# Evaluating Russian Adj-Noun Word Sketches against Dictionaries: a Case Study

Maria Khokhlova

St Petersburg State University
Universitetskaya emb. 7-9-11
199034 St Petersburg, Russia
m.khokhlova@spbu.ru

**Abstract.** The paper discusses adj-noun word sketches produced for 20 Russian headwords. We analysed the differences between the output and collocations extracted from Russian dictionaries and also validated the collocates by expert evaluation. The aim was to study to what extent their data coincide with each other and to investigate how collocations presented in dictionaries are reflected in a large Web corpus. The comparison with the gold standard shows low precision whereas expert evaluation gives higher values. LogDice tend to extract more peculiar examples compared to joint frequency according to human assessment.

**Keywords:** Word sketches, Collocations, Evaluation, Dictionaries, Russian language.

## 1 Introduction

Lexicography and corpus linguistics become more user-oriented. Sketch Engine was one of the first systems to facilitate research by taking over most of the routine procedures [13]. The word sketches greatly influenced applied linguistics helping to represent collocational behaviour in the form of convenient tables, which until then had to be found by separate queries or with other filters. More than 20 years have passed since the appearance of the first word sketch grammar. Therefore, we see the need to re-evaluate them, analyze possible issues, and also outline ways to solve them. We currently work on creating a gold standard of collocability for the Russian language [6], which can be used as a source of reference data. We consider collocations extracted in Russian dictionaries and analyse how they are reflected in a corpus of contemporary Russian and hence are presented as word sketches. Thus, the purpose of our research is to compare word sketches with verified lexicographic data, i.e. trace the intersection between data collected by experts and automatically extracted from an up-to-date corpus.

The paper is structured as follows. The Introduction presents the basic idea of the research. The next section provides a brief overview of the related work.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2020, pp. 125–131, 2020. © Tribun EU 2020

Section 3 discusses the methods and relevant notions, i.e. the word sketch rules and gold standard used during the analysis. The next section examines the results of the experiment while the last one concludes the paper and offers future perspectives.

## 2 Background

A profound evaluation of word sketches was presented by A. Kilgarriff et al. [8] for four languages (Dutch, English, Japanese and Slovene). The authors differentiate between developer and user approach concentrating on the latter. The article [5] introduces Russian word sketches and describes Russian sketch grammar. The Russian language can be seen as underestimated in various kinds of analysis therefore the evaluation of the output of Russian word sketches may yield nontrivial results.

## 3 Methods

In the paper we also adhere to 'user evaluation' perspective [8] and will address to the dictionaries as sources of expert data.

### 3.1 Gold Standard

In contrast to the approach presented in [8] we use dictionaries as verified sources, i.e. we consider them as containing data that has already been approved by experts. As mentioned above, during our work on the gold standard of Russian collocability [6,7], we collected examples from six different Russian dictionaries. During the study we scrutinized the following ones:

1. two explanatory dictionaries, i.e. the Dictionary of the Russian Language [2]; the Large Explanatory Dictionary of the Russian Language [10];
2. three collocation dictionaries [1,11,12];
3. an online dictionary [9].

Within the analysis we considered attributive collocations built according to the "adjective / participle + noun" model. At the moment, the database includes more than 15 thousand units of such a type. The dictionary data from the gold standard is suitable for evaluating the precision of word sketches output. Since word sketches are shown for a headword, we decided to consider a number of nouns that form this type of collocations.

We analysed the most frequent headwords presented in collocations from the gold standard, namely the ones having a variety of syntagmatic relations. The 20 selected headwords turn to highly productive, i.e. form a wide range of collocations (the precise number is given in parentheses): *sila* 'force' (97), *uspekh* 'success' (59), *bor'ba* 'fight' (55), *toska* 'boredom' (54), *lyubov'* 'love' (49), *interes* 'interest' (46), *delo* 'case' (43), *bolezn'* 'illness' (42), *radost'* 'joy' (40), *pamyat'* 'memory' (40), *krasota* 'beauty' (38), *znacheniye* 'meaning' (37), *chuvstvo* 'sense' (36), *sistema* 'system' (36), *nenavist'* 'hate' (36), *um* 'intellect' (35), *strast'* 'passion' (34), *rol'* 'role' (34), *kholod* 'cold' (33), *usiliye* 'effort' (32). As one can see the majority of the nouns refer to emotions and abstract notions.

### 3.2 Experiment: Settings

As authors in [8] rightly note, it is also necessary to evaluate those collocations that could be seen as potential candidates for the inclusion in the dictionary, so we will evaluate recall in two ways: 1) by comparing to the dictionary data; 2) by an expert's assessment.

We confine ourselves to the first 50 examples produced by word sketches. From a user's perspective it is reasonable to process brief lists of collocates. Sketch Engine enables ranging word sketches according to two measures (namely, logDice and joint frequency). Since we analyse top-50 collocations, the results can differ; hence we decided to use both types of output in our evaluation. In other words it can seen as an evaluation of not only word sketches but also of two measures. Following the approach by S. Evert [3] we compute precision as proportion of collocations (identified either in the gold standard or by expert evaluation) in the output and recall as proportion of collocations from the gold standard that were correctly extracted from the corpus.

RuTenTen corpus is one of the largest Russian corpora [4], so we chose it for our experiment and expect to see the widest range of collocations extracted from it.

