Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2022”

Moscow, June 15–18, 2022

Refining Criteria of Paronymy for Building Computer Dictionaries of Russian Paronyms

Bolshakova E. I.
Lomonosov Moscow State University
HSE, Moscow, Russia
eibolshakova@gmail.com

Telegina A. D.
Lomonosov Moscow State University
Moscow, Russia
anya519@mail.ru

Abstract

Paronyms are words that have some similarity in sounding and spelling, but differ in meaning and usage (e.g., *sensitive − sensible, излишек – излишество*). In morphologically rich languages like Russian, paronymy is rather frequent phenomenon and one of the sources of speech difficulties. However, known dictionaries of Russian paronyms are not complete enough to help language learning or to support automatic correction of paronymy errors, and they do not provide precise definition of paronymy, which is necessary for constructing more extensive computer dictionaries. Aiming to clarify the concept of paronymy and to refine the previously proposed formal affix criterion of paronymy, we have performed a statistical study of paronyms taken from two printed dictionaries of Russian paronyms. Formal and semantic similarity of paronymy pairs were numerically estimated across various dimensions: proximity in affixes, in sounding, and in word meanings (the latter with the aid of neural models of distributive semantics and with an extensive base of Russian word combinations). Based on results of the study, refined criteria of paronymy and thresholds were proposed, which can be useful to automatically construct computer dictionaries of Russian paronyms, as well to replenish them by diagnostic contexts.

Keywords: paronymy conception; paronyms; criteria of paronymy; computer dictionary of paronyms
DOI: 10.28995/2075-7182-2022-21-61-69

Уточнение критериев паронимии для построения словарей паронимов русского языка

Большакова Е. И.
МГУ имени М. В. Ломоносова
НИУ ВШЭ, Москва, Россия
eibolshakova@gmail.com

Телегина А. Д.
МГУ имени М. В. Ломоносова
Москва, Россия
anya519@mail.ru

Аннотация

Паронимы − это слова, имеющие некоторое сходство в звучании и написании, но различающиеся по значению и употреблению (*sensitive − sensible*, *излишек - излишество*). В морфологически богатых языках, как русский, паронимия является довольно частым явлением и служит одним из источников речевых трудностей. Известные печатные словари русских паронимов недостаточно полны для помощи в изучении языка или для автоматизированного исправления паронимических ошибок, и они не дают точного определения паронимии, необходимого для построения более обширных компьютерных словарей. С целью уточнения понятия паронимии и ранее предложенного аффиксального критерия паронимии, нами было проведено статистическое исследование паронимов, взятых из двух печатных словарей русских паронимов. Формальное и семантическое сходство паронимических пар оценивалось численно по различным аспектам: близость в аффиксах, в звучании и в значениях слов (последнее с помощью нейронных моделей дистрибутивной семантики и обширной базы русских словосочетаний). По результатам исследования были предложены уточненные критерии паронимии и пороговые значения, полезные для автоматического построения компьютерных словарей русских паронимов, а также способ их пополнения диагностическими контекстами.

Ключевые слова: понятие паронимии; паронимы; критерий паронимии; компьютерный словарь паронимов

1 Introduction

Paronymy is linguistic phenomenon existing in many natural languages, as a relation between two or more words, which are similar in form (sounding and spelling), but differ in meaning and usage [7, 17], e.g.: Eng. *hare – hair*, Rus. *исправить – поправить,* Germ. *original – originell*. Such words-paronyms may be easily confused thus causing difficulties in speech understanding and creating [15, 16]. Mistakes, when one word is replaced by another word similar to it, but with different meaning, are called malapropisms in Western linguistics (e.g., *sensual news* instead of *sensational news*), while in Russian they are known as paronymy errors.

As a rule, such mistakes are typical for foreigners, but they also may appear in speech of native speakers. Several scientific works propose ways to automatically reveal and correct them [2,5], and for this purpose, appropriate computer dictionaries of paronyms are required, they are also obviously useful for teaching foreign languages. However, there are a few printed dictionaries of paronyms, for example, [1, 15], they have very limited size and rely on paronymy conception, which is not formal enough and varies significantly. Most dictionaries imply that paronyms differ only in affixes, i.e., prefixes and suffixes (e.g., *одеть − надеть, massive − mass*), the others indicate only similarity in sounding and spelling, with substantial difference in their meaning (*hare − hair*).

