Antidictionary

You

18 сентября 2018 г.

Аннотация

In this paper, we describe a database of Russian words that can be found on the Internet but are not included in dictionaries, for example new loanwords, slang words, misspelled words, intentionally misspelled words, proper names etc. We also report a pilot study of words with multiplicated letters and the orthography-phonology connection of letter multiplication.

1. Introduction

In this work, we are describing the dictionary of out-of-vocabulary words (”Antidictionary”), based on The General Internet-Corpus of Russian. The primary purpose of this dictionary is providing the possibility to track and research linguistic phenomena which are not reflected in dictionaries and corpora that are based on fiction, newspapers or academic texts or spoken language. The aim of this work, therefore, is constructing a database that such dictionary should be based on.

There is a traditional distinction between two modes of language and communication: written and spoken. Since the 1990s, however, it is possible to talk about the new dominant attitude that describes the Internet language as the third mode of language [Cry01].

The Internet Russian is a real mine of information about communication in total and allows us to study the language like never before. On the other hand, it is underexplored, and there is a lot of areas which have to be dealt with in a particular way, e.g., syntax or some orthographic features. We believe that the essential discoveries in this area are only waiting to happen.

According to [Cry01], there are four main perspectives for the further investigation of the Internet language: the sociolinguistic perspective, the educational perspective, the stylistic perspective and the applied perspective. All these perspectives are linked with each other and provide a breeding ground for reflecting the language in various directions. We suppose it is necessary to have a linguistic source that combines all research directions mentioned above.

Our work is based on the materials provided by The General Internet-Corpus of Russian, namely, a list of words that were not recognized by the parser that they use from the part of the corpus based on livejournal.com blogs. This list is truly gigantic (about two gigabytes of text). We cleaned this list from duplicates and words that actually belong to Standard Russian: the final list that we worked with consists of 7 998 443 words, which makes manual annotation barely possible. Our aim is to turn the rest into a database of out-of-vocabulary words, automatically annotated for some of their morphological and other features. This database is further called the dictionary of out-of-vocabulary words.

THere is a lot of dictionaries that deal with the Internet language, slang and neologisms, but our dictionary is distinct in its core function. It is not an explanatory dictionary, but the database which can be used for research of the Internet language in different ways.

Here are some examples of research areas where our database might be useful:

1. From the sociolinguistic perspective, we can observe the transformation of communication. The Internet changed the way how people communicate. It affects the language and gives it new stylistic varieties, mostly in the field of the informal style. Moreover, language changes can be influenced by shifts in society and economy (e.g., globalization). There is a lot of possibilities to observe this and study such phenomena like language calques, loan words and their localization, gendered nouns and genderless language, the slang of particular groups and Internet memes.

2. From the educational perspective, it is possible to research two main areas: the influence of the Internet on language education and the research of spelling mistakes. Nowadays people deal with a lot of documents created by humans online. Under the best of circumstances these texts are automatically spell-checked, and it does take its toll on language education. The formal and academical style and written forms of informal language are blending together. Transferring of the informal communication into the Web also affects the spelling and grammar of what people write, and many linguists stick to the opinion that the Internet has a negative influence on the language standard [Bar13]. On the other side, there is a high possibility to research all kinds and types of errors in written language in connection with their frequency, the diachronic change and synchronous preferences and typical patterns [KB13].

3. From the stylistic perspective, there is a possibility to research new and different forms of language creativity [Cry01]. There are not only neologisms, but also phenomena connected only with the written kind of language, like multiplicated letters or the use of upper case for emotional expression.

It must be stressed that this is only a non-inclusive list of the possible research problems that can be addressed when using our dictionary. In our work, we not only describe the foundations of developing the “Antislovari”, but also provide an example of the possible linguistic research.

2. Implementation

In this section, we describe our data and the ways we dealt with it.

2.1. Background and Sources

As the primary source for our dictionary we have used the data collected by The General Internet-Corpus of Russian (GICR). This corpus consists of texts from Russian Internet which are automatically gathered and tagged. It includes a lot of text materials from the blogs, social networks, and magazines (mostly from Vkontakte, Live Journal, and Mail.Ru Blogs). Nowadays it is one of the most significant corpora, which reaches the size of 20 billion words.

