Mikhail Mikhailov 2021. Corpus-based analysis of Russian translations of the Animal Farm by George Orwell. In Meng Ji and Michael Oakes (eds.). *Corpus Exploration of Lexis and Discourse in Translation*, Routledge, 56-82.

Corpus-based analysis of Russian translations of the Animal Farm by George Orwell.

Mikhail Mikhailov

Abstract

The phenomenon of multiple translations of same classical works has been discussed extensively since 18th century. The dominant approach however is to study retranslating as a cultural, not a linguistic phenomenon. For unknown reasons, little use has been made of corpus data in research on this topic, although corpora of retranslated texts would seem to be a natural source of empirical data. Studying multiple translations with the help of corpus-based methods makes it possible to obtain the general picture of the data and to find its critical points. The quantitative data can be used for developing criteria for evaluation of the texts.

In this paper, six Russian translations of George Orwell's Animal Farm are studied. The translations are compared against an unedited machine translation. A multidimensional scaling of the frequency-list-based distance matrix was performed. The analysis demonstrated that the most frequently republished translations are the most distant from the MT. The keyword analysis of the translations confirmed the findings of the MDS analysis and gave concrete clues on the lexical items typical for certain translations.

Keywords: retranslation, keyword analysis, distance measure, lexical similarity

1. Is it possible to study retranslations with corpus methods?

Habent sua fata libelli. The fate of most literary works is to be published only once and to fall after that into oblivion. Very few are republished and are still read by the next generation. And only in extremely rare cases the work becomes classics and is republished many times, is read by many generations and sometimes even survives the language it was written in. A national classic may turn into a world classic, if it is translated into other languages and is popular in different cultures.

The fate of a classical work inside its own culture is dull: its text after a number of editions reaches a certain stability, becomes canonised and mummified and does not change afterwards. This canonical version is reproduced faithfully in academic editions and only suffer very minor changes

in orthography and punctuation in editions for the general public, the main change being the number of footnotes and endnotes with explanations for the readers.

The fate of a classical work in other cultures is more exciting. After it becomes a word classic, its translations into other languages are also republished, however, in many cases old translations are revised and corrected, and after some time entirely new translations appear. Some literary works are translated many times and new translations continue to appear. For example, Shakespeare's Othello was translated into German 38 times from 1766 to 2010 (Alharbi et al 2015: 1). Other works, after being translated several times, reach their canonical form in the target language. It is difficult to say, whether this is a matter of pure luck the work had, or a matter of interest towards the work in the receiving culture.

The reasons for retranslating might be very different. Old translations were often done via third language and got the savour of this intermediary language. They were often abridged and some important passages may be missing. The translators of the good old days did not have modern dictionaries, encyclopedias and corpora and sometimes made errors in their translations. Some of the old translations are very domesticated and as result the original text is hard to recognise. In some cases, the translations were censored or self-censored which made the translation differ dramatically from the original. Last but not least: the language of the old translations may be hard to understand due to changes in the target language.

According to the retranslation hypothesis expressed by Goethe and in the 1990 explicitly formulated by Antoine Berman (see e.g. Deane-Cox 2014: 3) retranslations are continuous attempts to reach the ingeniousness of the source text in the target language. They pass various stages: heavily domesticated translations to make the readers acquainted to the work, foreignised translations to make the readers acquainted to the language and structure of the work, and finally – the optimal representation of the source text in this language. However, the hypothesis is difficult to confirm on empirical data. It cannot be taken for granted that each new translation is a step forward: that depends on the translator and other factors. Besides, the quest for producing the best translation is not the only reason for retranslating a work. Many researchers criticise Berman's theory and show that it does not work on their data (e.g. see Deane-Cox 2014, Kuusi 2014). Anyway, this hypothesis shows the interrelations between different translations of the same work and the possibility of their influence on each other.

The phenomenon of retranslation of literary works is a topic that has been discussed extensively in translation studies (Cadera & Walsh 2017, Koskinen & Paloposki 2015, etc.). A series of conferences with the topic “Retranslation in Context” was initiated in Istanbul in 2013 and successfully continued in 2015 (Istanbul), 2017 (Ghent), and 2019 (Madrid) (Pouke & Gallego

2019: 13). The special issue 27:1 (Voice in Retranslation) of the journal Target was devoted to retranslation. The dominant approach of most publications on the topic is to study retranslating as a cultural rather than as a linguistic phenomenon. For unknown reasons, little use has been made of corpus data in research on this topic, although corpora of retranslated texts would seem to be a natural source of empirical data. The electronic data makes it possible to compare different versions, to show how close or distant they are, did they influence each other, etc.

However, many researchers have studied retranslations by means of manual comparison (e.g. Brounlie 2006, Desmidt 2009, Deane-Cox 2014). Quite a few attempts to use corpus methods for studying multiple translations have been made so far: Jeremy Munday used corpus methods for comparing of two English translations of a newspaper article by García Marquez (Munday 1998), Tom Cheesman, Kevin Flanagan et al study in their project "Version, Variation, Visualisation" German translations of Shakespeare's works (Cheesman, Flanagan et al 2017, Alharbi et al 2015), Henry Jones does corpus-based analysis of English translations of Thucydides (Jones 2020). There are at least two projects on creating a massively parallel Bible corpus with thousand or more versions of the Bible in more than 800 languages (Mayer & Cysouw 2014, McCarthy et al 2020). The scholars do not however approach the Bible translations from the point of view of translation studies (at least at the current stage) but rather use the resource for typological and contrastive research.

In this paper, six Russian translations or George Orwell's Animal Farm are examined with the help of corpus-based methods. To compare the texts, frequency-list-based distance measure and keywords lists are used. A machine translation of the text performed with Microsoft Translator is used as a baseline for comparison.

2. Methods of studying lexical similarity of texts

To study and compare groups of texts distance measures can be used. This method can help to get a general picture, to find which of the texts are closer to each other without offering any explanations. To get deeper into the matter, keyword search can be helpful. While browsing lists of keywords, one can find the actual words that make the text (or a group of texts) in question different from other texts. The keyword search may offer good hints, but it compares texts only pairwise. In my study, I will try to combine these two methods.

2.1. Distance measures

The most straightforward way to find out how close the texts are is to compare their lexicons and frequency lists comparison is often used to study corpora or separate texts (see Kilgarriff 1997 and 2001, Piperski 2017 and 2018). The comparison can be based on unlemmatised or lemmatised

frequency lists. Using the complete frequency lists does not produce stable results, because a frequency list drawn from a large text would be much longer than that of a short text and thus long and short texts would be incomparable, and even for the texts of the same length low-frequency items would impede the comparison (see also NN 2019, 167-168). Therefore, the first X most frequent words (hereafter – MFW) are usually taken. Very short frequency lists would not help to find differences between texts, while very long ones would not work on short texts. In stylometric research MFW 100 (i.e. the first 100 words from frequency lists) is popular (see e.g. Eder et al 2016).