For Russian with its plenty of various of affixes, paronymy is a more vivid phenomenon, and there are three informative dictionaries of Russian paronyms [1, 10, 20], but the largest [10] contains only about 2,5 thou. words. Almost all paronyms in these dictionaries have the same root and part of speech (POS), thus having similar but somewhat different meaning. Differences in meanings of paronyms are explained and illustrated by diagnostic contexts. Meanwhile, the dictionaries do not give a precise definition of paronymy, which is necessary to build a more complete dictionary of paronyms.

The first computer dictionary of Russian paronyms constructed in [6] contains only word pairs differing in one or two letters (such as *комплекс − комплект*), whereas numerous word pairs that differ in several suffixes were not included. The work [4] describes a method to build computer dictionary of paronyms, which is intended to correct paronymy errors in Russian texts and based on the proposed formal affix criterion of paronymy. Only those pairs of words of the same root and POS are recognized as paronyms, for which the differences in affixes (suffixes and prefixes) are within the particular fixed limits. The volume of the built dictionary (about 135 thou. pairs of paronyms) is larger than all known dictionaries, but it also contains many word pairs that hardly be attributed to paronyms, in particular, dissimilar couples (such as *ходули − перерасходы*).

Thus, to build computer dictionaries more accurately, criteria of paronymy are to be further investigated. The difficulty of formalizing paronymy is related to several various aspects of similarity, as well as subjectivity of its perception. In order to refine criteria of similarity, we have performed a computational study of a representative set of paronyms (hereafter, etalon set) that were taken from two dictionaries of Russian paronyms [1, 10]. They were manually compiled by linguists and thus guarantee paronymy in its intuitive sense. In our work, only word pairs with the same root and POS were considered, and the following statistics for these pairs were evaluated:

- closeness in affixes, depending on POS (nouns, verbs, and adjectives, including participles);
- proximity in sounding by applying Soundex algorithm [9];
- difference in semantics, with the aid of neural distributive models [14] and CoSyCo corpus of syntactically related words [8].

Based on the revealed features of the paronymy pairs, we have formulated a refined affix criterion of paronymy that depends on POS, and also have proposed cut-off thresholds for proximity in sounding and in semantic similarity, which can be useful for building computer dictionaries.

We also have performed experiments to automatically build dictionaries of Russian paronyms and to estimate coverage of the etalon set. For the study and the experiments, we have exploited datasets[1] with Russian words split into morphs, open source programming tools[2], as well as our own tools[3] developed for the task.

Thus, the contributions of this paper are the following:

[1] https://github.com/cmc-msu-ai/NLPDatasets
[2] https://github.com/cmc-msu-ai/ParonStatistics
[3] https://github.com/annatelegina/ParonymsAnalyzer

- we present the results of the statistical study of paronyms from the etalon set, their closeness in affixes, sounding, semantics, and combinability;
- we propose refined criteria of paronymy that account for various aspects of word similarity and depend on preset thresholds;
- we report on the experiments for building dictionaries of Russian paronyms undertaken with the developed open-source tools and various combinations of the proposed criteria.

In the next section, we shortly describe and compare paronymy dictionaries [1, 10, 20], consider the affix criterion of paronymy used in [4], and based on this, clarify our understanding of paronymy. The results of the statistical study of formal and semantic proximity of paronyms from the etalon set are reported in sections 3 and 4, respectively. Section 5 describes experiments on building several computer dictionaries of Russian paronyms. Conclusions are presented in Section 6.

2 Paronymy Conception and Paronymy Dictionaries

Printed dictionaries of Russian paronyms [1, 10, 20] contain only nouns, verbs, adjectives (including certain participles), and most paronyms have the same root and the same part of speech (POS). Meanwhile, conception of paronymy somewhat vary: for example, in [20] the same gender for nouns and aspect for verbs, the same number of syllables and similar place of accent are indicated as additional features for true paromyms, thereby narrowing the conception. Paronyms are gathered either by pairs (~1000 pairs in [20]) or by so-called paronymy groups of 2–7 semantically close words (about 200 groups in [1] and 1100 groups in [10], such as *земельный – землистый – земляной – земной*). The relation between paronymy conception and semantics of the compared word roots remains unclear: in [1] similar words with different roots (e.g., *индейка – индианка*) called quasi-paronyms, in [10] words with homonymous roots (e.g., *платный – платяной*) are not considered as true paronyms, but this dictionary includes synonyms (*патетический – патетичный*), as well as words differing only in combinatorics (*туристский – туристический*). Nevertheless, the dictionaries present many diagnostic contexts for distinguishing meanings of particular paronyms (e.g., *игорный бизнес – игральный стол – игривый щенок – игристое вино – игровая зона*).