GICR is an educational and scientific project, which can **serve the solving** of many linguistics tasks, such as word distribution, the influence of gender and age, frequency of words, fixed expressions or the language of the social networks studies.

Our dictionary is based on the data generously provided by the General Internet Corpus of Russian, which represents the segment of LiveJournal with automatically disambiguated homonymy. LiveJournal is one of the most prominent Russian social networking service, which is a unique combination of a community platform and social media. Users of the service can keep their blog or join communities based on shared interests. LiveJournal represents the majority of Russian blogosphere, hosting over 80 of the top 100 Russian blogs [1]. The journals address such topics as politics, entertainment, sports, fashion, subcultures, literature, and design, and provide the possibility to study the language of different social interactions and people with various lifestyles.

For our work, we used the list of words, based on the LiveJournal blog posts and comments that were not recognized by ABBYY Compreno Dictionary, and the source text file is more than 2Gb in size. This list doesn’t contain the information about the word contexts and frequency, but we are discussing adding this information to the dictionary as future lines of our project.

2.2. Data (pre)processing

As stated above, our database consists of 7 998 443 words at the moment. This is the list that we have after removing the duplicated and words that are included in the MyStem dictionaries, as MyStem dictionaries are larger than those of ABBYY Compreno. A brief description of the process is given below:

Cleaning

- removed duplicates

[1] https://www.livejournal.com/about/

- removed words that can be recognized by the Mystem parser. The Mystem parser uses several dictionaries including the Zalizniak's grammatical dictionary and their own dictionaries, however, their dictionaries are not openly distributed, therefore we cannot tell exactly what is the size of its built-in dictionary. That turned out to be complex adjectival forms of verbs, some infrequent words and some new words like 'офигеть'. Therefore, some words that are not yet in dictionaries, but are known to Mystem, were omitted from our work.

2.3. Classification

At the moment we have 703 635 forms with one or more category assigned. Examples of categories are given in Table 1 (see Appendix 1). To determine categories that we want to annotate first, we considered several factors:

- how many words approximately would fall into this category?
- is it possible to assign a category by searching for a substring (e.g., if a word has a substring 'super' it falls into a category *loan affix*) or another algorithmically simple way

To get more independent opinions on the list of categories, we used crowdsourcing. Around 16 linguistics students of the Higher School of Economics were involved. First, we chose the random list of words and classified it manually, where it was possible, and then confirmed our classification with the survey.

Table 1 (see Appendix 1) illustrates the structure that our database has at the moment. A brief description of the operations we performed on the initial data and the full list of categories are given below:

1. Morphological annotation (what we do now):
    - clear-case suffixes that show up at the end of a word like -ость (as кость, трость etc are dictionary words, they would not appear in our lists and we wouldn't mark them as having such suffixes)
    - several not so clearly determinable suffixes and non-word-final suffixes (like '-[июауеы]шк-'), with respect to the surrounding material
    - clear-case prefixes (like 'квази-')
    - several not so clearly determinable prefixes (like 'пере-'), with respect to the surrounding material
2. Categorization:
    - multiplied letters by its phonological parameters: vowels, liquids, sonorants and fricatives (like м,з,р), stops and affricates (like б,т,ц) - total of 194148 entries
    - "glued up"phrases - 10 117 entries
    - loan words - 81 915 entries
    - diminutives and augmentatives - for now around 400 000 entries, the selecting algorithm still needs improvement.
    - non-words (banging on the keyboard) - 1548 entries
3. Normalization (the initial form is also preserved):
    - deleted multiplied letters
    - (quasi)lemmatized
    - transformed capitalization and uppercase strings to lowercase
    - normalized variant of a word if we suppose it was mistyped (real normalization)
4. What we are planning to add:
    - categorization of typos and mistakes
    - categorization of feminitives
    - categorization of proper nouns

2.4. Data Lemmatization and Normalization

To enable easier search on our data, we considered (pseudo)lemmatization and normalization of our words would be of help. A (pseudo)lemma for a word from our list is a nominative singular/infinitive form of the word one of which forms would look like a word from our list if such word existed. In other words, a (pseudo)lemma for a word 'каровы' from our list would be 'карова', as this word looks like a genitive form of a word that ends in 'a'. A norm in our work is the actual word that was mistyped, so for a word 'каровы' the norm would be 'корова'.

