Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2017"

Moscow, May 31—June 3, 2017

ARBITRARINESS OF LINGUISTIC SIGN QUESTIONED: CORRELATION BETWEEN WORD FORM AND MEANING IN RUSSIAN

Kutuzov A. B. (andreku@ifi.uio.no)

University of Oslo, Norway

In this paper, we present the results of preliminary experiments on finding the link between the surface forms of Russian nouns (as represented by their graphic forms) and their meanings (as represented by vectors in a distributional model trained on the Russian National Corpus). We show that there is a strongly significant correlation between these two sides of a linguistic sign (in our case, word). This correlation coefficient is equal to 0.03 as calculated on a set of 1,729 mono-syllabic nouns, and in some subsets of words starting with particular two-letter sequences the correlation raises as high as 0.57. The overall correlation value is higher than the one reported in similar experiments for English (0.016).

Additionally, we report correlation values for the noun subsets related to different phonaesthemes, supposedly represented by the initial characters of these nouns.

Keywords: distributional semantics, word2vec, semantic similarity, edit distance, vector space models, phonosemantics

ПРОИЗВОЛЬНОСТЬ ЯЗЫКОВОГО ЗНАКА ПОД ВОПРОСОМ: КОРРЕЛЯЦИЯ МЕЖДУ ГРАФИЧЕСКОЙ ФОРМОЙ СЛОВ И ИХ ЗНАЧЕНИЕМ В РУССКОМ ЯЗЫКЕ

Кутузов А. В. (andreku@ifi.uio.no)

Университет Осло, Норвегия

Ключевые слова: дистрибутивная семантика, word2vec, семантическая близость, векторные репрезентации лексики, фоносемантика, расстояние Левенштейна

1. Introduction

The arbitrariness of linguistic sign is one the foundational principles in the studies of language since [De Saussure 1916]. It assumes that there is no relationship between the word forms (phonetic or graphematic) and their meanings: any meaning can theoretically be conveyed by any sequence of sounds or characters, and they are mutually independent. This assumption is important for many linguistic problems, and for understanding language as a system in general.

However, there are well-known exceptions from this law. Many languages feature clusters of words with similar meaning, in which some part of their surface form (for example, initial sounds) is consistently reproduced. These reproduced patterns are called *phonaesthemes* [Firth and Strevens 1930] and seem to violate the principle of sign arbitrariness. Examples of phonaesthemes in English include initial sequence "*gl-*" related to vision or light [Bergen 2004]; in Russian one can note the sequence "*-стр-*" related to quickness or streaming [Mikhalev 2008], etc.

Another exception is *onomatopoeia*: cases when phonetic form of a word is motivated by the actual sound related to the denoted notion (Russian "*мяукать*" *to meow*). In this case, the linguistic sign becomes to some extent iconic and one observes the emergence of clear relationship between the form and the meaning.

It seems obvious that this iconicity can be manifested both in localized phonosemantic sets (groups of words with similar meanings and surface forms) and in the vocabulary of language as a whole (systematicity). It seems interesting to attempt testing the actual robustness of the arbitrariness principle and to measure the degree of systematic iconicity in different languages.

This can be done by measuring the correlation between the semantic and "surface" differences between word pairs. While surface differences can be easily represented with the so-called edit distances, semantic representations of words are more difficult to obtain. However, recent achievements in distributional semantics (manifested by the advent of prediction-based and other machine learning approaches to producing vector representations of words) provided computational linguists with efficient and robust lexical meaning models which can be trained on very large corpora. These models, including the so called *neural embeddings*, exhibit substantial performance in various natural language processing tasks, including prediction of the pairwise similarities between words [Baroni 2014].

In the presented pilot study I attempt to employ neural embeddings to measure the degree of systematic iconicity in the Russian language. I describe a series of experiments with correlations of semantic and orthographic distances between frequent Russian nouns. The results seem to support the hypothesis that there is a statistically significant systematicity in the Russian language, expressed even stronger than that reported for English in [Monaghan et al. 2014].

The paper is organized as follows. In Section 2 I briefly put the research in the context of the previous work. In Section 3, the experimental design is described, together with the data sources I used. Section 4 presents the results of experiments on several datasets and discusses them. In Section 5 I conclude and outline the possible future work.

