Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2018”

Moscow, May 30—June 2, 2018

CORPUS SIZE AND THE ROBUSTNESS OF MEASURES OF CORPUS DISTANCE[1]

Piperski A. Ch. (apiperski@gmail.com)
Russian State University for the Humanities / National Research University Higher School of Economics, Moscow, Russia

This paper studies the impact corpus size has on the robustness of various frequency-based measures of corpus distance (or similarity, respectively), such as Euclidean distance, Manhattan distance, Cosine distance, χ^2, Spearman’s ρ, and Simple-Maths Keyword distance. An experiment performed using the British National Corpus shows that Euclidean distance is least influenced by corpus size and thus is best suited for the purpose of comparing corpora.

Keywords: corpus similarity, distance between corpora, evaluation, British National Corpus

РАЗМЕР КОРПУСА И УСТОЙЧИВОСТЬ МЕР РАССТОЯНИЯ МЕЖДУ КОРПУСАМИ

Пиперски А. Ч. (apiperski@gmail.com)
РГГУ / НИУ ВШЭ, Москва, Россия

В статье рассматривается вопрос о том, какое влияние размер корпуса оказывает на устойчивость различных мер сравнения корпусов на основе частотных списков. Для анализа берутся шесть мер: Евклидово расстояние, манхэттенское расстояние, косинусное расстояние,

[1] This work is supported by the Russian Science Foundation under grant 17-78-10196.

χ^2, ρ Спирмена и сходство по ключевым словам. Эксперимент на материале Британского национального корпуса продемонстрировал, что Евклидово расстояние наименее подвержено влиянию размера корпуса и поэтому лучше всего подходят для сравнения корпусов.

Ключевые слова: сходство корпусов, расстояние между корпусами, эвалюация, Британский национальный корпус

1. Introduction

The problem of text and corpus similarity is extremely important for Natural Language Processing as well as for corpus linguistics[2]. Measuring similarity (or distance, respectively) is used for information retrieval, text classification, document clustering, machine translation evaluation, and many other applications. A survey of text similarity measures by [Gomaa & Fahmy 2013] includes two types of measures: character-based and term-based measures. Character-based measures treat texts (or text corpora) as sequences of characters which can be transformed into each other by using allowed edit operations, by finding an optimal alignment of two strings, etc. The best-known character-based measure of text similarity is the Levenshtein distance. However, character-based similarity measures imply that similar texts are obtained from each other by some simple operations, which is true in case of shorter texts like misspelled words deriving from the intended correct ones, but it does not conform to our intuition as to how longer texts are produced. For this reason, term-based measures, which can also be called frequency-based measures, are more widely used for measuring similarity between longer texts. To apply these measures, texts are represented as frequency lists, which are then being compared using various ways of measuring distances between vectors, the best-known of them being geometric measures such as Manhattan distance, Euclidean distance, and Cosine simlarity (or distance, respectively), and set-theoretic measures such as Jackard distance.

In corpus linguistics, the problem of text and corpus similarity has gained attention starting with [Kilgarriff 1997] and [Kilgarriff and Rose 1998]. Since then, many measures of corpus similarity have been proposed (see [Kilgarriff 2001, 2009]; [Fothergill et al. 2016]). However, no definitive measure for corpus similarity has yet been found. Specific approaches to computing similarity have been adopted in machine translation evalution, such as BLEU [Papineni 2002], NIST [Doddington 2002], and METEOR [Lavie and Agarwal 2007]. In most other applications that make use of measuring distances between texts, there is no measure that has become a de facto standard, though geometrical measures are generally preferred.

[2] In this paper, I refrain from pursuing the question whether we should using different measures of similarity for individual texts and for collections of texts, i.e., corpora.