Next we produced a pseudolemma for each word, again with Mystem, which can guess the lemma of a word it doesn't know. Mystem analyzes a given form of a word and suggests a lemma for this form based on Russian morphology rules.

Next we constructed a spell-checker using the open-source code for the module in python that constructs neighbour strings for a given string and checks whether it is known by Mystem. There were four levels of proximity: the closest was a substitution by a neighbour key or a key that denotes a phonetically close sound, the second was substitution/deletion/insertion of random letters and transposition, the next was double substitution by a neighbouring letter, and the fourth was any other operation on strings produced on the first and the second level. This way, the closest actual word to a mistyped form was found.

Each normalization was provided with the description of errors that were made to construct the given form from its normalization, so that the line for a word 'австраилию', for example, contained its norma 'австралия' and the errors 'del:и'

2.5. Annotation of word-formation structure

For our data we analyze each word and search productive word-forming affixes in it. This way, a search can be conducted on a given affix that would return all non-dictionary words (new formations, etc.) that were made by speakers using this morpheme. In the present version of the database we only annotate morphemes that do not pose too many ambiguity problems (large morphemes that are not likely to coincide with any other substrings, morphemes that allow excluding the coincidence with other substrings using a regular expression and the position of the morpheme in a word).

The annotation process is based on the search of a substring. We use self-composed lists of morphemes written down as regular expressions and search for each element of the list in a given the word. As for now, the lists of morphemes are made by us based on the list of productive morphemes by [S^{+}80](http://rusgram.narod.ru/morf1t.html) with respect to the morpheme frequency and possible ambiguity.

We perform the search and marking in three steps - from the edges of the word inwards. Each found morpheme becomes enclosed with hash symbols #, which also helps to catch roots or a suffix or prefix that happens to be between tho other suffixes or prefixes. We do not mark morphemes that consist of a single letter (e.g., *о-, а- or -н-, -а-* etc.). In case of "nested" morphemes such as suffixes *-изирова-, -ирова-, -ова-* we mark the largest one, for example in the verb *популяризировать* only the *-изирова-* suffix will be marked, therefore no non-morphemes like *-из-* or *-ир-* will be accidentally marked.

The three steps are:

1. Mark the suffixes and prefixes that never (or almost never) coincide with substrings of the words (e.g., супер-, экстра-, -мейстер, -изирова-) without considering their context.

2. Mark the morphemes that are usually on the edge of the word (e.g., *-ый, -ться, -ость or лже-, сверх-, недо-*), obviously, in the context of the edge of a word.

3. Mark the other morphemes with respect to their possible special context (i.e., special part of speech suffixes, other suffixes/prefixes, position in a word). The hash symbols that were inserted at the steps one and two play an important role here - for example, we know that prefixes may be in the very beginning of the word or between two other prefixes or between a root and another prefix. Once we have some boundaries marked, we can search for the prefix *пере-* in a context of the edge of a word or o hash symbol, so we do not get the word *оперетта*, where *пере-* is not a prefix, in our results.

We do not consider the grammar rules for Standard Russian to mark the morphemes. For example, a diminutive suffix *-чик* should not appear after stem-final velars, however it often appears after any kind of stem-final e.g., *лайкчик, блогчик* even after the consonants with very close phonological features e.g., *борщ-чик*. We only use constraints on substring search to minimize the hits where the substring of interest is not a morpheme. For example, in case of the prefix *архи-* we exclude all words where it is followed by *-в-*, therefore we do not mark this prefix in *архив* and cognate words.