All in all, these limited-size dictionaries do not give a precise definition of paronymy needed to build a more complete computer dictionary of paronyms. For this purpose, the work [4] proposed a formal affix criterion of paronymy that takes into account statistics of affix proximity of paronyms from the largest dictionary [10]. The affix similarity of two words of the same root and POS is estimated separately for prefixes and suffixes, by a pair of integers (N_p, N_s), N_p is the number of different prefixes, i.e. the minimum number of elementary editing operations [11] transforming chain of prefixes of one word into prefixes of another word. The number N_s for suffixes is computed similarly. The affix criterion is written as $(N_p = 0)\ \&\ (N_s <= 3) \vee (N_p = 1)\ \&\ (N_s <= 2)$ – either the prefixes in the compared word pair are the same, and there are no more than three differences in suffixes, or words have one different prefix, and there are no more than two differences in suffixes, for example: *о-дар-ённ-ый – дар-овит-ый* ($N_p = 1$, $N_s = 1$), *такт-ик-а – такт-ичн-ость* ($N_p = 0$, $N_s = 2$). Though this affix criterion covers almost 99% of paronyms from the dictionary [10], and its application to words from [3] produced a volume computer dictionary of paronyms (~135 thou. pairs), the resulted dictionary has some drawbacks. It turned out that the criterion allows antonyms (*типичный – атипичный*) and near synonyms (*патетический – патетичный*), some outwardly dissimilar word pairs (*седловина – сиденье*), and pairs distinguished only by a diminutive suffix (*мел – мелок*). At the same time, some similar pairs (e.g., *мерзость – омерзительность*) were not recognized as paronyms.

Aiming to refine the formal criterion of paronymy, we have studied various aspects of similarity for the set of paronyms taken from [10] and enlarged by paronyms from the dictionary [1], additional features of the paronymy pairs were studied as well. We considered pairs of words with the same root and POS, not paronymy groups, because some words within a particular group may vary significantly in affixes and semantics. Since surprisingly many noun pairs within paronymy groups have different gender, unlike [1, 4], we do not exclude such pairs from consideration.

The preliminary analysis of paronyms from all the dictionaries [1, 10, 20] has showed that allomorphism of the roots (owing to alternating consonants and fluent vowels, e.g., *отчий – отеческий*) is allowed, but there are no pairs distinguished by:

- antonymic prefixes (*не-, а-, анти-, контра-, против-,* etc.);
- foreign prefixes (e.g., *hypo, infra*) and prefixoids (*меж-, само-,* etc.);
- diminutive and magnifying suffixes (*-ушк, -онок, -ёнок, -ища,* etc.);
- postfixes (*-ся, -сь*) or verb aspect (perfect and imperfect).

Therefore, such word pairs should not be included in paronymy dictionaries. In our opinion, synonyms also do not belong to true paronyms, although there are several pairs in [10] (e.g, *крохотный – крошечный*). At the same time, similar to the work [4] it is reasonable to admit quite similar pairs with homonymous roots (such as adjectives *бурный – буровой*). But unlike [4], in our study:

- we consider a more representative set of paronyms encompassing two dictionaries [1, 10];
- the words are split into morphs (prefixes, root, suffixes, ending) according to the derivational dictionary [19] and the work [13] (as splitting of suffixes used in [3, 4] is not conventional);
- statistics for word pairs were collected and evaluated separately for different POS;
- besides affix proximity of paronyms, their semantic similarity was estimated in order to exclude synonymous pairs and also pairs differing by ambiguous diminutive suffixes used not in diminutive sense (e.g., *цвет – цветок*).

In total, our etalon set of paronyms encompasses 2704 words, most of them belong to adjectives: 54% of adjectives (including participles), 33% of nouns and 13% of verbs. All paronymy pairs (2013) were estimated for their formal similarity and semantic proximity.