To measure distances between texts, vectors with normalised frequencies of MFW for each text are composed and a distance matrix of such vectors is calculated (see Kilgarriff 1997, 2001 and 2008 for details). The next step is to analyse the distance matrix using cluster analysis, multidimensional scaling or some other method. There also exists a Stylo package on R that performs stylometric routines on groups of texts (Eder et al 2016).

In my earlier research (NN 2019) I have demonstrated that translations of the same works tend to be lexically very close, and a frequency-list-based distance measure makes it possible to distinguish retranslations of the same texts from other translated and non-translated texts (e.g. same author, same topic etc.). This finding is not surprising: the texts based on the same source text are of approximately the same length and have many overlaps in vocabulary. Besides, when studying retranslations, differences between texts are more important than similarities. Why some texts become outliers? Usually these are old translations, but can there be exceptions? Frequency-list-based distance measures yield only a very general picture without telling the researcher what is particularly different in the texts compared.

2.2. Keyword analysis

A more detailed analysis of the differences in vocabulary between retranslations can be performed with the help of keyword lists. Keywords are words from the research data that have frequencies significantly different from their frequencies in the reference corpus (NN 2016: 133-144). Keyword analysis is very popular: it is used in many fields of linguistics, digital humanities and translation studies (see e.g. Cermakova & Farova 2010, Fidler and Cvrček 2015, Johnson & Esslin 2006, Kemppanen 2004, 2008, Milizia 2010, Seale et al 2006, Wilkinson 2014).

The popular software package WordSmith Tools has Keywords utility that can be used for getting lists of keywords for single texts or collections of texts (Scott & Tribble 2006, https://lexically.net/downloads/version8/HTML/keywords.html). The SketchEngine online tool also has a utility of the same name that can be used for querying the ready-made corpora or users' own text collections (https://www.sketchengine.eu/my_keywords/keyword/). Also, keywords lists can be

extracted from two frequency word lists with the help of any statistical package (R, SPSS etc) or even with spreadsheet software like Microsoft Excel or Libre Office Calc.

In fact, there exist many alternative ways to compute the "keyness": chi-square and log-likelihood, among others. The simplest method available is surprisingly effective; this is Adam Kilgarriff's simple measure (Kilgarriff 2009), which is calculated with the formula:

$K = (Fe + N) / (Fc + N)$,

where Fe and Fc are the relative frequencies of the item in the experimental and control data (e.g. expressed in items per million, ipm), and N is a smoothing parameter, a constant. Smaller N (e.g. 1, 10) emphasises high-frequency words, larger N (e.g. 100, 1000) puts more weight on medium- and low-frequency words.

This measure is widely used in the Sketch Engine (sketchengine.eu) and other corpus tools.

Keywords are easily applicable to the comparison of translations. A list of keywords for a pair of translations of the same source text reveals what is particularly different in these texts. The list shows what words the given translator loves to use, what words he/she uses more often than the other translator, and what words he/she is trying to get rid of. The length and the structure of the list can expose the extent of dependence of the new translation on the old one: whether the new translation is just an edited and corrected old translation, or the new translation is heavily based on the old translation, or whether the new translation is really new.

3. Comparing retranslations using corpus methods. The "Animal Farm" in Russian.

When comparing a group of objects, it is good to have a starting point, an ideal object fto compare with. It is very difficult to choose such a starting point for a group of translations of the same text. Each case is unique: some of older translations influence new translations, some are completely forgotten, some new translations are based on previous translations, some are performed without consulting any previous works. The quality might be improving in new translations, or it may remain the same, or even degrade. Therefore, choosing the first or the last translation as a starting point for comparison would not work in many cases.

The best starting point would have been the original text, but it is written in other language, and this makes comparison impossible. A literal translation of the original into the target language would have been a good solution, but this is too difficult technically, especially with long texts. A machine translation of the text is obviously the closest to a literal translation and can be used as a kind of 'projection' of the source text onto another language. This projection is not ideal: a machine-translated text contains lexical and grammatical errors, as well as wrong translation equivalents.

Still, it has certain strengths: it is standardised, neutral and easy to obtain. For these reasons I decided to add a machine-translated version to the human translations and use it as a baseline for comparison.

3.1. The research data.

As it has been already mentioned, this paper is devoted to analysis of the Russian translations of George Orwell's *Animal Farm* (1945). The original text was aligned at the sentence level with a machine translation and six human translations. The software used for aligning was LF Aligner (https://sourceforge.net/projects/aligner/). The aligned texts were parsed with universal dependencies grammar parsers (https://universaldependencies.org/). The corpus software I used for querying the data was TextHammer, a web-based corpus manager that I developed myself (puolukka.rd.tuni.fi/texthammer).

The machine translation was performed with Microsoft Translator via WordFast translation memory programme. The resulting translation in the form of aligned bitexts was exported to a TMX file that was easy to parse and upload to the corpus database.

The human translations that are studied in this mini-research are the translations by Struve and Kriger (1949), Pribylovskij (1986), Task (1988), Polotsk (1989), Kibirskij (1989) and Bespalova (1989) (see Appendix for details). According to FantLab website (https://fantlab.ru/work9633), there exist at least two more Russian translations of Orwell's story, but these were not available.

It is easy to notice that five of these six translations were made almost at the same time and it is quite possible that the translators even did not know that other new translations of the same book were being prepared. The translators of the 1980-ies might have been familiar with the first Russian translation by Struve & Kriger, which was published in West Germany by Posev publishing house and some copies were smuggled to the Soviet Union (the book was prohibited in the USSR until Perestroika), and might have been consulting it while translating. To check this, a search for matching segments in the translations was performed. For this purpose a PHP-script was developed by the author of this paper. The search shows that none of the later translations contain extensive borrowings from the first translation. The largest amount of closely matching sentences longer than two words with Dice index greater than 70% was found in the translations by Pribylovskij (101 sentences) and Kibirskij (108 sentences). The total number of sentences in the translations vary from 1733 to 1804, thus the percentage of such possible borrowings is very low, a little more than 5%. The translation by Polotsk contains much less matches – 60, and even less were found in the translations by Task (41) and Bespalova (27).

Among the matches found are sentences that could have been produced independently by the translators themselves, e.g. *Vse životnye ravny* 'All animals are equal'. However, some longer extracts do not look as mere coincidences.

1 No animal must ever live in a house, or sleep in a bed, or wear clothes, or drink alcohol, or smoke tobacco, or touch money, or engage in trade. (Orwell)

1a Не живите в домах, не спите на кроватях, не носите одежды, не пейте спиртного, не курите, не занимайтесь торговлей, не берите в руки денег. (Bespalova)

1b Ни одно животное не должно жить в доме, спать в кровати, носить одежду, пить алкогольные напитки, курить табак, прикасаться к деньгам, заниматься торговлей. (Kibirskij)

1c Ни одно из животных не должно жить в доме, спать в постели, носить одежду, пить алкоголь, курить табак, притрагиваться к деньгам или заниматься торговлей. (Polotsk)

1d Ни одно животное не должно жить в доме, спать в постели, носить одежду, не должно употреблять алкоголь и курить табак, заниматься торговлей и вести денежные расчеты. (Task)

1e Ни одно животное не должно жить в доме, спать в кровати, носить одежду, пить спиртное, курить, прикасаться к деньгам, торговать. (Pribylovskij)

1f Ни одно животное не должно жить в доме или спать в постели, носить одежду, пить спиртное, курить табак, прикасаться к деньгам или торговать. (Struve & Kriger)

In example (1) the translations 1b, 1c and 1e almost coincide with 1f. Theoretically, the sentence of the source text has simple and transparent structure and vocabulary, and therefore can stimulate similar translations, especially if translators use the same dictionaries and have similar background. However, the possible use of standard solution by the translators in the example (2) does not look very convincing.