2. Related work

For English, the initial statistically rigorous experiments in phonosemantic systematicity are described in [Shillcock et al. 2001] and [Monaghan et al. 2014]. They used the Levenshtein distance [Levenshtein 1966] between orthographic word forms and the semantic distances produced by various distributional vector space models, in order to test whether differences in form are accompanied by differences in meaning. Their findings confirmed that there is a statistically significant (though low) correlation between semantic and orthographic distances in the set of mono-morphemic monosyllabic English words. Thus, the form space and the meaning space seem to be related.

Moving to more recent works, I was strongly inspired by the research of [Gutiérrez et al. 2016] in which it was proven that word embedding models can be helpful in studying violations of the arbitrariness principle in English. They also developed a new kernel-based algorithm for learning weights for different operations in the Levenshtein algorithm, which allowed finding local clusters of phonosemantic systematicity with the higher accuracy.

A different vein of research in this direction (not employing distributional semantic models) is represented by [Blasi et al. 2016]. They used Swadesh lexicons for several thousand world languages to trace bias in the frequency with which words denoting certain concepts tend to carry specific phonemes in contrast to their baseline occurrence in other words. They came to the conclusion that strongly expressed sound-meaning associations indeed exist even cross-linguistically.

For Russian, experiments related to systematic iconicity were performed by Alexander Zhuravlev (see, for example, [Zhuravlev 1991]). However, at that time it was impossible to employ large-scale distributional models, and thus opinions of limited number of informants were used to quantify semantic properties of words, rendering the results unstable and difficult to verify. I am not aware of any publications studying correlation between word embedding based semantic distances and graphematic distances for Russian.

Distributional semantic models are essentially based on the assumptions that word meaning is strongly related to the word's typical contexts [Firth 1957]. The meaning of words is represented with the so called *word embeddings*: dense real-valued vectors derived from word co-occurrences in large text corpora. They can be of use in almost any linguistic task related to semantics, and have recently become a buzzword in natural language processing (especially those trained using shallow neural networks). Their increased popularity is mostly due to new prediction-based approaches, which allowed to train distributional models with large amount of raw linguistic data very fast.

Some of the most popular word embedding algorithms in the field are highly efficient *Continuous Skip-Gram* and *Continuous Bag-of-Words*, implemented in the well-known *word2vec* tool. For more details, I refer the reader to [Turney and Pantel 2010] and [Mikolov et al. 2013]; application of these models to Russian is described, among others, in [Kutuzov and Andreev 2015].

3. Experimental setting

3.1. Data sources

I employed Russian National Corpus[1] [Plungian 2005] as the primary source of Russian texts. I also limited myself to nouns in this particular research, leaving other parts of speech to future work.

In order to test systematicity, a set of test words is needed. Ideally, it should consist of mono-morphemic words to exclude the influence of affixal word formation: in the case of "*отдел*" *department* and "*раздел*" *section*, phonetically similar words are generated to denote similar concepts, following straightforward and transparent derivation rules, not some arbitrary connection between the sound/graphical form and the meaning. It means words with shared roots should be avoided in this task.

Automatic morphemic analysis of Russian words is a difficult problem in itself [Lyashevskaya et al. 2009], so for this pilot experiments I assumed that the set of monosyllabic nouns can serve as a sort of proxy to the set of mono-morphemic nouns. I defined a "monosyllabic" word as containing one and only one vowel, and with this in mind, compiled 4 sets of nouns:

1. **Mono**: all monosyllabic nouns with frequency 100 and more in the RNC (1 729 words in total);
2. **Bi**: monosyllabic and bisyllabic words with frequency 1,000 and more in the RNC (2,900 words in total);
3. **Bi_NoDim**: the same as the previous one but excluding the nouns ending with the diminutive suffixes "*-ок*", "*-ек*" and "*-ка*" (2,633 words in total);
4. **All**: all nouns with frequency 1,000 and more in the RNC (6,715 words in total).

In all the datasets, I excluded very short words (less than three characters) and the words containing non-Cyrillic characters and digits. I also filtered out proper names and toponyms as detected by *Mystem* [Segalovich 2003].

The different choice of frequency thresholds is explained by the fact that I strive to achieve two contradictory aims: on the one hand, I need as many words in each dataset as possible (for the detected correlations to be statistically significant), and on the other hand it is desirable for the words to be as frequent as possible, in order for their distributional vectors (embeddings) to be well-trained. The chosen thresholds were selected as a good trade-off, resulting in datasets in the order of several thousand words, similar to the ones used in [Gutiérrez et al. 2016] and other related studies for English.