2. Measures of corpus distance

In this paper, I will discuss six measures of distance between corpora. Three of them are based on geometrical notions, namely Euclidean distance, Manhattan distance, and Cosine distance. Two further measures are closely linked to the established statistical procedures; these two measures are χ^2 and Spearman's ρ, which were especially popularized by [Kilgarriff 2001], who showed that they are by far superior to perplexity-based measures. A further measure is Simple-Maths Keyword distance, introduced by [Kilgarriff 2009] and implemented in the Sketch Engine corpus management system ([Kilgarriff et al. 2014]; http://the.sketchengine.co.uk). In this paper, all measures are computed based on the frequencies of 200 most common words in the aggregated frequency distribution of the two corpora being compared (for the first five measures), or on the keyness score of the top 200 keywords in the aggregated keyword list for Simple-Maths Keyword distance.

3. Corpus size and corpus distance

The measures of corpus distance are typically evaluated ussing the Known-Similarity Corpora (KSC) approach [Kilgarriff 1997, 2001]; [Kilgarriff and Rose 1998][3]. A KSC-set is built starting with two corpora X and Y that are deemed to be sufficiently distinct. These original corpora are split into equal-sized chunks that are randomly allocated to new corpora Z_0, Z_1, …, Z_M, each of them consisting of M chunks. Z_0 includes 0 chunks from X and M chunks from Y, Z_1 includes 1 chunk from X and M—1 chunks from Y, Z_2 includes 2 chunks from X and M—2 chunks from Y, etc. The similarity of these corpora is known: for instance, one can assume that Z_3 is closer to Z_5 than Z_2 is to Z_8. Thus, for any $k \leq l < m \leq n$ ($k \neq l$ or $m \neq n$) a good distance measure must satisfy the inequality $d(Z_l, Z_m) < d(Z_k, Z_n)$. One can test whether such an inequality is satisfied for all possible values of k, l, m, and n, and the proportion of inequalities captured correctly indicates how well a distance measure performs. In case of $M = 10$, a total of 660 KSC judgments of the kind need to be tested.

More recent studies have continued this approach, using a wider range of measures and larger amounts of test data [Piperski 2017a, 2017b]. However, no investigations have yet adressed the question of how corpus size influences the robustness of distance measures.

This is a question that plays a significant role in many areas of corpus linguistics. Namely, we can trust a measure of distance only if it yields comparable results when comparing samples from the same populations regardless of sample size or, at least, starting from a certain corpus size. Otherwise, the results might turn out to be untrustworthyб especially when different-sized corpora are being compared. For this reason, the aim of the present study is to test the robustness of the six measures listed in Section 2 with respect to corpus size. Even though it was shown by [Piperski 2017a] that levels of analysis and segmentation other than individual words, first and foremost character ngrams, are better suited for assessing distance between corpora,

[3] Another approach to this problem is [Forsyth ad Sharoff's 2014] anchor-text method.

in the present study I stick to the word level, since a word is the largest unit that seems to be more or less easily identifiable in a text as well as linguistically significant.

4. Experiment design

For the purpose of the experiment, 200,000-token subcorpora from 11 sources from the British National Corpus (BNC) were taken. The sources are listed in Table 1:

Table 1. List of sources

ID	Source	BNC file IDs
art	The Art Newspaper	CKT–CKY, EBS–EBX
bel	The Belfast Telegraph	HJ3–HJ4, K29–K35
bio	The Dictionary of National Biography: Missing persons	GSX–GTH
han	Hansard Extracts	HHV–HHX
ind	The Independent	A1D–A5X
kee	Keesings Contemporary Archives	HKP–HLT
law	The Weekly Law Reports	FBS–FE3
mir	The Daily Mirror	CH1–CH3, CH5–CH7
nsc	New Scientist	ANX, B71–B7N
sco	The Scotsman	K56–K5M
uni	Unigram X	CMW–CN0, CS8–CTV

Obviously, if one is to trust the results of ths study, one must assume that these sources are homogeneous. There is no way of measuring this unless we have a good measure of similarity at our disposal, since homogeneity is closely related to similarity; for this reason, we are forced to take the suitability of the sources for granted.