3 Formal Similarity

3.1 Affix Distances

To estimate affix proximity for pairs from the etalon set of paromyms, morpheme segmentation of words were taken from RuMorphs-Lemmas[4] dataset (created from [19]) and were additionally modified according to [13], as this is more relevant for comparing word pairs separately within the considered POS: nouns, adjectives, verbs. In particular, the so-called thematic vowels of verbs *a, i, e* were attached to the verb suffixes and suffixes of participles, some adjacent suffixes were also concatenated (e.g., *норм-ирова-ть* instead of *норм-ир-ов-а-ть*). Statistics of affix distances for paronymy pairs were automatically computed, the results are given in Table 1. The first two columns show all discovered combinations of distances in prefixes (N_p) and suffixes (N_s), respectively. The other columns present the number of paronymy pairs for the particular POS, according to the corresponding combinations. The affix distance equals the minimal number of editing operations [12] on affixes (their deletion, insertion or substitution) that transform one word of the compared pair into the other (word endings are not taken into account).

N_p	N_s	Nouns	Adjectives	Verbs
0	0	12	11	4
0	1	**273**	**544**	9
1	0	89	30	**331**
1	1	8	26	0
0	2	206	357	4
2	0	0	3	1
1	2	3	11	0
0	3	34	57	0
	2013	625	1039	349

Table 1: Statistics of affix distances for paronymy pairs

One can notice that the statistics differ for various POS: most of verb pairs (around 94%) are prefix paronyms with distance $N_p = 1$, only 4 pairs have suffix distance $N_s = 2$ (e.g, *хозяйничать – хозяй-*

[4] https://github.com/cmc-msu-ai/NLPDatasets

ствовать, etc.) and only one pair has prefix distance $N_s = 2$ (*изобразить – отразить*). In contrast, for adjectives and nouns, most pairs are suffix paronyms with distances $N_s = 1$ and 2, and prefix paronyms with $N_p \geq 1$ are relatively rare or absent.

This statistics corresponds to the well-known fact that Russian derivation for verbs is predominantly prefixal, while for nouns and adjectives, it is suffixal. Thus, the criteria of affix similarity for paronyms should obviously depend on the particular part of speech.

Some features are common for all POS: most paronymy pairs differ only in one affix, and many pairs differ in two affixes. Some couples turned to be at the minimum distance (0, 0), when the words do not differ in affixes and have the same root morph, in particular, verbs with allomorphic roots (*огородить – оградить*), adjectives differing only by ending (*временный – временной*).

It seems acceptable do not account for rare pairs with distances (2, 0), so the refined affix criterion of paronymy can be set as follows, separately for various POS:

for nouns and adjectives: $(N_p = 0)\ \&\ (N_s <= 3)\ \vee\ (N_p = 1)\ \&\ (N_s <= 2)$

for verbs: $(N_p = 0)\ \&\ (N_s <= 2)\ \vee\ (N_p = 1)\ \&\ (N_s = 0)$.

This affix criterion cover 99.4% pairs from the etalon set. When applying the refined affix criterion for building a computer dictionary of paronyms, it is necessary to additionally exclude pairs that differ only i) by any antonymic prefix; ii) by any foreign prefix or prefixoid; iii) by postfix; iv) by a nonambiguous diminutive suffix (e.g., *аргумент – контраргумент, сетевой – межсетевой, умывший – умывшийся, кот – котик*, etc.). For this purpose, a pre-compiled list of such prefixes, prefixoids, and suffixes is used, but it does not contains ambiguous diminutive suffixes (*-чик, -ик,* etc.), because it is necessary to separate pairs with suffixes in diminutive sense (*дом – домик,*) from those without diminutiveness (*перевод – переводчик, такт – тактика*), and additional techniques are required for their recognition.

