5,482	33,97
głupi	8	5,801	48,833	wiek	5	5,467	28,197
mądry	5	5,65	29,47	dwa	27	5,339	148,491
mąż	8	5,627	46,916	chłopiec	7	5,175	36,723
biedny	9	5,6	52,461	cztery	6	5,138	31,168
stary	56	5,591	327,808	prawo	5	5,04	25,306
piękny	20	5,571	115,938	wiele	14	4,986	70,037
odrzec	12	5,52	68,658	należać	6	4,981	29,9
przyglądać	9	5,42	50,24	kochać	5	4,963	24,783
krzyczeć	5	5,409	27,824	obok	6	4,938	29,547
poznać	9	5,395	49,932	praca	5	4,908	24,414
posłuchać	5	5,365	27,52	znać	9	4,899	43,901
drobny	6	5,36	32,993	także	6	4,878	29,069
kość	5	5,334	27,313	świat	10	4,843	48,032
kochać	13	5,327	70,965	stary	16	4,798	76,038
dziewczyna	14	5,32	76,305	u	10	4,777	47,15
niezwykły	5	5,313	27,168	kilka	9	4,692	41,401
odezwać	12	5,163	62,821	lub	6	4,692	27,571
milczeć	5	5,079	25,58	biały	8	4,538	35,147
jakiś	48	5,052	245,402	zawsze	9	4,531	39,469
				dla	18	4,495	78,31
				inny	16	4,475	69,147
				bez	11	4,465	47,304
				patrzeć	6	4,443	25,585
				siedzieć	7	4,248	28,053
				nad	11	4,19	43,306
				chodzić	6	4,15	23,26
				dom	11	4,132	42,471
				życie	8	4,129	30,818
				taki	16	4,08	60,808

<b>dziewczyna</b>	<b>Freq</b>	<b>MI</b>	<b>LL</b>	<b>dziewczyny</b>	<b>Freq</b>	<b>MI</b>	<b>LL</b>
jasnowłosa	8	10,095	98,399	piękny	8	7,568	68,593
ładny	11	7,456	92,437	młody	5	6,622	36,219
miły	9	6,933	69,016	dwa	10	6,211	67,206
piękny	32	6,846	242,373	jeden	8	5,269	43,355
młody	31	6,532	221,224	czas	5	4,744	23,446
chłopak	6	6,297	40,684	który	15	4,621	69,055
mieszkać	9	6,05	57,996	za	9	4,484	39,362
córka	5	6,025	32,02	z	35	4,315	152,805
ostrożnie	6	5,859	37,065	powiedzieć	9	4,296	37,117
włos	15	5,802	91,64	jak	8	3,721	26,93
obcy	5	5,8	30,479	być	37	3,606	127,551
podać	6	5,698	35,742	ten	10	3,59	32,124
wspaniały	6	5,638	35,253	mieć	8	3,588	25,569
podejść	10	5,637	58,785	co	7	3,576	22,215
imię	8	5,62	46,824	do	13	3,55	41,383
grać	5	5,582	28,988	w	27	3,52	87,785
odejść	8	5,463	45,103	a	9	3,496	27,776
opowiadać	5	5,432	27,963	na	20	3,426	61,529
zwrócić	8	5,43	44,748	to	14	3,387	41,764
kochać	9	5,394	49,906	o	6	3,375	17,469
dziewczyna	9	5,28	48,515	się	28	3,361	85,546
obok	10	5,258	53,615	tak	5	3,317	14,161
unieść	5	5,25	26,724	i	26	3,299	76,982
tamten	8	5,155	41,757	ja	6	3,256	16,579
poznać	5	5,144	26,007	nie	16	2,922	38,67
moment	6	5,038	30,357	że	7	2,678	14,459
spać	6	5,036	30,341	on	15	2,608	30,478
siedzieć	16	5,024	80,83				
pewien	12	5,008	60,302				
kobieta	14	4,944	69,176				
przy	25	4,919	123,025				
wzrok	8	4,912	39,133				
ramię	12	4,793	56,842				
zapytać	17	4,789	80,527				
kiedyś	5	4,758	23,407				
zauważyć	7	4,683	32,08				
spotkać	5	4,678	22,874				

<b>мужчына</b>	<b>Freq</b>	<b>MI</b>	<b>LL</b>	<b>хлопец</b>	<b>Freq</b>	<b>MI</b>	<b>LL</b>
мажны	11	10.238	137.639	неблаг	5	10.565	63.904
пажылы	6	9.476	67.910	прыгожы	5	6.625	36.158
ніводзін	5	8.378	48.520	добры	10	6.288	67.883
Гэта	5	7.345	41.194	мог	6	5.302	32.566
жанчын	8	7.345	65.942	адказаць	10	5.260	53.863
жанчына	61	7.317	504.643	нешта	5	4.880	24.269
рост	5	6.815	37.486	той	15	4.683	69.447
сядзіць	5	6.815	37.486	пачаць	5	4.627	22.580
напрыклад	6	6.566	42.918	яго	9	4.505	39.326
мужчына	16	6.515	113.523	сказаць	17	4.178	67.446
старэйшы	6	6.400	41.535	ўжо	5	3.893	17.751
трыццаць	6	6.391	41.458	гэта	18	3.870	64.255
належыць	5	6.362	34.341	свой	11	3.826	38.346
высокі	11	6.341	75.315	з	40	3.821	143.688
валаса	9	5.766	54.487	што	41	3.789	145.730
сапраўдны	11	5.659	65.004	ты	10	3.708	33.307
любы	6	5.574	34.725	як	18	3.677	59.758
хацумома	6	5.531	34.367	так	8	3.462	24.076
паміж	11	5.437	61.673	а	16	3.448	48.313
сядзець	19	5.432	106.600	за	11	3.429	32.762
ўсе	8	5.336	43.719	на	34	3.343	100.415
нейкі	14	5.323	76.371	ён	44	3.327	130.741
моцны	6	5.282	32.341	тое	7	3.323	19.818
прыгожы	6	5.222	31.848	да	14	3.236	38.447
кінуць	5	4.992	24.980	калі	12	3.195	32.260
побач	10	4.967	49.686	і	56	3.189	159.240
твар	16	4.959	79.442	яшчэ	5	3.132	12.937
цела	7	4.828	33.448	ў	24	3.117	63.191
малады	5	4.819	23.815	быць	21	3.102	54.694
выйсці	8	4.807	38.007	ж	5	3.100	12.737
хлопчык	8	4.749	37.391	не	32	3.081	83.754
выгляд	6	4.746	27.999	які	10	3.040	24.902
адзін	29	4.736	135.751	я	26	3.029	65.813
адзіны	5	4.724	23.178	у	17	2.985	41.586
дзяўчына	5	4.722	23.167	толькі	5	2.912	11.596
іншы	21	4.714	97.483	гэты	5	2.854	11.242
чакаць	11	4.621	49.563	яны	8	2.575	15.433
пачуць	7	4.597	31.284	але	7	2.446	12.445
падобны	7	4.591	31.231	яна	8	2.162	11.715
другі	8	4.580	35.577				
пры	6	4.537	26.326				
сярод	5	4.486	21.592				
такі	10	4.475	43.107				
павярнуцца	5	4.452	21.370				
павінен	6	4.429	25.467				
вядома	7	4.417	29.606				
заўважыць	8	4.391	33.574				
любіць	5	4.382	20.905				
напэўна	5	4.358	20.743				
кожны	11	4.333	45.361				
які	75	4.281	309.924				