2 As he had said, his voice was hoarse, but he sang well enough, and it was a stirring tune, something between 'Clementine' and 'La Cucaracha'. (Orwell)

2a Голос у него, и верно, был сиплый, но пел он неплохо. И мотив, помесь "Клементины" и "Кукарачи", брал за сердце. (Bespalova)

2b Голос его и вправду звучал сипло, но пел он довольно хорошо. Мотив был бодрый и волнующий (нечто среднее между мелодиями "Клементайна" и "Ла Кукарачча"). (Kibirskij)

2c Как он и говорил, голос у него был хриплый, но волнующая мелодия, нечто среднее между "Клементиной" и " Кукарачей " звучала достаточно чисто. (Polotsk)

2d Хотя голос у него и вправду был уже не тот, однако пел он довольно прилично, и мелодия сразу западала в сердце, напоминая одновременно "Клементину" и "Кукарачу". (Task)

2e Как он и предупреждал, голос у него был хриплый, но пел он совсем неплохо, и мотив у песни был бодрый, что-то среднее между "Клементиной" и "Кукарачей". (Pribylovskij)

2f Как он сам сказал, голос у него был хриплый, но пел он совсем не плохо, а мотив был бодрящий – нечто среднее между "Клементиной" и "Кукарачей". (Struve & Kriger)

The source sentence in the example (2) is longer and more complicated both lexically and syntactically, but still 2a, 2c and 2e match 2f, while 2a and 2d are different.

We can therefore suppose that Pribylovskij, Kibirskij and Polotsk used to some extent the translation by Struve and Kriger, while Task and Bespalova evidently worked only with the source text.

3.2. The visual impressions.

Unlike a normal corpus-based study, our research data is just seven versions of one relatively short text and it is therefore possible to cast a glance on the translations and to form an opinion on their quality.

3.2.1. The machine translation.

The machine translation is (quite expectedly) unacceptable for publishing, although one must admit that the level achieved by the programme in some passages is surprisingly good. In a short extract demonstrated in Table 1, only segment (2) is acceptable, albeit it is somewhat heavy stylistically. Other segments need minor or major corrections. Segment (1) contains a grammatically incorrect construction *byla prinjata ... rešenie* (the verbal clause is in feminine while the object is a neuter noun, the correct form: *bylo prinjato ... rešenie*). Segment (3) has a redundant *sozvali ih vmeste*, the verb *sozvat'* 'to call together' has a seme of gathering included into its meaning and thus the adverb *vmeste* 'together' only complicates and disbalances the sentence. Segment (4) contains constructions that are grammatically unacceptable: *eto polovina šestogo* (should be *sejčas polovina šestogo*) and *u nas est' dlinnyj den' pered nami* (correct version: *pered nami dlinnyj den'*). The punctuation of the sentence is not quite correct as well. In Russian, a direct speech is signalled with m-dashes and the punctuation in (4) should have been like this: *"Direct speech", – indirect speech, – "direct speech continued"*. In the segments (5) and (6), wrong lexemes are chosen: *urožaj* 'crops' (should be *uborka* 'gathering' or uborka *urožaja* 'gathering of crops') and *prinjat* 'decided' (should be *rešen* 'settled').

Table 1. A fragment of a bitext with machine translation.

	Source text (Orwell)	Machine translation (Microsoft Translator)
1	A unanimous resolution was passed on the spot that the farmhouse should be preserved as a museum.	На месте была принята единогласное решение о том, что фермерский дом должен быть сохранен как музей.
2	All were agreed that no animal must ever live there.	Все были согласны с тем, что ни одно животное никогда не должно жить там.
3	The animals had their breakfast, and then Snowball and Napoleon called them together again.	Звери позавтракали, а потом Снежок и Наполеон снова созвали их вместе.
4	"Comrades," said Snowball, "it is half-past six and we have a long day before us.	"Товарищи", сказал Снежок, "это половина шестого, и у нас есть длинный день перед нами.
5	Today we begin the hay harvest.	Сегодня мы начинаем урожай сена.
6	But there is another matter that must be attended to first."	Но есть еще один вопрос, который должен быть принят в первую очередь ".

As a whole, the machine translation is in most cases readable and understandable (although some passages sound comically), but it does not make a cohesive text: each segment is handled by the programme separately, without taking into account the information from previous segments. As result, proper names are not translated consistently, e.g. names of the characters like *Boxer* or *Clover* regularly turn into common nouns. Gender of the characters floats from masculine to feminine and back. The style does not meet the standards of a literary text, it is not consistent and

grades from official to colloquial. Also, when translating passages with complex syntax, the programme makes grammar errors and confuses equivalents.

Still, in spite of its insufficient quality and numerous errors, the machine translation closely follows the structure of the source text and makes no omissions, which makes it suitable for the role of a starting point for comparison of the human translations.

3.2.2. The human translations.

It was fairly difficult to evaluate the Russian translations by reading impressions. The earliest translation by Struve and Kriger looks rather old-fashioned, some words used in the translation have changed their meaning in the modern language (e.g. *šosse* in the meaning 'any road' which means in the modern Russian 'a motorway'), there are small omissions because the translators evidently used an earlier edition of the Orwell's work. The later translations are roughly on the same level and all have own strengths and weaknesses. The translation by Sergei Task and S. Kibirskij are more domesticated, the translation by Ilan Polotsk is more foreignized, anyway, none of the translations can be called clearly domesticated or clearly foreignised: all contain both trends, the translators are switching constantly between copying English syntax and rewriting some passages completely, transliterating some names and translating others, replacing English realia by Russian realia and preserving English realia, etc.

3 Remove Man from the scene, and the root cause of hunger and overwork is abolished for ever. (Orwell)

3a Если мы уберем человека, мы навеки покончим с голодом и непосильным трудом, ибо человек – их причина. (Bespalova)

3b Удалите Человека – и основная причина голода и рабского положения животных будет устранена навеки. (Kibirskij)

3c Уберите со сцены человека, и навсегда исчезнет причина голода и непосильного труда. (Polotsk)

3d Уберите с подмостков Истории человека, и вы навсегда покончите с голодом и рабским трудом. (Task)

3e Уберите Человека – и коренная причина голода и изнурительных трудов будет устранена навеки. (Pribylovskij)

3f Уберите Человека, и коренная причина голода и переутомления будет устранена навеки. (Struve & Kriger)

In the example (3) one can see how the same phrase is treated by different translators. As in the previous examples, a certain similarity can be traced between the translation by Struve & Kriger (3f), Pribylovskij (3e), and Kibirskij (3b). The translations by Kibirskij (3b), Polotsk (3c), Pribylovskij (3e) and Struve & Kriger (3f) follow the structure and even try to preserve some features of orthography (capitalised *Man*) and punctuation (comma before *and*) of the original. In contrast, Bespalova (3a) makes changes in syntax and makes the statement more explicit. Kibirskij (3b) changes *overwork* of the source text to *rabskoe položenie* 'slavery conditions'. Task (3e) preserves syntactic structure but makes radical semantical changes (e.g. *scene* → *podmostki Istorii* 'the stage of History').