3.2 Proximity in Sounding

Another aspect of word similarity is their sound proximity and to estimate it, we use Soundex algorithm [9] and its open implementation[5] for Russian. For a given word pair, this algorithm produces an integer that indicates the degree d of sound proximity: the smaller it is, the more similar the words sound. The sound proximity was evaluated for paronymic pairs from the etalon set, separately for various POS, the results are given in Figure 1. Though most paronymy pairs has d equal to 2 or 3, the situation for verbs is again different from that for nouns and adjectives: the most verb pairs have $d < 5$. The "tail" of statistical distribution corresponds to rare dissonant pairs (such as *натуралистический – натурный* with $d = 10$). If sound proximity is important, such pairs can be eliminated from a computer dictionary, with the aid of a preset threshold P_d: pairs with $d > P_d$ are to be excluded. In particular, $P_d = 5$ for verbs and $P_d = 7$ for nouns and adjectives exclude 19 pairs of adjectives, 10 pairs of nouns, and only 1 pair of verbs (*сигнализировать – сигналить*) from our etalon set.

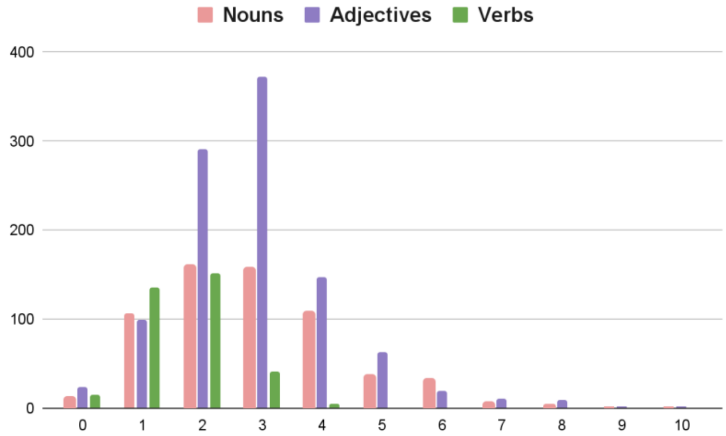

Figure 1: Statistics of sound proximity

[5] https://pypi.org/project/ru-soundex/

4 Semantic Proximity

4.1 Word Embeddings

Neural models of distributive semantics are widely used in modern computational linguistics, so we used word-level models relevant for our task, Word2Vec [14], to estimate semantic proximity of paronyms. Words are represented as vectors (embedding) in the space formed by such models, and similarity (or difference) in words meaning are evaluated by cosine distance between their vectors: if it is large enough, then the words have similar meanings. Thus, the models are suitable to identify synonyms (in order to exclude them from a dictionary), and also to reveal diminutive sense of ambiguous suffixes *(-чик-, -ик-, -к-, -ок-, -ец-, -иц-, -ц-* , etc.) in particular words.

Among pre-trained Word2Vec models from RusVectores project[6] [11], for our experiments we have chosen RusCorpora and Tayga – cf. Table 2, since they contain more words[7] from the etalon set (the number of absent word are given in the third column of Table 2).

Text Collection For Training	#Words	Training Method	#Absent Paronyms
RusCorpora	278 million	CBOW	110 (171 pairs)
RusCorpora + Wikipedia	778 million	SkipGram	133 (202 pairs)
Tayga	5 billion	CBOW	106 (162 pairs)
Russian News	2.6 billion	SkipGram	400 (434 pairs)

Table 2: Word2Vec models from RusVectores

The statistical distribution of computed cosine distances *(dcos)* for pairs of paronyms is presented in Figure 2. Values *dcos* vary greatly, the right "tails" correspond to pairs that are close in meaning and may be synonyms. In experiments we have found the threshold $P_s = 0.8$ (for Tayga model), such that pairs close in meaning ($dcos > P_s$) can be considered synonyms and so be excluded from the dictionary of paronyms: *крохотный – крошечный* (*dcos=0.98*), *завесить – занавесить* (*dcos=0.81*), *целебный – целительный (dcos = 0.81*). However, words that seem synonymous may differ by combinations with other words used in speech, and for word pairs with *dcos* close to the cut-off threshold, it seems reasonable to additionally compare their combinability with other words.

To identify diminutive meaning of an ambiguous diminutive suffix, we have performed additional experiments with 450 arbitrary taken pairs differing by such suffixes, and have empirically chosen a threshold for *dcos*: $P_a = 0.55$. Word pairs with $dcos \geq P_a$ are excluded (e.g., *блокнот – блокнотик, dcos=0.75*), while pairs with $dcos < P_a$ are recognized as paronyms (*макт – тактика, dcos = 0.14*).