For each of the sources, 50 random parts having the length of 20,000 tokens[4] (1/10 of the total), 40,000 tokens (2/10), 60,000 tokens (3/10), …, and 180,000 tokens (9/10) were taken, which mimics the approach of Tweedie and Baayen (1998) to measuring lexical diversity. For each pair of sources and for each part size, 50 distances between corresponding random parts were computed, their mean was calculted, and then the 95% confidence interval for the "true" mean of the distance value was estimated using 10,000-sample bootstrapping. For instance, for parts from *The Daily Mirror* and *Unigrams* comprising 100,000 tokens, the 50 Manhattan distances are as follows:

> 0.562; 0.564; 0.567; 0.567; 0.569; 0.569; 0.57; 0.57; 0.571; 0.572; 0.573; 0.573; 0.573; 0.574; 0.576; 0.576; 0.577; 0.578; 0.58; 0.58; 0.58; 0.582; 0.584; 0.585; 0.585; 0.587; 0.592; 0.592; 0.593; 0.598; 0.599; 0.6; 0.606; 0.607; 0.613; 0.614; 0.617; 0.624; 0.625; 0.626; 0.627; 0.628; 0.63; 0.631; 0.632; 0.637; 0.638; 0.64; 0.642; 0.642,

[4] All manipulations with the BNC were performed using Python 3.6, and, more specifically, the BNCCorpusReader class from NLTK [Bird et al. 2009]. Punctuation marks were treated as separate tokens, and word processing was case-sensitive.

the mean is 0.596, and the 95% confidence interval for the "true" mean is [0.5892; 0.6033]. This confidence interval can be further compared to the best estimate of the distance between the two sources, namely the distance between the two 200,000-token corpora—in this case, 0.5895. This distance falls withing the estimated confidence interval, which is an indicator of the fact that this measure is robust with respect to corpus size, because it yields a good estimate of corpus distance even with a relatively small corpus size. If the best estimate (i.e., the estimate based on 200,000-token portions) falls within the confidence interval for a certain pair of sources and for a certain corpus size, the measure gets 1 point; otherwise, it gets 0 points. Because we have 11 sources, the number of pairs is $11 \times 10 / 2 = 55$; for each pair, we work with 9 corpus sizes, which makes a total of $55 \times 9 = 495$ test cases for each of the six measures. The winning measure is the one that gets the most out of 495 possible points.