Interestingly, in general the translations look after all surprisingly different. There are different ways of translating the title of the book (see Table 2), the names of the personages (e.g. *Snowball*: *Snežok* 'snowball' (Struve & Kriger, Pribylovskij, Kibirskij), *Ciceron* 'Cicero' (Task), *Obval* 'avalanche' (Bespalova), *Snouboll* (Polotsk)), other proper names (e.g. the *animalism* doctrine: *skotizm* 'cattle + ism' (Struve & Kriger, Bespalova), *zverizm* 'beast + ism' (Pribylovskij), *animalizm* (Kibirskij, Polotsk, Task)).

However, publishers seem to prefer two translations: by Larisa Bespalova (29 editions) and by Sergey Task (10 editions) (see Table 2). In spite of this fact, one cannot claim that the most popular Bespalova's translation has met all the standards of an ideal translation.

Visual comparing of different translations of the same text yields many interesting observations. Yet, it is not possible to obtain a general picture of the continuum made up of attempts to acquire the best translation. Only quantitative data may give a clue.

3.3. Descriptive statistics.

Let us have a look, if descriptive statistics can shed more light on the matter. In the Table 2 the information on the original text of the story and its translations can be found. I have provided four measures: number of words, number of characters, length of lemmatised word list and standardised type-token ratio per 1000 words (STTR, see explanation below).

Eugen Nida and Charles Taber claim that a translation will be always longer than an original text, because the translator has to make explicit many things that are evident to the readers of the source text (Nida & Taber 1974, 163). This heuristic is very difficult to confirm or refute, because different languages have their own ways of 'packing' the information: short words vs. long words, synthetic vs. analytical grammar forms, use of composite words, use of particles, use of articles, etc. In any case, the data from this case study does not confirm this heuristic. The number of words in the Russian translations is much less than in the original text, the main reason being that Russian is an articleless language. The difference in number of characters is smaller, even so the Russian translations are 'shorter' than the English original.

It follows from the Nida & Taber heuristic that a machine translation (that is generated mechanically and is incapable to make implicit information explicit) should be shorter than a human translation (that is created having in mind the background of the audience). Strangely, in our data the machine translation is much longer than any human translation (see Table 2). Probably, the reason is that human translators are able to find alternative ways of translating complicated constructions while the machine translator has no options but go the straight way and is therefore forced to use long and clumsy solutions. The lengths of the human translations vary between 122,690 characters (Task) and 134,519 characters (Polotsk), which confirms that the length of a

translation has something to do with preserving the structure of the source text in the translation and with smoothing of angles. As it has been mentioned above, the first is the most domesticated and the last – the most foreignised translation.

Table 2. Animal Farm and its translations.

Author (Translator)	Title	Year	Editions	Number of Words	Number of characters	Length of lemmatised word list	STTR
Orwell, George	Animal Farm	1945		30,437	138,269	3556	48.4
Microsoft Translator	Ferma životnyh	-		25,027	136,521	3859	64.39
Struve, Gleb & Kriger, Marina	Skotskij hutor	1949	2	22,746	125,839	4280	67.76
Pribylovskij V.	Ferma životnyh	1986	3	23,729	133,657	4830	69.65
Task, Sergey	Skotskij ugolok	1988	10	21,399	122,690	5038	73.33
Bespalova, Larisa	Skotnyj dvor	1989	29	23,004	125,250	4655	67.49
Kibirskij S.	Ferma životnyh	1989	2	23,085	131,400	4472	70.3
Polotsk, Ilan	Skotskij hutor	1989	1	23,879	134,518	4319	67.42

The STTR index (the mean of the ratio of number of unique words (types) to number of different words (tokens) calculated for fix-length extracts, e.g. 1000 words, see NN 2016: 116-121 for the detailed explanation) shows diversity of vocabulary and thus reflects repetitiveness, readability, and lexical richness. Texts with low STTR are more simple, more straightforward, easier to read, but dull and repetitive, while texts with high STTR are more compact, less repetitive, more attractive, but more difficult to read. STTR values for different languages are different, therefore the lower value of STTR of the original English text in Table 2 does not mean that the Russian translations are 'more beautiful'. After comparing STTR values of the translations we can make an interesting observation: the machine translation has much lower STTR than the human translations. This is what we could have expected. The STTR values of the human translations vary from 67.49 (Bespalova) to 73.33 (Task). Strangely, these two translations with the extreme values of STTR are also the most often published (see 3.2). Anyway, the difference in STTR values of human translations is not significant.

To sum up the findings, the numeric data reveals much more variation that might be expected from the translations of the same work. Still, no conclusions can be drawn yet from these number without more sophisticated data processing.

3.4. Studying retranslations with distance measure.

One of the ways to get deeper into the matter would be measuring distances between the translations as it was described in 2.1. To obtain the data for the distance measuring, I generated lemmatised frequency lists of all seven texts (machine translation and six human translations), than loaded the lists in R.[i] As it was already mentioned in 2.1, comparing complete lists is not very effective even for related texts, therefore the lists were truncated to the 100 most frequent words.

The truncated lists were merged into a single table, the table was rotated and the final dataset was a data frame with data on texts in rows and words in columns (see a fragment in Table 3). The frequencies were normalised to items per thousand.[ii] Full outer join was used for merging of the tables, i.e. items that did not occur in all frequency lists were also copied to the new table. The size of the resulting data frame was 7 X 155.

Table 3. A fragment from the joint frequency table.

	а: CCONJ	боец: NOUN	бой: NOUN	больше: ADV	большой: ADJ	бы: PART	быть: AUX	быть: VERB
Bespalova	9.61	4.48	1.39	0	0	3.43	8.87	2.39
Kibirskij	5.93	0	0	1.52	1.6	1.91	18.45	2.6
MT	2.64	0	0	0	2.24	2.56	37.2	3.32
Polotsk	4.4	0	0	0	0	1.93	19.81	2.68
Pribylovskij	7.33	0	0	0	1.39	2.36	15	3.41
Struve & Kriger	5.14	0	0	0	0	2.42	23.17	3.3
Task	6.87	0	0	0	0	2.66	12.9	2.29

On the next stage, I generated a distance matrix and performed on it the multidimensional scaling (MDS). The software used were R Studio and its packages *cluster* and *smacof*. The MDS analysis worked well, the stress value was 1.3, which generally means that the fit of the two-dimensional model is good. The resulting visualisation can be seen on Fig. 1. The texts are placed into a two-dimensional space, the geometrical distances between the dots reflect the differences between the frequencies of their MFWs. The lowest value on the x-axis get the machine translation, the highest – the translations by Bespalova and Task, which, as it was already mentioned, are also the most popular (see 3.2).