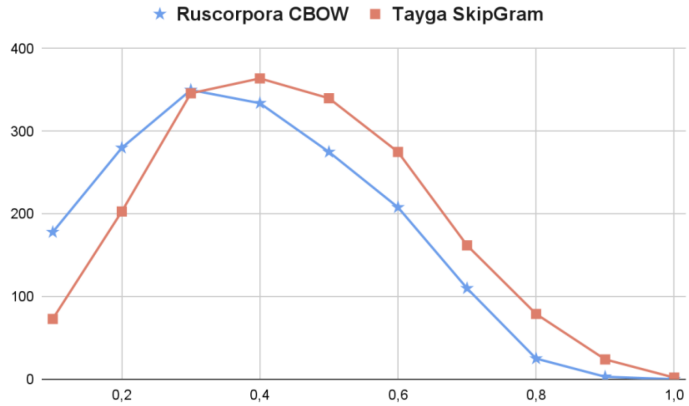

Figure 2: Distribution of semantic proximity *dcos*

[6] https://rusvectores.org/ru/

[7] Amost all such models do not include vectors for words that are either rare or not encountered in the text collection for training. FastText still constructs vectors for such word, but we cannot rely on their properties.

4.2 Word Combinability

Since paronyms are often distinguished by the words combined with them, another way to compare their semantics implies revealing differences in context words syntactically related with them. Corresponding diagnostics contexts are usually presented in paronymy dictionaries, such as *болотный цвет* but *болотистая местность* for paronyms *болотный* and *болотистая*.

Based on the assumption that the closer the words are in meaning, the greater the number of identical words combined with them, and vice versa, we evaluate their semantic similarity (or difference) by comparing their combinations with words of other POS, taking them from CoSyCo [8], a large corpus of syntactically related Russian words (~1.75 million combinations). The corpus was automatically constructed from large text collections and encompasses word combinations (bigrams) of several types, along with frequencies of occurrences for each particular combination and its words-components.

Since CoSyCo contains a plenty of word combinations (up to 2–3 thou.) for each particular word, and most of them are neither stable nor idiomatic, we need to reveal the most stable and typical among them. After extraction from CoSyCo all combinations: *Adjective + Noun* for each nouns or adjective and *Verb + Noun* for each verb, they are ranged by applying logDice association measure [18]:

$$logDice = log_2 \frac{2*Freq(a,b)}{Freq(a) + Freq(b)},$$

where *Freq(a, b), Freq(a), Freq(b)* are frequencies of the word combination and its components, respectively. This measure is well-known and well-performing for revealing stable combinations, and it is suitable for our task, because it does not require the size of source texts (on which statistics were collected), such information is absent in CoSyCo.

Having two ranged lists of word combinations for two given words W_1 and W_2, we take N-tops of the lists (the most stable combinations), and compare sets S_1 and S_2 of words combined respectively with W_1 and with W_2, by computing proportion of common words in them (more precise, cardinality of intersection of these sets, divided by N):

$$sim = \frac{|\, S_1 \cap S_2 \,|}{N}$$

In such a way, semantic similarity *sim* of the words can be estimated, but what is more important, words that are not included in the intersection (i.e., are not common) distinguish the meaning of the compared words W_1 and W_2 and thus can be considered as diagnostic contexts. For example, if N=10, for paronyms *дружеский* and *дружественный* semantic similarity equals to 0.3 (they are somewhat similar), and their diagnostic contexts include: *дружеский – вечеринка, участие, пирушка, шарж* but *дружественный – интерфейс, страна, держава, государство*.

To find the number N that is appropriate for extracting diagnostic contexts for paronymy dictionary, we have evaluated *sim* values for paronyms from the etalon set, for various N (up to 80) and separately for various POS, the resulted mean values are shown in Figure 3.