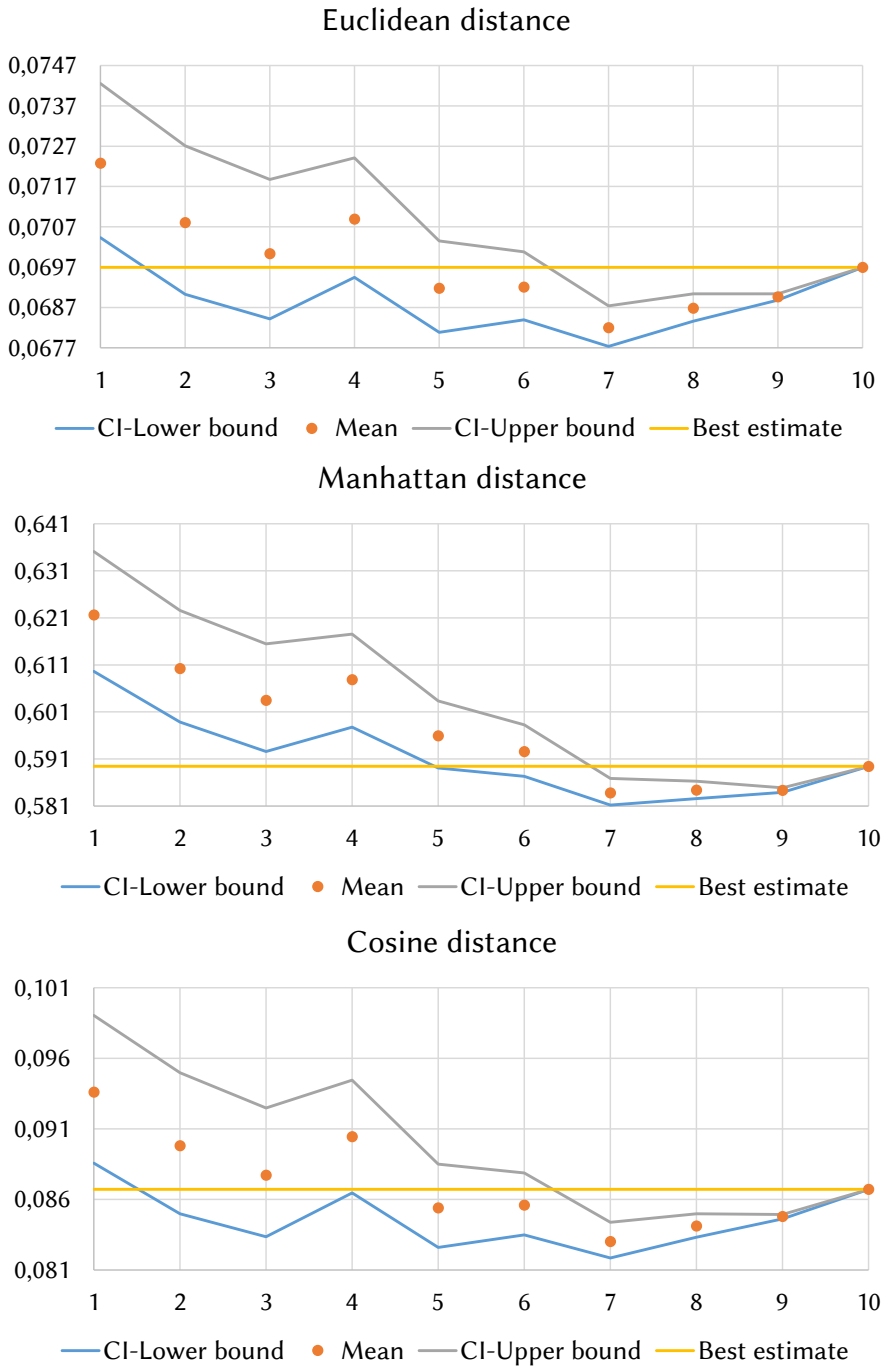
Obviously, when we measure a distance between two corpora, we never now whether the resulting distance falls close to the "true" dustance based on the whole populations. However, if we know in advance that a measure often comes close the best estimate we have, this might speak in favor of this measure.