Probably, the x-axis is related with preserving the structure of the original text: the extreme case is the machine translation with minimal changes, while the translations by Bespalova and Task contain changes in structure, possible omissions and explicitations for making the text more transparent and more readable.

As it was already mentioned in 3.2, the visual inspection of the texts left the impression that Bespalova tried to adapt the structure of the translation to the norms of Russian style, while Task's translation is very free and changes on semantic level happen fairly often. It is possible therefore that the y-axis shows the gradation from grammatical changes to lexical changes.

The human translation closest to MT both on x and y axes is the translation by Struve & Kriger. The position on the x-axis shows that this is the most literal translation in our group, and the position on the y-axis means that the structure of the source text is mostly preserved and some minor lexical changes may be found.

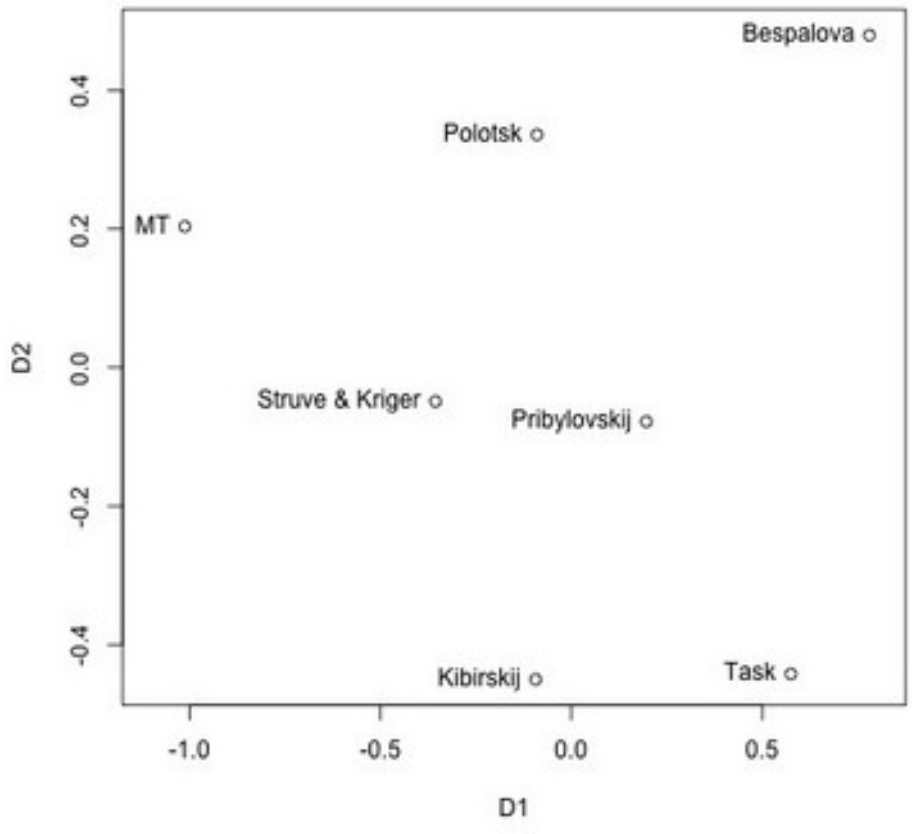

Fig. 1. Distance measure visualisation (MDS)

3.4. Keywords analysis of retranslations.

Keywords help to figure out, what are the items that are specific for the research data. Keywords are not an absolute concept, they are always relative to the reference data. When performing searches for keywords, it is vitally important to choose a relevant reference data set, the resulting list largely depends on this choice. The selection of the reference corpus should be made keeping in mind the objectives of the study. When comparing the research data against a large corpus of general language, the keyword list would tell about the topic and the text type, when the reference corpus are texts of the same text type, the keywords would show the specifics of the subgenre or of the style of the author.

In this study, the task is to find out, how the translations of the same work differ one from another. Using a large corpus of language for general purposes as reference data would not provide an answer to the question: the lists obtained from retranslations would be close to each other and the differences would be blurred. Using one of the translations should be more effective, but the question is which translation to take. Fortunately, we have a machine translation, which is the most

literal and has neither omissions nor additions. So, we can use the machine translation as the reference data.

Keyword search was run on lemmatised word lists using Kilgarriff's simple measure with N=100 (see 2.2.) and the items with K ≥ 2 were taken. Proper names were removed from the results. A fragment of one of the keyword lists can be seen in Table 3.

3.4.1. The keyword lists: general information.

One would expect that the lists obtained this way would be very close. This did not happen however. Only four(!) items are present in all six keyword lists: *vot* 'this is, interj.', *zerno* 'grain, n', *korovnik* 'cowshed, n', *tvoj* 'your, pron. 2 pers. Sg'.

Keyword lists often help to find interesting lexical items in the data and observe its specific features. Indeed, it is easy to notice that the keyword list of the translation by Struve and Kriger contains many archaic words like *ibo* 'because', *glasit'* 'to announce', *javit'sja* 'to appear'. As for other translations, their keyword lists do not seem to contain anything special. Let us try therefore to study those lists as a whole.

The shortest list was the list of keywords for the translation by Struve & Kriger (57 items) and the longest were the lists for the translations by Task (118), Polotsk (84) and Bespalova (83) (see Table 3). Thus the keyword search goes in line the results of the MDS analysis of the frequency lists carried out in 3.3.

Table 3. A fragment of the keyword list from the translation by Struve & Kriger.

Token	Experimental data, ipm	Reference data, ipm	K
сражение 'battle'	923.00	79.00	5.72
работник 'worker'	483.00	39.00	4.19
всякий 'any'	967.00	159.00	4.12
коровник 'cowshed'	615.00	79.00	3.99
прежний 'of the previous times'	439.00	39.00	3.88
дерево 'tree'	527.00	79.00	3.50
заказ 'order'	659.00	119.00	3.47
уметь 'be able'	483.00	79.00	3.26
употреблять 'to use'	351.00	39.00	3.24

We can anticipate that the more differences are between experimental and reference data, the larger keyness values (i.e. values of K) they would have. The keyness values for the top ten keywords in our lists differ from 5.72 to 3.24 (Struve & Kriger), 6.56 to 3.42 (Kibirskij), 9.44 to 3.42 (Polotsk), 7.87 to 2.83 (Pribylovskij), 12.65 to 4.16 (Bespalova), 7.71 to 4.21 (Task). Hence, the keywords

from the translation by Struve & Kriger has the lowest keyness values, while the values of those from the translations by Bespalova and Task are the highest.