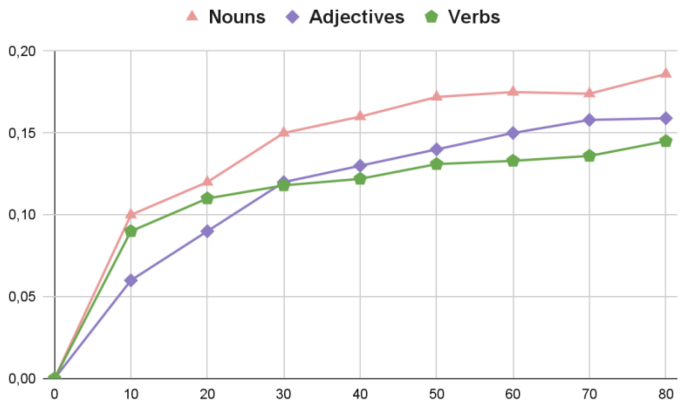

Figure 3: Distribution of similarity in word combinability

The growth of similarity values *sim* with the growth of N are explained by increase of common words for the considered paronyms, but at the same time the stability of the corresponding word combinations decreases – this means that distinguishing words (diagnostic contexts) are appeared more rarely and within less stable combinations. Our analysis of extracted word combinations for 90 paronymy pairs has showed that $N = 20$ is appropriate for revealing typical diagnostic contexts to be included them into a dictionary of paronyms being constructed.

5 Constructing Dictionaries with Refined Criteria

Due to ambiguity in understanding of paronymy and various purposes of paronymy dictionaries, it is reasonable to consider the above-proposed cut-off thresholds P_s, P_a, P_d as parameters of procedure for building a dictionary – this makes it possible to set more or less strict restrictions on paronymic pairs, depending on the task. In order to build different dictionaries, the developed programming tools[8] provide API for setting needed thresholds and other parameters.

In our experiments to build dictionaries of Russian paronyms, as input data we have used RuMorphs-CrossLexica[9] dataset (26 thou. words taken from [3]). To estimate built dictionaries (and thus the refined criteria), we have evaluated their volumes, as well as coverage of the etalon paronymy set, the coverage degree is calculated as follows:

$$C_{measure} = \frac{| D \cap E |}{| E |}$$

where D is the set of paronymic pairs in the built dictionary and E are pairs in the etalon set. Table 3 presents results obtained for different combinations of criteria and the proposed thresholds: $P_s = 0.8$, $P_a = 0.55$, $P_d = 5$ for verbs and $P_d = 7$ for nouns and adjectives. The first column corresponds to particular combinations of the criteria, while the second and the third show the coverage and volume (thou. paronymic pairs) for the built dictionary.

Criteria	$C_{measure}$	Volume
Refined Affix Criterion	99.7	100.7
Refined Affix Criterion + P_S	99.6	100.6
Refined Affix Criterion + P_a	99.5	101.0
Refined Affix Criterion + P_d	98.2	98.1
All Criteria	98.0	97.4

Table 3: Comparison of the built dictionaries

One can notice that coverage degree slightly decreases with the introduction of the thresholds, as does the volume of the resulted dictionary of paronyms. With all refined criteria, the volume of the dictionary is about 97.4 thou. paronymy pairs (instead of 135 thou. pairs and 99% coverage degree for the dictionary built in [4]). Decrease of the volume and the coverage does not mean that the refined criteria work worse, as only word pairs are excluded that are less typical for paronymy.

In addition, we manually reviewed some fragments of the resulted dictionary, to estimate the appearance of non-paronymy pairs in them. The most pairs correspond to our refined understanding of paronymy, however, there are some pairs of participles (such as *увиденный – увидевший*, passive and active forms), but whether such pairs should be excluded from paronyms is an open question that requires further research.

6 Conclusions

Based on the study of paronyms taken from the representative etalon set, their features of formal and semantic proximity, we have refined affix criterion of paronymy in Russian, as well as have proposed the ways to estimate proximity of potential words-paronyms in sounding and semantics. It turned out that similarity in affixes and in sounding differ for nouns and adjectives, on the one hand, and verbs,

[8] https://github.com/annatelegina/ParonymsAnalyzer

[9] https://github.com/cmc-msu-ai/NLPDatasets

on the other hand, thus resulting in the affix criterion depending on POS of words being compared. For estimating semantic proximity, two ways were considered: neural models of distributive semantics and the large database of word combinations, the latter enables to reveal diagnostic contexts (typical word combinations) that distinguish meaning of compared words. The proposed cut-off thresholds for proximity in sounding and semantic similarity can be changed according to particular tasks.