The main object of our experiments is the **Mono** dataset, as it is supposed to be least influenced by word formation (most words in it do not share roots) and thus its systematicity should best reflect the real relationship between form and meaning in Russian. The other three datasets were compiled for reference and to test what is the amount of influence of word-formation patterns on the phonosemantic systematicity.

[1] Further RNC.

3.2. Distributional model

For computing orthographic distances between words, I used the well-known Levenshtein (edit) distance algorithm [Levenshtein 1966] implemented in *Python*. However, to be able to calculate semantic distances a more sophisticated approach is needed.

To this end, I used the *Continuous Skipgram* distributional algorithm [Mikolov et al. 2013] which learns vectorial representations for words (neural embeddings) based on their co-occurrences in the training corpus. I trained the model on all the RNC texts, using vector size 300 and symmetric context window of 10 words to the left and 10 words to the right, leaving other hyperparameters at their default values. Prior to training, the corpus was tokenized, split into sentences, lemmatized and PoS-tagged using *Mystem*. For training itself, I employed the *Continuous Skipgram* implementation in the *Gensim* library [Řehůřek and Sojka 2010].

Distributional models can be intrinsically tested for their sanity, for example, using semantic similarity or analogy test sets. For the former, I employed the Russian part of *Multilingual SimLex999* [Leviant and Reichart 2016] which contains human judgments on the relative semantic similarity of word pairs, and the task for the model is to mimic the rankings produced by humans. This test set is known to be difficult for distributional models: its authors managed to achieve Spearman rank correlation only as high as 0.26 for Russian with the model they trained on Wikipedia. At the same time, the model used in this research showed a higher correlation of **0.36**.

The analogy test sets pose models with the task to guess one word in a "semantic proportion" (for example, "Rome is related to Italy as Moscow is related to ???"). On the translated Russian version of the *Google Analogies* dataset [Mikolov et al. 2013] the employed model showed accuracy **0.65** (using only semantic sections of the data set). There are no known published results with this translated test set for other Russian models, but the value is comparable to state-of-the-art results for English[2].

Both results also fit well into the average performance of the Russian models featured at the *RusVectores* web service [Kutuzov and Kuzmenko 2017]. Thus, I presuppose that the trained model is good enough and in general outputs correct predictions on the semantic similarities and dissimilarities of Russian words (at least comparable to state-of-the-art).

3.3. Measuring correlation

In order to measure the degree of dependency between the form and the meaning, I first calculated pairwise orthographic (string) and semantic distances between all words in the datasets. The orthographic distance was calculated as Levenshtein edit distance, while the semantic distance was equal to $1 - CosSim$, where *CosSim* is the cosine similarity between word embeddings in the vector space of the model trained on the RNC. In the rare cases when cosine similarity was negative (about 1.5% of all

[2] See http://www.aclweb.org/aclwiki/index.php?title=Google_analogy_test_set_(State_of_the_art)

the pairwise similarities), I assigned it zero value, so as the range of the cosine distance was within [0...1]. The number of pairwise distances for the dataset of n words is equal to $n\,(n-1)/2$, so I got two sets (edit and cosine) of 1,493,856 distances for the **Mono** datasets, with this number being about 3.5 million for the **Bi_NoDim** dataset, more than 4 million for the **Bi** dataset, and about 22.5 million for the **All** dataset.

Then, it is trivial to calculate any suitable correlation coefficient between the edit distances and cosine distances, that is, to what extent it is true that one of the parameters grows with the growing of another. Linguistically speaking, high correlation would mean that word pairs similar in form tend to be similar in meaning, and vice versa. Zero correlation would mean that the form and the meaning are absolutely unrelated. As the sets are quite large, I expected the calculated coefficients to be statistically significant, which proved to be true (see below).

The ordinary correlation coefficients are however not enough: they presuppose that the values in the data sets under analysis are independent, and this is not the case for the pairwise distances (changing the representation of one word influences several distances, not only one). Thus, I followed the previous work in testing the significance of the correlation using Mantel permutation test [Mantel 1967].

Mantel test essentially performs random shuffling of the value assignments in one of the two sets (for example, in the semantic distances). It generates a predefined amount of such “possible lexicons” (randomly drawn from the space of all possible permutations), and then computes the ordinary correlation coefficients between orthographic and semantic distances in these generated “lexicons”, as well as the correlation for the real lexicon. Then, the proportion of the lexicons that produced higher correlations than the real one is calculated; based on this, the veridical (true) correlation in the real lexicon is found, together with the significance measures. The idea behind this approach is that if the correlation is not accidental, one will very rarely find a higher correlation in randomly generated lexicons.