2.6. Classification 2.0

We assign categories to words basing on morphemes they contain and other factors. At the moment, the morpheme-based categories are:

1. words with loan affixes, e.g., *супер-круто* or *нечит-абельный*
2. diminutives and augmentatives

We also assign following categories:

1. phrases without spaces, e.g., *тупойкомпьютер* - assignement is based
   - on the length of a word
   - on the presence in the middle of the word bigrams or trigrams that usually appear in the end of a word.
   - on the number of consecutive consonants (five and more)
2. non-words (i.e., somebody just typed with random characters) are classified basing on their length and combinations of characters that do not usually appear together or in this particular order in words of Russian language (e.g., *лр, фп, щж etc*.
3. multiplicated characters, e.g., *прррривееет*. We assign different categories to:
   - multiplicated (or elongated) vowels
   - multiplicated (or elongated) continuant consonants
   - mutiplicated stops.

2.7. Pipeline

All of our data transformations were formed into one Python script, which runs as follows. For each word, the script:

- shrinks multiplied letters, if there are any
- checks whether a word is a glued up phrase; if the answer is yes, further fields remain blank
- checks whether a given string is a nonsense string (bashing on the keyboard, a random combination of letters)
- finds productive affixes
- if all previous steps didn't give any results, normalize and lemmatize.

3. Multiplicated letters in connection with speech phonology.

3.1. Background and problem statement

The standard approach to the written mode of language is to connect it with the spoken language, and its phonology, e.g., punctuations and intonation seem to be closely interrelated. [[Cha88]] The aim of this research is to analyze the connection between words with multiplicated letters and the phonology of Russian language, especially its prosodic features. We suppose the Internet mode of language to conform the same rules as the written mode and to be influenced by such phonetic and prosodic elements like the length of sound, word and prosodic stress and intonation.

3.2. Word stress

According to Reformatskij, word stress is the stress placed on one or more syllables in a word, which is achieved by various means:

1) Dynamic or force stress is the word stress, where the prominence in a stressed syllable is achieved by the intensity of articulation. 2) Quantitative type of stress is the word stress, where the prominence in a stressed syllable is achieved by the changes in the quantity of a vowel. 3) Musical or tonic type of stress is the word stress, where the prominence in a stressed syllable is achieved by the change of musical tone compared to other syllables. A distinction is made between falling, rising and compound types of pitches. [Р□04]

Standard Russian combines the elements of the dynamic and quantitative types of stress. A stressed syllable is the most intense and long syllable in the word.

Based on this facts the multiplicated letters are supposed to be connected with the word stress and to express the length of the stressed vowel.

3.3. Prosodic stress and its connection with multiplication

Prosodic or sentence stress refers not to syllables or individual words, but to prosodic or intonation units, which are the segments of speech with a single prosodic contour. So far, many linguists consider the current concepts of the accent structure of phrase insufficient and distinguish between emphatic, tonal and lengthening accents [К□17]. Prosodic elements of language express communicative and pragmatic components of speech and can change the meaning of the whole sentence. For this reason, prosodic elements have to be expressed not only in the spoken mode of language but the written speech through punctuation and other graphic patterns [□□14].

In our research, we focus on the lengthening accent and other prosodic features, which are connected with the length of words. Our dictionary doesn't provide the information about the word context, discourse or other factors, which are significant for prosodic elements of speech. But it's important to note that multiplication can be caused not only by word stress but also by various prosodic factors, which are graphically expressed.

The most detailed description of the lengthening stress in standard Russian can be found in works of [К□96]. According to his works, there is a direct correlation between the emphatic and tonal accents and vowel length. Some combinations of two tones in the same syllable are mostly followed by a long vowel which is known as the lengthening accent:

In some cases this accent could be repeated at the end of the word, like in this sentence: – Какой-то он ничего не понима-ающи-ий, ничего не жела-ающи-ий …

The functions of quantity accent are described in another work of Kodzasov. According to Kodzasov, the length of vowel represents the remoteness, extension or atelicity of the object. The lenghting stress and it's functions are described in further detail in his works [□□00] [К□00].