Clearly, it is hardly possible to fully formalize the concept of paronymy, since it is associated with diverse aspects of word similarity, as wel as subjectivity of its perception. Nevertheless, dictionaries of Russian paronyms can be automatically built for various applied tasks, in particular, for teaching languages or automatic correction of paronymy errors. The dictionaries may differ in volume, degree of formal and semantic similarity of paronymy pairs, which can be achieved with the developed programming tools, by setting appropriate thresholds.

References

[1] Belchikov Yu. A., Panjusheva M. S. (2004), Dictionary of Russian Paronyms [Slovar' paronimov russkogo jazyka], Russkij Jazyk, Moscow.

[2] Bolshakov I.A., Gelbukh A. (2003) On Detection of Malapropisms by Multistage Collocation Testing. // A. Düsterhöft, B. Talheim (Eds.) Proc. Int. Conf. Applications of Natural Language to Information Systems NLDB´2003, June 2003, Burg, Germany, GI-Edition**,** LNI V. P–29, p. 28–41.

[3] Bolshakov I.A. (2013), CrossLexica – Universum of links between Russian words [CrossLexica – universum svyazey mezshdu russkimi slovami], Business Informatics [Biznes Informatica], No 3 (25), pp. 12–19.

[4] Bolshakova E. I., Bolshakov I. A. (2015), Affix criterion of paronymy for building a computer dictionary of Russian paronyms [Affiksal'nyj kriterij paronimii dlya postroeniya komp'yuternogo slovarya paronimov russkogo yazyka], Scientific and technical information [Nauchno-tekhnicheskaya informaciya], Ser. 2, № 11, pp. 28–35.

[5] Costin-Gabriel C. (1998), Malapromisms detection and correction using a paronyms dictionary, a search engine and Wordnet, Search, Bucharest, pp. 364-373.

[6] Gusev V. D., Salomatina N. V. (2001), Electronic dictionary of paronyms [Elektronnyj slovar' paronimov, ver. 2], Scientific and technical information [Nauchno-tekhnicheskaya informaciya], Ser. 2, № 7, pp. 26–33.

[7] Hartmann R. R. K., James G. (1998), Dictionary of Lexicography. London, Routledge.

[8] Klyshinsky E., Lukashevich N., Kobozeva I. Creating a corpus of syntactic co-occurrences for Russian. In: Computational Linguistics and Intellectual Technologies: Proc. of the International Conference "Dialogue 2018", Issue 17 (24). Moscow, 2018, pp. 311-324.

[9] Knuth D. E. (1973). The Art of Computer Programming: Volume 3, Sorting and Searching. Addison-Wesley, pp. 391–392.

[10] Krasnykh V.I. (2007), Explanatory Dictionary of Russian Paronyms [Tolkovyj slovar' paronimov russkogo jazyka], AST Astrel, Moscow.

[11] Kutuzov A., Kuzmenko E. (2017), WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Analysis of Images, Social Networks and Texts. AIST 2016. CCIS, vol. 661, Springer, Cham.

[12] Levenshtein V.I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". Soviet Physics Doklady. 10 (8): 707–710.

[13] Lopatin V. V., Ulukhanov I. S. (2016). The dictionary of word-forming affixes of the modern Russian language [Slovar' slovoobrazovatel'nyh affiksov sovremennogo russkogo yazyka], Azbukovnik, Moscow.

[14] Mikolov T., Chen K., Corrado G. (2013), Dean J. Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at ICLR, Arizona.

[15] Müller W. (1973), Easily confused words [Leicht verwechselbare Wörter], Duden Taschenwörterbücher, vol. 17, Mannheim, Bibliographisches Institut.

[16] Péchoin D., Dauphin B. (2001) Dictionnaire des dificultés du français d'aujourd'hui. Larousse.

[17] Rosenthal D. E., Telenkova M. A. (1976), Dictionary-handbook of linguistic terms [Slovar'-spravochnik lingvisticheskih terminov], Prosveshchenie, Moscow.

[18] Rychlý P. (2008) A Lexicographer-Friendly Association Score: RASLAN, Brno.

[19] Tikhonov A. N. (1990), Word Formation Dictionary of Russian language [Slovoobrazovatel'nyj slovar' russkogo yazyka], Russkij yazyk, Moscow.

[20] Vishnyakova O. V. (1984), Dictionary of Russian Paronyms [Slovar' paronimov russkogo yazyka] Russkij yazyk, Moscow.