3.4.2. Cohesion words

The weak point of MT is text cohesion. A machine translator handles pronouns, conjunctions and particles of the source text in the same way as any other items (nouns, verbs, adjectives), while human translators obviously choose first a suitable syntactic construction and use the grammatical markers it needs: for compound sentence a conjunction would be needed, a nominalised clause would do with a preposition. That is why all six keyword lists contain pronouns, conjunctions, particles and prepositions. The amount of such items signals about the extent of adaptation of syntax to the norms of the target language. The study of keyword lists shows that the list from the translation of Struve & Kriger contains least cohesion words. The lists of keywords from the translations by Polotsk, Kibirskij and Pribylovskij also do not contain many words of these kinds, while among keyword lists from the translations by Bespalova and especially by Task many pronouns, particles and even prepositions can be found (see Table 4).

Table 4. Part of speech statistics in the keyword lists.

Tokens	Bespalova	Kibirskij	Polotsk	Pribylovskij	Struve & Kriger	Task
a	4	10	7	9	4	12
adv	10	9	9	6	6	11
conj	1	1	0	1	1	1
n	26	23	28	23	16	47
particles	9	7	4	6	4	9
prep	3	2	2	2	0	3
pron	7	3	6	6	4	9
v	23	22	28	14	22	26
Total	**83**	**77**	**84**	**67**	**57**	**118**

Emphasizing particle *vot* 'here is' can be found in all keywords lists. English has much less particles than Russian and uses different means for highlighting elements of the text. Therefore *vot* occurs only three times in the machine translation and is quite frequent in the human translations, especially in the translations by Bespalova (61 occurrences) and Task (34 occurrences).

4 Man is the only real enemy we have. (Orwell)
4a Человек – вот кто наш истинный враг. (Bespalova)
4b Человек – наш единственный подлинный враг. (Kibirskij)
4c Вот кто наш единственный подлинный враг – человек. (Polotsk)
4d Вот он, корень зла – человек. Другого врага у нас нет. (Task)
4e Человек – вот наш единственный подлинный враг. (Pribylovskij)
4f Человек – вот наш единственный настоящий враг. (Struve & Kriger)
4e Человек – единственный настоящий враг, который у нас есть. (MT)

In example (4) only Kibirskij does not use *vot* in the translation. Obviously without it the statement loses energy. The machine translation (4e) is grammatically and lexically correct and is the closest variant of translation, still, no one of human translators chosen this way of translating this sentence.

The particle *vot* is obviously a favourite of Bespalova, the relative frequency of the word in her translation is 2651.71 ipm which is much higher than in the Russian National Corpus (1785.1 ipm, see Lâševskaâ & Šarov 2009).

5 The words ran: (Orwell)
5a Вот эта песня: (Bespalova)
5b Слова же были такие: (Kibirskij)
5c Слова были таковы: (Polotsk)
5d Впрочем, ближе к тексту: (Task)
5e Слова были таковы: (Pribylovskij)
5f Слова же были следующие: (Struve & Kriger)
5e Слова побежал: (MT)

In the example (5) only Bespalova uses *vot* in translation. The original text looks very simple, although the direct translation suggested by MT (5e) is impossible, and not only because of disagreement of the noun and the verb: the verb *pobežat'* 'to start running' is never used in the meaning 'to express by means of language'. Still, only two translators, Polotsk (5c) and Pribylovskij (5e) do not use any particles in their translations. Bespalova uses *vot*, Kibirskij (5b) and Struve & Kriger (5f) use another particle *že*, Task (5d) starts the sentence with a modal word *vpročem* 'however'.

These examples show that the cohesion words are in most cases added by translator and not transferred from the original text. Use of these markers belongs to individual style.

3.4.3. Nouns and verbs.

The faithfulness of the translations to the original correlates with the numbers of nouns in the lists of keywords: the smallest number is in the translation by Struve & Kriger, the largest is in the translation by Task (see Table 4). An interesting feature of the keyword lists of the translations by Task and by Pribylovskij is that the number of nouns is almost two times bigger that that of verbs, while for the remaining four translations number of nouns and verbs in the keyword lists is almost equal. Where do all these nouns come from?

One of the most frequent noun of the keyword list from the translation by Task is *massa* 'mass': it occurs only twice in the MT and the translation by Polotsk, only once in the translation by Kibirskij and is never used in other translations. Task uses the word *massa* in two meanings: 'a large group of people' and 'physical mass'. The first usage (in plural) is very typical for marxist literature (at least in Russian), and Task uses the word to link the Animal Farm to a socialist state. However, Orwell

does not use the word *mass* in this meaning, therefore Task's use of the word *massa* signals changes in the contents of the source text.

6 These two had great difficulty in thinking anything out for themselves, but having once accepted the pigs as their teachers, they absorbed everything that they were told, and passed it on to the other animals by simple arguments. (Orwell)

6a Они ничего не могли придумать самостоятельно, но, раз и навсегда признав свиней своими учителями, буквально впитывали каждое их слово и доходчиво передавали другим животным. (Bespalova)

6b Им обоим было трудно самостоятельно всё продумать, но, однажды признав свиней своими учителями, они впитывали всё, что им говорилось, и затем передавали это простыми словами другим животным. (Kibirskij)

6c Сам процесс мышления доставлял им немалые трудности, но раз и навсегда признав свиней своими пастырями, Кловер и Боксер впитывали в себя все, что было ими сказано и затем терпеливо втолковывали это остальным животным. (Polotsk)

6d Эти двое были не способны дойти до чего-либо своим умом, но после того как им все разжевали, они стали самыми надежными проводниками свинских идей; они внедряли их в сознание масс с помощью простейших формулировок. (Task)

6e Они с огромным трудом могли бы что-нибудь придумать сами, но признав однажды свиней своими учителями, они принимали все, что те говорили, на веру и в доходчивых выражениях объясняли это другим животным. (Pribylovskij)

6f Этим двум не легко было мыслить самим, но, раз уже признав свиней своими учителями, они усваивали все, что им говорили, и с помощью простых доводов передавали усвоенное другим животным. (Struve & Kriger)

In example (6) Task replaces *other animals* by *massy* 'masses', while other translations just keep to the original using direct equivalents: *drugie/ostal'nye životnye* 'other/remaining animals'.

7 Napoleon sent for pots of black and white paint and led the way down to the five-barred gate that gave on to the main road. (Orwell)

7a Наполеон распорядился принести по банке черной и белой краски и повел их к тесовым воротам, выходящим на большак. (Bespalova)

7b Наполеон послал за черной и белой краской и подвел всех к тяжелым воротам, отгораживавшим ферму от дороги. (Kibirskij)

7c Наполеон послал за банками с черной и белой красками и направился к воротам, за которыми начиналась основная дорога. (Polotsk)

7d Наполеон послал Делового за масляной краской, черной и белой, а сам повел массы к главным воротам. (Task)

7e Наполеон послал за ведрами с черной и белой краской и повел всех вниз, к выходившим на главную дорогу воротам, которые были окованы пятью железными скрепами. (Pribylovskij)

7f Наполеон послал за банками черной и белой краски и провел животных к калитке, которая выходила на шоссе. (Struve & Kriger)

In example (7) the word *massy* appears in Task's translation out of nowhere: in the original text the object is implicit, *led the way*. Struve & Kriger (7f) explicitate it as *provel životnyh* 'led the animals', Polotsk (7c) remove the object completely changing the verb to *napravilsja* 'went', other translators explicitate the object by adding a pronoun: *povel vseh* 'led all'. Task changes the neutral source text to pathetic *povel massy* 'led the masses'.