According to this, we can conclude, that the vowel length tightly correlates with the word and prosodic stress. It tends to appear mostly in stressed syllables, and we expect to observe the same association in the Internet mode of language.

3.4. General results

In this section, we provide general numbers that characterize our data on words with multiplicated letters.

We have 194 148 words with multiplicated letters from the total 7 998 444 words in the database. In Table one shows how many words there are with every specific number of multiplicated letters. For example, there are 170 367 words with one multiplicated letter (e.g., Привеееет), 19 742 words with two multiplicated letters (e.g., Прииииввееет) etc.

Таблица 1: Number of multiplicated letters per word

number of multiplicated letters	number of entries (words)
1	170367
2	19742
3	3044
4	644
5	209
6	70
7	35
8	17
9	7
10	5
11	2
12	2
13	1
15	1
16	1
18	1

3.5. Pilot study results

For the goals of this pilot study, there were selected 100 random words with duplicated letters. The words, which true meaning was impossible to determinate, were replaced by other random words.

To estimate the influence of phonology on the multiplication, we distinguished 16 criteria, which are based on the theoretical works described in the previous part:

1. The coincidence between the word stress and duplicated letters.
2. Duplicated letters in the first syllable.
3. Duplicated letters in the middle syllable.
4. Duplicated letters in the last syllable.
5. Duplicated letters in two or more syllables.
6. Only one syllable in the word, stressed.
7. Duplicated letters at the beginning of the word.
8. Duplicated letters at the end of the word.
9. Long consonants.
10. Short consonants.
11. Iotized vowels.
12. Two or more duplicated letters (vowels or consonants)
13. Typing errors
14. Spelling errors
15. Phrase
16. Prephonological and phonological errors

The choice was aimed at getting a general picture of duplicated letters and their origin. All the words were examined and marked manually.

For the purpose of analyzing only words with intentional multiplication we removed three words with typing errors like “нефтедолллары” or “рассстреливать”.

The other 97 words were tested for the correspondence between the word stress and duplicated letters. Only 44 words illustrate the connection between the word stress and multiplication (“волшееееееебно”). 6 of them consist of one syllable(“гииииииииииииииииипс”), so we can’t be certain about the correspondence in these examples.

The rested words with duplicated vowels have four multiplications in the middle(”волшеееееееебно”), 7 in the first (”Аааааааафигеть”) and 16 in the last syllable (”демоныыыыы”) by typing. In addition to this fact, there are 23 reduplications at the end of the word (”вегетарианцыыыыыы”) and only 6 in the beginning(”Аааааааафигеть”). From this sample we can suggest a hypothesis that the multiplication tends to appear in the last syllable or at the end of the word, which could be tested in further researchesю

Furthermore, our estimate shows that 11 of 100 words have vowel multiplication in two or more syllables (”ГООООДООООМ”).

In attempt to find grounds for the hypothesis for the multiplication dependency of the quantitative stress we discovered 21 words with duplicated consonants. From the phonology perspective, 16 of them are long consonants (”десятьтыЩЩЩ”), and only six are short (”завтракккккккккккккк”).

13 words contain prephonolical and phonological errors (”ваааабщето”), which seem to be intentional, and four words contain de-iotized vowels (”раждЭЭЭЭнья”)

To be sure that the multiplications were made intentionally we checked the words with spelling mistakes. 79 words have no spelling errors apart from reduplications. It allows us to say that duplicated letters belong to the semantic and stylistic sphere of the Internet language.

Overall, the results presented below show that multiplications from our set are only partially connected with phonetic and prosodic elements of speech. There is no full identity with the usual written language, and the Internet mode conforms to its own rules. Multiplicated letters are not spelling errors, but the particular way to represent semantic and stylistic components of the speech graphically. It could be the ground for further researches in this field.

4. Typos types

We randomly selected 200 mistyped words that didn’t fall into any of the pre-normalization categories. Of them 137 had only one norm option, so we consider them corrected.