The most popular correlation measure in the literature is Pearson correlation coefficient. However, there are two reasons against using it with my data:

1. The distances in the sets are not distributed normally: for example, for the semantic distances in the **All** dataset, the normality statistics [D’Agostino and Pearson 1973] is equal to 4,775,600, with $p=0$ (zero probability that this data can come from a normal distribution).
2. The distances are strongly skewed to the right (see the Figure 1): this is arguably related to the well-known problem of *hubness* in vectorial spaces [Dinu et al. 2014].

Pearson coefficient is known to become non-robust when the data is not normally distributed and particularly when it is skewed. Thus, with my Mantel tests I employed Spearman rank-order correlation coefficient. In fact, for the experiments below, Pearson returned the same results, but I still report Spearman to be on the safe side.

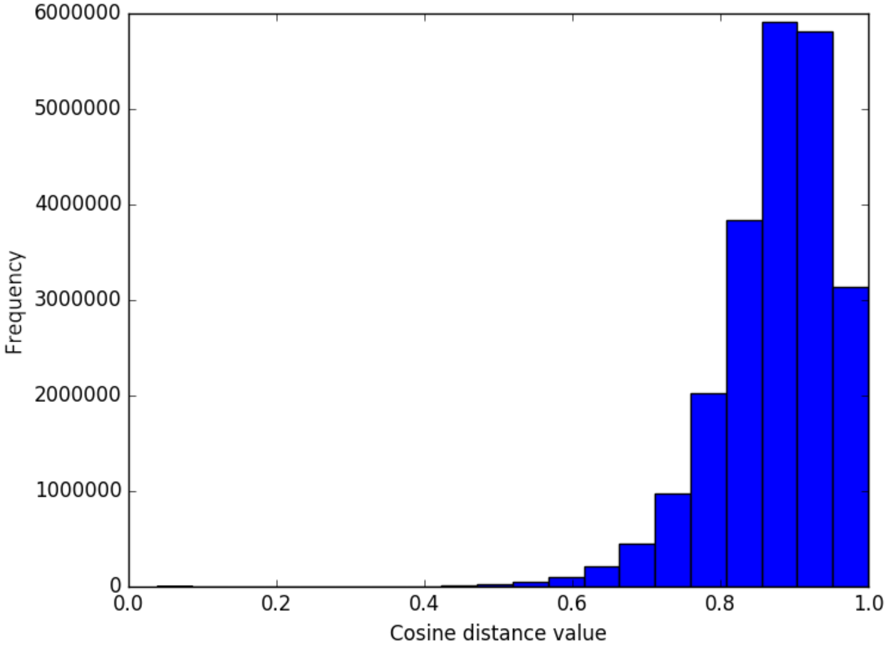

Fig. 1. Distribution of the values of pairwise cosine distances in the **All** dataset

In the next section, I describe the results of the experiments.

4. Results and discussion

I calculated Spearman correlation for all the datasets, using Mantel test with 1,000 random permutations. The results are presented in the Table 1.

As one can see, for the set of monosyllabic words, the correlation between the semantic distances and the orthographic edit distances is about 0.03, with the correlations for the less restricted datasets expectedly higher, reaching 0.08 in the case of all nouns. The value of the correlation coefficient itself is not high, but the Mantel test shows that it is strongly significant: $p = 0.001$ here means that only one lexicon from 1,000 tested has produced the correlation equal or higher to the real one[3]. Of course, this was precisely the real lexicon. Thus, all the random lexicons showed lower correlations, and it is extremely unlikely that the link between edit distances and semantic distances in the real lexicon is accidental.

[3] 10,000 permutations showed exactly the same results ($p = 0.0001$).

Table 1. Correlations between orthographic edit distances and semantic distances

Dataset	Spearman correlation	Mantel test upper-tail *p*-value
Mono	0.0310	0.001
Bi_NoDim	0.0519	0.001
Bi	0.0586	0.001
All	0.0800	0.001

The interesting fact is that in a similar experiment, [Monaghan et al. 2014] reported the correlation of only 0.016 for the set of English mono-morphemic words. The results of the experiments seem to suggest that Russian possesses at least as strong systematicity as English, and probably even stronger. This of course does not disprove the principle of the arbitrariness of linguistic sign in general; however, it is clear that there are some regular exemptions from this law, manifested throughout the lexicon.