8 Without halting for an instant, Snowball flung his fifteen stone against Jones 's legs. (Orwell)

8a Но Обвал не дрогнул и всей своей шестипудовой тушей двинул Джонса по ногам. (Bespalova)

8b Ни на мгновение не замедляя бега, Снежок врезался всеми своими девяноста пятью килограммами в колени врага. (Kibirskij)

8c Ни на мгновенье не останавливаясь, Сноуболл всем своим внушительным весом сбил Джонса с ног. (Polotsk)

8d Не сбавляя хода, Цицерон всей своей массой врезал Джонсу по ногам. (Task)

8e Ни на секунду не останавливаясь, Снежок налетел на фермера и все шесть пудов своего веса бросил ему под ноги. (Pribylovskij)

8f Не останавливаясь ни на секунду, Снежок бросил свою пятипудовую тушу под коленки Джонсу. (Struve & Kriger)

In example (8) Task uses the word *massa* in the different meaning, and it is used as equivalent for a culture-specific word *stone* (an English unit of weight). None of the translators try to preserve this word in the translation. Bespalova (8a), Pribylovskij (8e) and Struve & Kriger convert stones into *puds* (pud is a traditional Russian weight measure, 16 kg), Kibirskij (8b) converts stones into kilograms, others just mention that Snowball was heavy: *svoim vnušitel'nym vesom* 'with his impressive weight' (Polotsk, 8c), *vsej svoej massoj* 'with all his mass' (Task, 8d).

No great variety can be observed in the numbers of verbs in the lists of keywords: all lists have between 20 and 30 verbs with the exception of Pribylovskij's translation, which has only fourteen. The verbs from the keyword list of Pribylovskij are more abstract than the verbs from other five lists. No verbs of action can be found, there are verbs of social activities (*rukovodit'* 'to manage', *gotovit'* 'to prepare', *sledit'* 'to spy'), verbs of speech (*uverjat'* 'to assure', *priznavat'sja* 'to confess'), verbs of state (*otnosit'sja* 'to belong', *predstojat'* 'to expect'). In contrast, the verbs of the keyword lists from other translations are not only more numerous, but also more diverse, e.g. in Bespalova's keyword list there are verbs of action (*snesti* 'to pull down', *razrušit'* 'to destroy'), movement (*obojti* 'to walk around', *podnjat'sja* 'to go up'), mental activity (*podumat'* 'to think', *ponimat'* 'to understand', *sčitat'* 'to suggest').

Interestingly, the proportion of nouns and verbs in the keyword list from Pribylovskij's translation is the same as in the translation by Task, but the numbers are smaller. The keyword lists show therefore that both Pribylovskij's and Task's translations are more 'static' than other translations, i.e. they pay more attention to objects than to actions.

4. Conclusions

The corpus-based analysis of the six translations of Orwell's *Animal Farm* made it possible to detect relations between different translation, measure the distances between them and even find some peculiarities of individual translations.

The frequency-list-based comparison of the texts proved to be very efficient and the multidimensional scaling method makes it possible to visualise relations between the texts. Using a machine translation of the source text as a 'starting point' seems to work well, it may be an alternative to a literal translation. The poor quality of MT may skew to some extent the results, but using one of human translations for the purpose would be much worse.

The keyword analysis gives additional data to the research and it confirms the findings of the MDS analysis. It also makes it possible to find lexical classes, parts of speech, or certain lexemes that may yield additional data. The study of keyword lists and the statistics drawn from these helps a researcher to get an idea what to look for.

The most interesting result of this mini-research is that, at least in this concrete case, the first translation was not the most domesticated, as it should be according to the retranslation hypothesis. The first translation was the most literal and the most close to the original text. The later translations are less literal and pursue readability and naturalness of the language of translation. The publishers tend to choose for publishing the translations that are written in more natural language, and this (sadly) means that the faithfulness to the original and the quest for giving the most exact picture of the original work does not interest the publishers and, most likely, the readers as well.

What is important is that the methods presented in this paper are applicable not only to prose, they work with translations of poetry and drama as well. The results of the analysis of concrete empirical data can also aid in the fields of language technologies, plagiarism detection and other disciplines that study similarities in texts.

The development of corpus-based methods to study retranslations can be of great use for translation studies in that they offer quantitative measures for comparing translations and their quality evaluation. It becomes possible to manage very large sets of data, to study large works that were translated many times.

In this particular study the alignments were used for concordancing purposes only. The distance measure and keyword searches could have been performed on unaligned texts as well. However, possessing aligned parallel texts opens many other possibilities for research and for comparing parallel texts: omissions or additions can be discovered, use of certain translation equivalents can be mapped, etc. Indeed, a large parallel corpus of retranslations aligned on sentence or even on word level would be of great use. Sadly, aligning multiple translations is still very difficult technically, the standard aligning software was not developed for such ambitious tasks, they are made for aligning pairs of technical manuals, agreements, or other documents written in standardised language and with clear structure. Aligning literary texts is much more difficult, even pairwise. It was possible to align eight parallel texts of *Animal Farm* with LF Aligner, but aligning forty translations of one fragment from Macbeth was very demanding task (see Cheesman et al 2017: 744). Huge parallel corpora of *the Bible* became technically possible only because the verses of *the Bible* are numbered and thus the aligning had been already performed manually a long time ago and by other people.

References

Alharbi, Mohammad, Robert S Laramee, and Tom Cheesman (2015) TransVis: Integrated Distant and Close Reading of Othello Translations. JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, 1-18, https://doi.org/10.1109/TVCG.2020.3012778

Brownlie, Siobhan. 2006. Narrative Theory and Retranslation Theory. *Across Languages and Cultures*. 7:2, 145-170.

Cadera, Susanne M. and Andrew Samuel Walsh (eds.) 2017. *Literary Retranslation in Context.* New Trends in Translation Studies. Oxford, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien, Peter Lang.

Čermakova A. & Farova L. 2010. Keywords in Harry Potter and their Czech and Finnish translation equivalents. In Čermak F. et al (eds.). *InterCorp: Exploring a Multilingual Corpus.* Prague: Nakladatelství Lidové Noviny/Czech National Corpus, pp 177-188.

Cheesman, Tom, Kevin Flanagan, Stephan Thiel, Jan Rybicki, Robert S. Laramee, Jonathan Hope, Avraham Roos (2017) Multi-Retranslation corpora: Visibility, variation, value, and virtue. *Digital Scholarship in the Humanities*, Volume 32, Issue 4, 739–760, https://doi.org/10.1093/llc/fqw027.

Desmidt, Isabelle. 2009. (Re)translation Revisited. *Meta*. 54:4, 669-683.

Deane-Cox S. (2014). *Retranslation: Literature and Reinterpretation.* London: Bloomsbury.

Eder, Maciej, Jan Rybicki and Mike Kestemont. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, 16 (1), 107-121.