5. Results

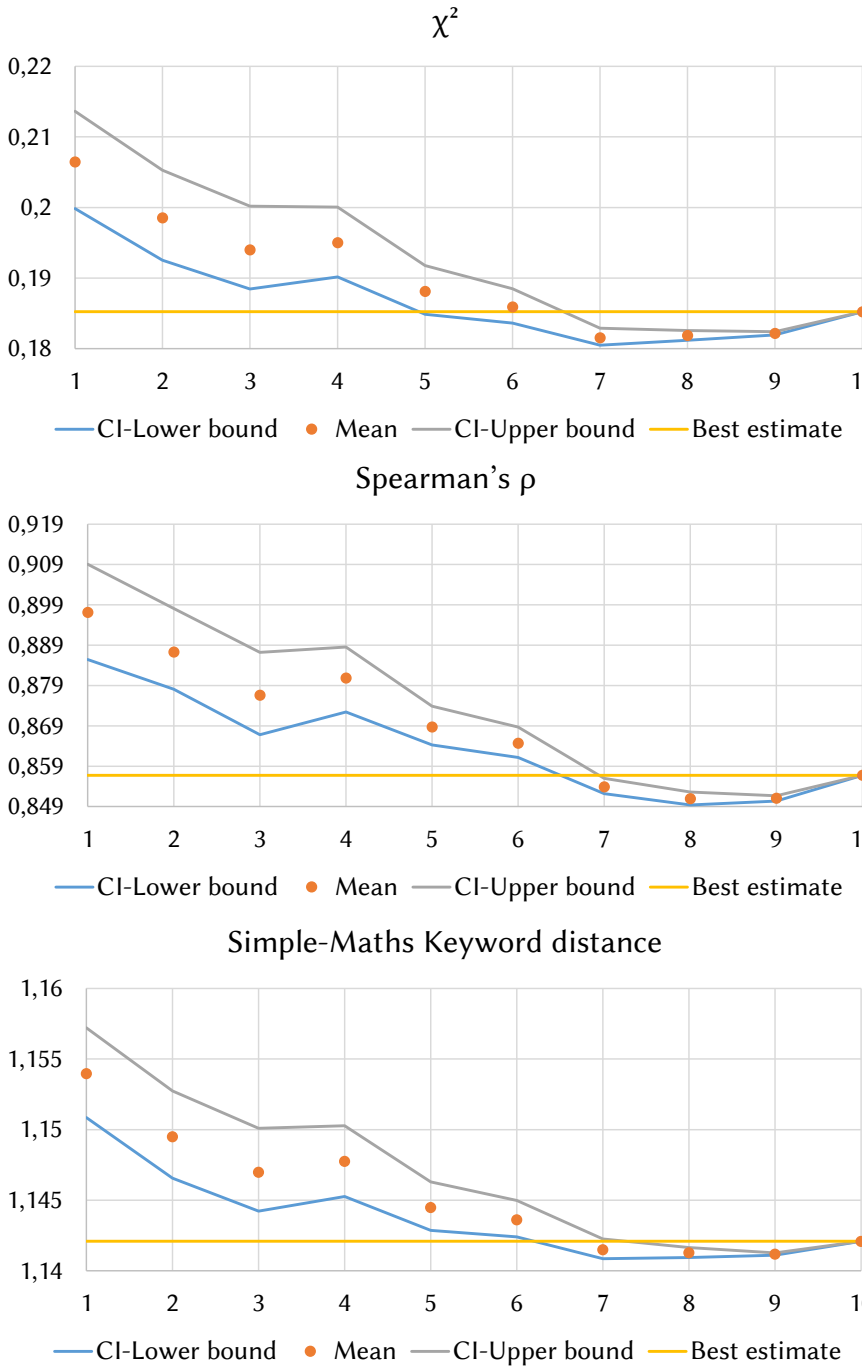
As an illustration, the result for one pair of corpora is shown in Figures 1 to 6 below. They visualize the comparison of *The Daily Mirror* with *Unigram X*. One can see from Figure 1 that the best estimate for the two sources as a whole falls within the bounds of the confidence interval in 5 cases out of 9 for Euclidean distance, because the orange line lies between the grey and the blue line for $x = 2, 3, 4, 5, 6$ (corresponding to 40,000-token, 60,000-token, 80,000-token, 100,000-token, and 120,000-token corpora). It also falls within the bounds of the confidence interval in 2 cases out of 9 for Manhattan distance, in 5 cases out of 9 for Cosine distance, in 2 cases out of 9 for χ^2, in 0 cases out of 9 for Spearman's ρ, and in 1 case out of 9 for Simple-Maths Keyword distance. Thus, in this case the two most robust measures are Euclidean distance and Cosine distance, whereas Spearman's ρ is the least robust one.

Interestingly, in all cases the distances obtained for smaller corpora are larger than the best estimate, which also holds true for other pairs of sources. The amount of variation in estimates for smaller corpora is not surprising, because larger parts must overlap with each other, whereas smaller parts do not necessarily do so. The form of the graphs is similar in all six cases (a fall, then a rise at 4/10 of the corpus, then a further fall followed by a rise), but this is an artifact of random sampling from *The Daily Mirror* and *Unigram X*; for another pair of sources, the graphs need not look the same.