Still, the correlation coefficient of 0.03 (and even 0.08) seems to be rather low. Considering that it is statistically very significant, the reason for this might be that the correlation is at least partly "localized" in some parts of the lexicon, not uniformly "dispersed" across all lexemes. In other words, for some nouns the connection between their form and their meaning is stronger than for the others.

One can attempt to trace this local systematicity by segmenting the original dataset into several subsets and measuring correlation for each of them. I performed this experiment on the initial two-character sequences in the **Mono** dataset, splitting it into 321 subsets corresponding to these sequences (for instance, a subset of nouns starting with "*ст-*", etc). Then, I filtered out 159 subsets containing less than three nouns, and 18 subsets with no variance in the pairwise edit distances (for example, the "*чи-*" subset containing the words "*чиж*", "*чик*", "*чин*", "*чип*", "*чиф*", and "*чих*", with all the pairwise edit distances equal to one, leaving no possibility to calculate correlation). This left me with 144 "valid" subsets.

Correlation coefficients were calculated for each of these datasets in the way described above. The distribution of correlation coefficients for all the subsets is shown on the Figure 2 (blue histogram). For some subsets the correlation was almost perfect (close to 1 or -1), but in most cases it was not statistically significant. One example of this phenomenon is the "*тв-*" subset (*"тварь" beast, "твердь" ground, "твист" twist*) with the veridical correlation equal to 1, and the *p*-value equal to 0.17, far above the 0.05 threshold of significance.

Note that it is difficult to mine anything useful from the negative correlation coefficients in this case. First, only 3 of them were statistically significant at 0.05 level. Second, even conceptually, negative correlation here means that words in the subset tend to become more similar in their meaning as the differences in their graphical form grow. This hardly makes any sense, thus I am inclined to consider the negative correlations a statistical fluctuation.

In general, it seems that grouping words by their initial characters indeed reveals local areas of high systematicity. To prove that it is not a statistical illusion, I sampled the **Mono** dataset to produce a comparable collection of 144 random subsets containing

12 words each (without replicating words across subsets), and measured correlations within these subsets. As one can see on the Figure 2 (red histogram), the distribution of correlations in these subsets is much narrower and more normal than in the initial letters based subsets. Correlation values are mostly concentrated around zero (as expected for random data), and what is important, we do not observe subsets with correlations higher than 0.4…0.5, and even those are rare. In contrast, the initial letters based subsets clearly feature many strongly correlated cases, breaking the normal distribution of correlations. This supports the point that the strength of connection between the form and the meaning of words is at least partly conditioned by their initial characters/phonemes.

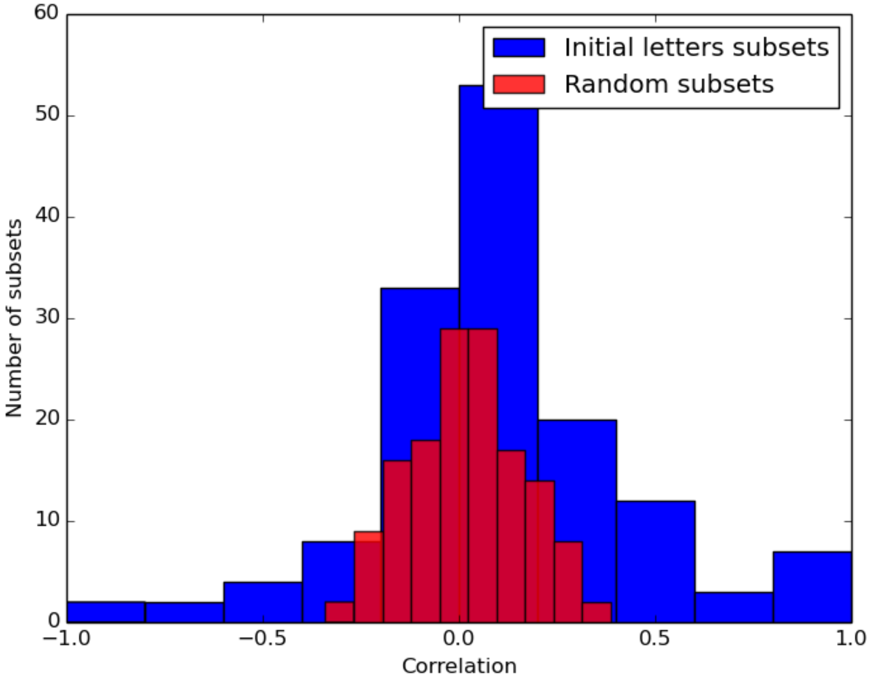

Fig. 2: Distribution of correlation coefficients in the subsets of the **Mono** dataset