We expect two different types of errors: typos, which were produced due to closeness of two keys on the keyboard, and actual spelling mistakes. We suggest that typos would be more frequent in our data, as the nature of the Internet communication is fast and provides for this kind of errors.

error	frequency
’repl:о,а’	8
’repl:а,о’	8
’del:о’	7
’repl:е,и’	6
’repl:и,е’	5
’repl:о,е’	5
’repl:и,м’	4
’del:с’	4
’del:и’	3
’del:а’	3
’del:ч’	3
’del:п’	3
’repl:а,я’	3
’insert:а’	3
’del:ь’	2
’del:л’	2
’del:к’	2
’repl:е,н’	2
’del:н’	2
’repl:е,э’	2
’repl:н,г’	2
’repl:в,ы’	2
’repl:т,ь’	2
’del:ю’	2
’del:р’	2
’repl:и,ы’	2
’repl:т,и’	2
’repl:ы,я’	2
’del:у’	2
’repl:о,р’	2
’insert:ь’	2
’insert:в’	2

The most frequent typos from the sample were replacing a with o or otherwise.This is clearly a case of mistakes that have some phonetical grounds.

Список литературы

[Bar13] Liana Barseghyan. On some aspects of internet slang. *Graduate School of Foreign Languages N*, 14:19–31, 2013.

[Cha88] Wallace Chafe. Punctuation and the prosody of written language. *Written communication*, 5(4):395–426, 1988.

[Cry01] D. Crystal. *Language and the Internet.* Cambridge: Cambridge University Press, 2001.

[KB13] Victor Kuperman and Raymond Bertram. Moving spaces: Spelling alternation in english noun-noun compounds. *Language and Cognitive Processes*, 28(7):939–966, 2013.

[S^{+}80] Natalia Y Shvedova et al. Russkaja grammatika [russian grammar]. *AN SSSR Publ, Moscow*, 1980.

[К□96] СВ Кодзасов. Просодический строй русской речи. *М.: Институт русского языка РАН*, 1996.

[К□00] СВ Кодзасов. Фонетическая символика пространства (семантика долготы и краткости). *Логический анализ языка. Языки пространств. М*, pages 227–238, 2000.

[К□17] Сандро Кодзасов. *Исследования в области русской просодии.* Litres, 2017.

[Р□04] Александр Александрович Реформатский. *Введение в языковедение.* Аспект-пресс, 2004.

[□□00] Нина Давидовна Арутюнова and И Б Левонтина. *Логический анализ языка: Языки пространств*, volume 13. Языки русской культуры, 2000.

[□□14] Ирина Михайловна Кобозева and Леонид Михайлович Захаров. Просодия как ключ к пониманию смысла и ее искажение в «Кривом зеркале» пунктуации. *Филология и культура*, (2 (36)), 2014.

5. Appendix 1: database format

Таблица 2: Формат итоговой базы (На данный момент, еще будем менять NB Конечно, данных добавить, чтобы хотя бы страничка была)

исходная форма	лемма	частотность леммы	норма	морфологический разбор	категории
архимудаическое	архимудаический	-	архимудаический	-архи-муда-ическ-ое	loan affix
почувствовуют	почувствововать	-	почувствовать	-по-чув-ств-овуют	-
ууууррррмурмурмур	ууууррррмурмурмур	-	урмурмурмур	ууууррррмурмурмур	vowel lengthening
Орифлеймика	Орифлеймик	-	Орифлеймик	Орифлейм-ик-а	dim
ДОБАВИТЬ	ПРИМЕРОВ	????	-	-	-

5.1. Appendix 2: study of multiplicated letters

Таблица 3: Number of entries with multiplicated letters per letter

letters	entries
а	33336
и	21460
о	19850
е	14795
у	10130
р	9840
я	8483
ы	7186
н	6424
с	5682
х	2733
з	2399
э	2334
ж	2319
ё	2266
м	2104
ю	1915
ш	1851
ф	1694
т	1551
щ	1534
л	1529
г	1254
ц	1191
й	992
п	979
ь	906
к	905
д	890
ч	706
в	609
б	375
ъ	145