The total counts of the best estimates falling within the counfidence intervals for smaller parts of corpora based on all 55 pairs of sources are presented in Table 2:



Figures 1–3. Robustness of the six distance measures as comapred using *The Daily Mirror* and *Unigram X*



Figures 4–6. Robustness of the six distance measures as comapred using *The Daily Mirror* and *Unigram X*

Table 2. Overall robustness of the six distance measures

Distance measure	Score	Percentage
Euclidean	91	18%
Manhattan	55	11%
Cosine	84	17%
χ^2	50	10%
Spearman's ρ	43	9%
Simple-Maths Keywords	42	8%

This table shows that Euclidean distance and Cosine distance are the most robust measures with respect to corpus size, whereas other measures, including both statistical measures and the keyword-based measure, are less trustworthy. This also conforms to the findings by [Piperski 2017b], who showed that geometrical measures of corpus distance perform best when assessed with the Know-Similarity Corpora approach.

6. Stability of the confidence interval

A further question arises from the fact that the evaluation technique presented above can be easily tricked. Namely, if a measure provides a wide confidence interval for some smaller corpus size, it is likely that the "true" estimate will fall within this interval. This suggests an additional requirement on the winning measure: it must not inflate the confidence interval for smaller sample sizes, i.e. its estimates for the same corpus size must not be too different from each other. The problem is that the width of the confidence interval is hard to compare across different measures. We cannot just express it as a percentage of the absolute value of the best estimate, because adding a constant to the distance would not change the measure as such, but it would change this percentage; for instance, the Simple-Maths Keyword distance as implemented in SketchEngine has a minimum value of 1, and if we were to subtract 1 from it, we would assess the relative width of the confidence interval differently.

To counter these difficulties, I propose two measures of stability of the confidence interval. First, as already mentioned in Section 5, it is evident that the variation of distance estimates between smaller corpora must be larger that the variation of distance estimates between larger corpora, simply because larger corpora drawn from the same 200,000-word population must necessarily overlap. We expect the confidence interval for 180,000-word corpora to be smaller than the confidence interval for 20,000-word corpora, and we can calculate how many times larger is the confidence interval for the mean for the smallest corpus size (20,000 tokens) as compared to the largest corpus size (180,000 tokens, since we do not have a confidence interval for 200,000-token corpus, but only a single estimate).

For example, if we apply Manhattan distance to the corpora sampled from *The Daily Mirror* and *Unigram X*, the confidence interval is [0.6097, 0.6347] for 20,000-token corpora and [0.5839, 0.5849] for 180,000-token corpora. This means that making the corpus 9 times smaller increases the confidence interval by 25 times. This value can be computed for each measure for all 55 pairs of sources. The results are summarized in Table 3.

Table 3. The increase of the width of the confidence interval from 180,000-token to 20,000-token corpora

Distance measure	Mean	Median
Euclidean	17.8	16.8
Manhattan	19.4	18.3
Cosine	21.7	19.4
χ^2	25.1	21.5
Spearman's ρ	17.0	15.4
Simple-Maths Keywords	22.7	20.8

Table 3 shows that the two best measures in this respect are Euclidean distance and Spearman's ρ. Cosine distance has performed well during the first test, but it might be due to the fact that it is likely to inflate the confidence interval.

Second, we also must check whether a distance measure is biased in some direction with respect to corpus size. Even if a measure has a relatively stable confidence interval, it may be the case that this interval is gradually drifting away from the best estimate the smaller our corpus becomes. This means that if we take one step further towards a smaller corpus, we must accept it that the confidence interval becomes larger, but we cannot tolerate if it steadily shifts in one direction. In the graphs above, there is a somewhat unsatisfying general upwards trend when looking from right to left. In order to quantify this trend, I propose the following way of computing an instability score: for $1 \leq n \leq 8$, if a distance measured for a pair of corpora containing $n \times 20{,}000$ tokens is larger than the upper bound of the confidence interval for the mean distance for $(n + 1) \times 20{,}000$ tokens, the measure is given $n / (n + 1)$ points[5]; if, on the contrary, a distance is smaller than the lower bound of the confidence interval, the measure loses $n / (n + 1)$ points. A good measure will behave symmetrically, i.e., it will receive approximately the same amount of points as it will lose, making the result close to 0.