The Table 2 presents 10 initial letters based subsets with the highest positive correlation among those which feature *p*-value less than 0.05 (as I was interested in the cases with the robust signal).

Some found subsets are quite interesting even with simple eyeballing. For instance, the highest correlation is demonstrated by the "*ха-*" subset featuring words like "*хай*" *loud speaking*, "*хам" mucker*, and "*харч*" *foodstuff* with this phonaestheme probably related to negative or derogatory connotations (but also "*хаббл*" *Hubble*, "*хадж*" *Hajj*, "*хан*" *khan*, "*хант*" *Khanty*, "*хань*" *Han*, "*хаш*" *khash*). The "*ше-*" subset contains "*шелк*" *silk* and "*шерсть*" *wool* (but also "*шейх*" *sheikh,* "*шельф*" *continental* *shelf*, "*шен"* as a surname, "*шень*" as a surname, "*шер*" as a proper name, "*шест*" *pole*, "*шеф*" *chief*), while in the "*гл-*" subset[4] one sees the nouns "*глубь*" *depth*, "*глушь*" *wilderness* and "*гладь*" *smooth surface*, all associated with natural substances and spaces (but also "*главк*" *department*, "*глад*" *hunger*, "*глаз*" *eye*, "*глас*" *voice*, "*глист*" *helminth*). At the same time, other subsets (like "*дж-*") seem to be not more than simple clusters of borrowed words: "*джей*" *Jay*, "*джим*" *Jim*, "*джин*" *Gin*, etc.

For certain, most (if not all) of these correlations can be explained with rigorous diachronic research: for example, some words in the pairs can be cognates. However, I still believe that these "pockets of sound symbolism" [Gutiérrez et al. 2016] deserve a closer look[5]. Whatever are the reasons for the statistically significant co-variation of the graphic form and semantics of Russian nouns, it is obvious that this co-variation exists in the present state of the language and it can be quantified. What follows is that the linguistic sign is not as arbitrary as we were used to thinking.

Table 2. Most systematic initial phonaesthemes in the **Mono** dataset

Initial	Correlation	*P*-value	Number of words in the subset
ха-	**0.57**	0.011	9
дж-	0.43	0.047	7
ше-	0.39	0.015	9
фо-	0.35	0.019	9
ва-	0.33	0.017	10
ло-	0.32	0.011	13
ле-	0.27	0.012	14
ка-	0.26	0.029	16
ку-	0.25	0.012	17
ба-	0.22	**0.005**	23

5. Conclusions and future work

I presented the results of preliminary experiments on finding the link between the surface forms of Russian nouns (as represented by their graphic forms) and their meanings (as represented by vectors in a distributional model trained on the Russian National Corpus). I showed that there is a strongly significant correlation between these two sides of word as a linguistic sign. This correlation coefficient is equal to 0.03 as calculated on a set of 1,729 mono-syllabic nouns.

In many subsets of words starting with particular two-character sequences, the correlation (statistically significant) raises as high as 0.3 and more, with one case of 0.57. The overall correlation value is higher than the one reported in similar experiments for English (0.016).

[4] Its *p*-value is 0.055, only slightly exceeding the threshold needed to get to the Table 2.

[5] All the raw data used in this paper is available at http://ltr.uio.no/~andreku/arbitrariness/.

In the future, I plan to refine the datasets by more accurate filtering of noise entities (first of all, abbreviations and proper names) and probably extract mono-morphemic words from one of the available Russian morphemic dictionaries ([Kuznetsova and Efremova 1986], [Tikhonov 2003]). I am also going to enrich the experiments to include other parts of speech except nouns.

Finally, it seems fruitful to employ string metric learning for kernel regression [Gutiérrez et al. 2016] to learn weights for different types of operations in edit distances and thus improve the sensitivity of the Levenshtein metric.