Fidler M. and Cvrček V. 2015. A Data-Driven Analysis of Reader Viewpoints: Reconstructing the Historical Reader Using Keyword Analysis. *Journal of Slavic Linguistics.* 23(2), pp. 197–239.

Jones, Henry. Retranslating Thucydides as a scientific historian. A corpus-based analysis. *Target* 32:1, 59–82. doi.org/10.1075/target.19082.jon

Johnson, S. & Esslin A. 2006. Language in the news: Some reflections on keyword analysis using Wordsmith Tools and the BNC. *Leeds Working Papers in Linguistics*, # 11. <http://www.leeds.ac.uk/arts/info/125154/leeds_working_papers_in_ linguistics_and_phonetics/1949/volume_11_2006>

Kemppanen H. 2004. Keywords and Ideology in Translated History Texts: A Corpus-based Analysis. *Across Languages and Cultures* 5 (1), 89-106.

Kemppanen, Hannu. 2008. *Avainsanoja ja ideologiaa: käännettyjen ja ei-käännettyjen historiatekstien korpuslingvistinen analyysi*. Joensuu: University of Joensuu.

Kilgarriff, Adam, 1997. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In Information Technology Research Institute Technical Report Series. 97-07. <http://aclweb.org/anthology/W97-0122>.

Kilgarriff, Adam, 2001. Comparing corpora. *International Journal of Corpus Linguistics*. 6(1), pp. 97–133. <https://www.sketchengine.eu/wp-content/uploads/comparing_corpora_ 2001.pdf>

Kilgarriff, Adam. 2009. Simple maths for keywords. In Mahlberg, M. et al (eds). *Proceedings of Corpus Linguistics Conference CL2009*, University of Liverpool, UK.

Koskinen, Kaisa ja Outi Paloposki 2015. *Sata kirjaa, tuhat suomennosta: Kaunokirjallisuuden uudelleenkääntäminen*. SKS.

Lutsenko E.M. 2019. Perevodčeskoe fiasko K.D. Bal′monta černovaâ redakciâ "Romeo i Džul′etty" U. Šekspira. *Šagi / Steps*. 5:3, 84-103. DOI: 10.22394/2412-9410-2019-5-3-84-103

Kuusi Päivi, 2014. Kääntämisen universaaleja uudelleenkäännöksissä. *MikaEL*, vol 4. https://sktl-fi-bin.directo.fi/@Bin/0ee65f8d4916a5370ceb8374ebf00d7e/1419945515/application/pdf/533414/Kuusi_MikaEL2014.pdf.

Lâševskaâ, O. N., Šarov S. A. 2009. *Častotnyj slovar' sovremennogo russkogo âzyka* (na materialah Nacional′nogo korpusa russkogo âzyka). Moskva : Azbukovnik. http://dict.ruslang.ru.

Mayer, Thomas & Michael Cysouw 2014. Creating a Massively Parallel Bible Corpus. In *LREC 2014, Ninth International Conference on Language Resources and Evaluation.* Reykjavik, Iceland, 3158-3163, www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf

McCarthy Arya D., Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, David Yarowsky 2020. The Johns Hopkins University Bible Corpus: 1600+ Tongues for Typological Exploration. In *Proceedings of The 12th Language Resources and Evaluation Conference*, Marceille, France, 2884–2892. https://www.aclweb.org/anthology/2020.lrec-1.352.pdf

Mikhailov, Mikhail. 2019. The Extent of Similarity: comparing texts by their frequency lists. In Jantunen, Jarmo Harri et al (eds.) Proceedings of the Research Data and Humanities (RDHum) 2019 Conference: Data, Methods And Tools. Oulu: University of Oulu, 159-178.

Milizia D., 2010. "Keywords and phrases in political speeches." In M. Bondi & M. Scott (eds.) Keyness in Text. Amsterdam & Philadelphia: John Benjamins, pp. 127-145.

Munday, Jeremy 1998. "A Computer-assisted Approach to the Analysis of Translation Shifts". *Meta*, Volume 43, Issue 4, 542–556.

Nida, Eugene A., and Charles R. Taber. 1974. *The theory and practice of translation.* Leiden: Published for the United Bible Societies by E.J. Brill.

NN 2016.

NN 2019.

Paloposki, Outi, Koskinen, Kaisa. 2010. Reprocessing texts. The fine line between retranslating and revising. *Across Languages and Cultures*. 11:1, 29-49.

Piperski Aleksandr, 2017. Sravnenie korpusov meroj χ^2: simvoly, slova, lemmy ili časterečnye pomety? [Comparing corpora with χ^2: characters, words, lemmata, or PoS tags?], In *Korpusnaja lingvistika–2017* [Corpus Linguistics–2017]. Saint Petersburg, Saint Petersburg State University, pp. 282–286.

Piperski, Aleksandr, 2018. Corpus size and the robustness of measures of corpus distance. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018"*. Moscow, May 30—June 2, 2018. <http://www.dialog-21.ru/media/4327/piperskiach.pdf>

Pouke van, Piet and Guillermo Sanz Gallego 2019. Retranslation in Context. *Cadernos de Tradução* 39(1):10-22.

Scott M. & Tribble C. 2006. Textual Patterns: Key Words and Corpus Analysis in Language Education. Amsterdam: John Benjamins.

Seale C, Charteris-Black J, Ziebland S. 2006. Gender, cancer experience and internet use: a comparative keyword analysis of interviews and online cancer support groups. Social Science and Medicine, 62, 10: 2577-2590.

Susam-Sarajeva, Şebnem. 2003. Multiple-entry visa to travelling theory: Retranslations of literary and cultural theories. *Target*: Volume 15, Number 1: 1–36.

Venuti, L. 1995. *The Translator's Invisibility:A history of translation*. London & New York: Routledge.

Wilkinson, M. 2014. Using the Keyword Tool to Explore Lexical Differences between British and American English in Specialised Corpora. *CALL-EJ* 15(1), 21-38. <http://callej.org/journal/15-1/Wilkinson_2014.pdf>

Appendix. Research data

Orwell: Orwell, George. *Animal Farm*, 1945.

Bespalova: Оруэлл, Джордж. *Скотный двор*. Пер. с англ. Беспалова, Лариса, 1989.

Kibirskij: Оруэлл, Джордж. *Ферма животных*. Пер. с англ. Кибирский С., 1989.

Polotsk: Оруэлл, Джордж. *Скотский хутор*. Пер. с англ. Полоцк, Илан, 1989.

Task: Оруэлл, Джордж. *Скотский уголок*. Пер. с англ. Таск, Сергей, 1988.

Pribylovskij: Оруэлл, Джордж. *Ферма животных*. Пер. с англ. Прибыловский В., 1986.

Struve & Kriger: Оруэлл, Джордж. *Скотский хутор*. Пер. с англ. Струве, Глеб и Кригер, Марина, 1949.

i The analysis could have been done with the *stylo* package for R (Eder et al 2016), but the package does not have special support for Russian. The processing would have been done without lemmatisation. For this reason it was decided to process word lists generated from the corpus with TextHammer.

ii It would have been possible to use absolute frequencies, since the texts are translations of the same source text and do not differ much in length. However, normalised frequencies give a better picture.