In the worst-case scenario, a measure may receive a score whose absolute value is $(8/9+7/8+6/7+5/6+4/5+3/4+2/3+1/2) \times 50 \times 55 \approx 16{,}970$ (the value would be negative the distances are decreasing with decreasing corpus size, and positive otherwise).

Table 4. Instability scores

Distance measure	Instability score
Euclidean	2,072.5
Manhattan	4,365.9
Cosine	2,345.4
χ^2	5,462.2
Spearman's ρ	4,990.3
Simple-Maths Keywords	6,453.3

[5] Since corpora of $n \times 20{,}000$ tokens are in generally more similar to $(n + 1) \times 20{,}000$-token corpora, falling outside the confidence interval must cost more for larger values of n. However, the proposed cost of $n / (n + 1)$ was selected *ad hoc* and has no external justification.

Table 4 shows that all measures have a positive instability score, i.e. they tend to yield larger distances when corpus size decreases. The two measures with the smallest value of instability score are Euclidean distance and Cosine distance.

7. Conclusion

This paper presents the results of testing the frequency-based measures of corpus distance the for robustness with respect to corpus size. In all three experiments (Tables 2 to 4), Euclidean distance was among the two best measures, which leads us to a conclusion that it is actually the measure that is most robust to corpus size among the six measures evaluated. Further possible directions of study include evaluating robustness of distance measures when measured for corpora of different sizes as well as taking into consideration languages other than British English only.

References

1. *Bird S., Klein E., Loper E.* (2009), Natural Language Processing with Python, Cambridge (Mass.), O'Reilly Media.
2. *Gomaa W. H., Fahmy A. A.* (2013), A Survey of Text Similarity Approaches, International Journal of Computer Applications, 68(13), April, pp. 13–18.
3. *Doddington G.* (2002), Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics, In Proceedings of the Second International Conference on Human Language Technology Research, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc, pp. 138–145.
4. *Forsyth, R. S., Sharoff S.* (2014), Document dissimilarity within and across languages: A benchmarking study, Literary and Linguistic Computing, 29:1, pp. 6–22.
5. *Fothergill R., Cook P., Baldwin T.* (2016), Evaluating a topic modelling approach to measuring corpus similarity, In Proceedings of the Tenth International Conference on Language Re-sources and Evaluation (LREC 2016), Portorož, Slovenia, pp. 273–279.
6. *Kilgarriff A.* (1997), Using word frequency lists to measure corpus homogeneity and similarity between corpora, http://aclweb.org/anthology/W97-0122.
7. *Kilgarriff A.* (2001), Comparing corpora, International Journal of Corpus Linguistics, 6(1), pp. 97–133.
8. *Kilgarriff A.* (2009), Simple maths for keywords, In Proceedings of Corpus Linguistics Conference CL2009, University of Liverpool, UK, July 2009.
9. *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, pp. 7–36.
10. *Kilgarriff A., Rose T.* (1998), Measures for corpus similarity and homogeneity, http://aclweb.org/anthology/W98-1506.
11. *Lavie A., Agarwal A.* (2007), Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments, In Proceedings of the Second Workshop on Statistical Machine Translation, pp. 228–231.

12. *Papineni K., Roukos S., Ward T., Zhu W.-J.* (2002). BLEU: A Method for Auto-matic Evaluation of Machine Translation, In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318.
13. *Piperski A.* (2017a), 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.
14. *Piperski A.* (2017b), Izmerenie rasstojanij mezhdu korpusami [Measuring distances between corpora], course given at Tampere Summer School in Multilingual Corpora, Tampere, Finland, 28 August–1 September 2017.
15. *Tweedie F. J., Baayen R. H.* (1998), How Variable May a Constant be? Measures of Lexical Richness in Perspective, Computers and the Humanities, 32(5), pp. 323–352.