References

1. *Baroni M., Dinu G., Kruszewski, G.* (2014), Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 238–247.
2. *Bergen, B. K.* (2004), The psychological reality of phonaesthemes. Language, 290–311.
3. *Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., Christiansen, M. H.* (2016), Sound–meaning association biases evidenced across thousands of languages. Proceedings of the National Academy of Sciences, 2016 113 (39) 10818–10823.
4. *D'Agostino R., Pearson E. S.* (1973), Tests for departure from normality. Empirical results for the distributions of b2 and√ b1. Biometrika, 60(3), 613–622.
5. *De Saussure F.* (1916), Course in General Linguistics. — New York, NY : Columbia University Press, 2011.
6. *Dinu, G., Lazaridou, A., Baroni, M.* (2014), Improving zero-shot learning by mitigating the hubness problem. arXiv preprint arXiv:1412.6568.
7. *Firth J. R., Strevens P. D.* (1930), The tongues of men and speech. — 1968.
8. *Firth J. R.* (1957), A synopsis of linguistic theory 1930–1955. Studies in Linguistic Analysis (Oxford: Philological Society): 1–32. Reprinted in F. R. Palmer, ed. (1968). Selected Papers of J. R. Firth 1952–1959. London: Longman.
9. *Gutiérrez, E. D., Levy, R., & Bergen, B. K.* (2016), Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 2379–2388.
10. *Kutuzov A., Andreev I.* (2015), Texts in, meaning out: Neural language models in semantic similarity tasks for Russian, Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue" (Moscow, May 27–30, 2015), issue 14 (21), Moscow, RGGU.
11. *Kutuzov A., Kuzmenko E.* (2017), WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham
12. *Kuznetsova A, Efremova T.* (1986), Dictionary of Russian morphemes: circa 52,000 words [Словарь морфем русского языка: около 52 000 слов] — Russkiy Yazyk, 1986.

13. *Levenshtein V. I.* (1966), Binary codes capable of correcting deletions, insertions, and reversals //Soviet physics doklady. — 1966. — T. 10. — № 8. — pp. 707–710.
14. *Leviant I., Reichart R.* (2015), Separated by an un-common language: Towards judgment language informed vector space modeling //arXiv preprint arXiv:1508.00106. — 2015.
15. *Lyashevskaya O., Grishina E., Itkin I., Tagabileva M.* (2009), Word-formation annotation of the Russian National Corpus — aims and methods [О задачах и методах словообразовательной разметки в корпусе текстов], Poljarnyj vestnik. — 2009. — T. 12. — pp. 5–25.
16. *Mantel N.* (1967), The detection of disease clustering and a generalized regression approach //Cancer research. — 1967. — T. 27. — № 2 Part 1. — pp. 209–220.
17. *Mikhalev A. B.* (2008), Psycholinguistic problems of phonaesthemes. Language being of humans and ethnic groups: cognitive and psycholinguistic aspects. [Психолингвистическая проблематика фонестемы. Языковое бытие человека и этноса: когнитивный и психолингвистический аспекты.], Proceedings of the 4th International Berezin Readings.–M.: INION RAN, MGLU, (14), 140–148.
18. *Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.* (2013), Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems.
19. *Monaghan P., Shillcock R., Christiansen M., Kirby S.* (2014), How arbitrary is language? //Phil. Trans. R Soc. B. — 2014. — T. 369. — № 1651.
20. *Plungian V. A.* (2005), Why we make Russian National Corpus? [Зачем мы делаем Национальный корпус русского языка?], Otechestvennye Zapiski, 2.
21. *Řehůřek R., Sojka P.* (2010), Software framework for topic modeling with large corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta.
22. *Segalovich I.* (2003), A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine, MLMTA, pp. 273–280.
23. *Shillcock R., Kirby S., McDonald S., Brew C.* (2001), Filled pauses and their status in the mental lexicon //ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech. — 2001.
24. *Tikhonov A.* (2003), Russian derivational dictionary in 2 volumes: more than 145 000 words [Словообразовательный словарь русского языка: в двух томах: более 145 000 слов], Astrel, 2003.
25. *Turney P. D., Pantel P.* (2010), From frequency to meaning: Vector space models of semantics, Journal of artificial intelligence research. — 2010. — T. 37. — pp. 141–188.
26. *Zhuravlev A.* (1991), Sound and meaning [Звук и смысл], M.: Prosveschenie. — 1991. — T. 160.