### 3.3 Word Sketch Grammar

Collocations based on the "adjective / participle + noun" model will be in the focus of our attention, e.g., *prakticheskoye znacheniye* 'practical meaning', *zhiznennyy uspekh* 'life success', *oslepitel'naya krasota* 'dazzling beauty', etc. In [2] we described the rules for the Russian language, which were implemented for generating word sketches. Below one can see a subset from the so called "word sketch grammar" which takes into account this type of collocations.

```
*DUAL
=amodifier/modifies
2:adj 1:noun & agree(1,2)
2:adj 3:adj 1:noun & agree(1,2) & agree(1,3)
2:adj 3:adj 4:adj 1:noun & agree(1,2) & agree(1,3) & agree(1,4)
2:adj 3:adj 4:adj 5:adj 1:noun & agree(1,2) & agree(1,3) & agree(1,4)
  & agree(1,5)
2:adj [word="  "|word=" "] 3:adj 1:noun & agree(1,2) & agree(1,3)
2:adj [word=","]? 4:adj [word="  "|word=" "] 3:adj 1:noun & agree(1,2)
  & agree(1,3) & agree(1,4)
2:adj [word=","]? 4:adj [word=","]? 5:adj [word="  "|word=" "] 3:adj
  1:noun & agree(1,2) & agree(1,3) & agree(1,4) & agree(1,5)
2:adj [word=","] 3:adj 1:noun & agree(1,2) & agree(1,3)
2:adj [word=","] 3:adj [word=","] 4:adj 1:noun & agree(1,2) & agree(1,3)
  & agree(1,4)
```

```
2:adj [word=","] 3:adj [word=","] 4:adj [word=","] 5:adj 1:noun
  & agree(1,2) & agree(1,3) & agree(1,4) & agree(1,5)
```

The Russian language is characterized by rich morphology and has a high number of inflections, therefore, an adjective or a participle must have the same gender and case with the noun to which they belong (in the above given rules this agreement is marked by 'agree' showing in parentheses the numbers of words involved in this relation). The above mentioned word sketch rules cover noun phrases and can be illustrated by the following examples:

1. adj-noun (e.g. *temperaturnyy rezhim* 'temperature regime');
2. adj-adj-noun (e.g. *vysokochastotnyy elektricheskiy tok* 'high-frequency electrical current');
3. adj-adj-adj-noun (e.g., *global'naya sputnikovaya navigatsionnaya sistema* 'global navigation satellite system');
4. adj-adj-adj-adj-noun (e.g., *kitayskiy zelenyy baykhovyy krupnolistovoy chay* 'Chinese green loose large leaf tea');
5. adj-conj-adj-noun (e.g. *mobil'nyy ili domashniy telefon* 'mobile or home phone number');
6. adj-,-adj-conj-adj-noun (e.g., *tekhnicheskaya, informatsionnaya i reklamnaya podderzhka* 'technical, information and advertising support');
7. adj-,-adj-,-adj-conj-adj-noun (e.g., *administrativnoye, pensionnoye, sotsial'noye i trudovoye zakonodatel'stvo* 'administrative, pension, social and labour law');
8. adj-,-adj-noun (e.g. *federal'nyy, regional'nyy uroven'* 'federal, regional level');
9. adj-,-adj-,-adj-noun (e.g., *neftyanaya, khimicheskaya, pischevaya promyshlennost'* 'oil, chemical, food industry');
10. adj-,-adj-,-adj-,-adj-noun (e.g., *doshkol'noye, obscheye, dopolnitel'noye, vyssheye obrazovaniye* 'preschool, general, supplementary, higher education').

These ten cases describe collocations of varying length taking into account a certain distance between nodes and collocates. Adjectives can be separated by commas or combined by conjunctions *i* 'and' and *ili* 'or'.

## 4 Results

The output showed collocates produced with morphological errors that can be accounted for several reasons. The results list token collocates belonging to the same lemmata but representing different cases, numbers or genders. For example, word sketches list *vol'nyy* 'free' (masculine gender) and *vol'naya* 'free' (feminine gender) as collocates for the lemma *bor'ba* 'fight'. This type of errors leads to the discrepancies in frequencies and hence to the false output and false ranging by both statistical measures.

For adj-noun collocations, representation of participles as verb forms can be seen as a certain problem. For example, one can find the following collocates for the headword *krasota* 'beauty': *zavorazhivat'* 'to bewitch' instead of *zavorazhivayuschiy* 'bewitching', *potrysat'* 'to amaze' instead of *potrysayuschiy* 'stunning'. However there are word sketches listing both verbs and participles as frequent collocates (e.g. *dominirovat'* 'to dominate' and *dominiruyuschiy* 'dominant' for the headword *rol'* 'role'). It could be more convenient for users (especially for Russian language learners) to see forms of the participle in the apropriate word sketch table (amodifier/modifies), i.e. *zavorazhivayuschiy* or *potrysayuschiy*.

Most of the errors in lemmatisation was found for the collocates with the headword *bolezn'* 'disease'. This can be due to the fact that they represent themselves special terms and therefore are absent in the morphological dictionary. Also a large number of incorrect results was produced for the headword *usiliye* 'effort'. The output shows verbs (instead of participles) and incorrect gender and case forms for collocates.

The total number of such errors equals to 7.4% for logDice and 4% for joint frequency respectively.

LogDice tend to extract more peculiar collocations. For example, among the first 50 results we found *tsepkaya pamyat'* 'tenacious memory' and *fotograficheskaya pamyat'* 'photographic memory', i.e. these collocations can be listed in entries of dictionaries for Russian language learners. The joint frequency measure yield yet less promising collocations among the top 50 ones.

Table 1 shows the results for the precision and recall computed when compared with the gold standard and expert assessment. The least number of the examples were found for the headword *pamyat'* 'memory'. This can be due to the fact that the corpus contains contemporary texts showing mostly occurrences for this noun with the meaning "computer memory" while dictionaries list examples for other meanings.

**Table 1.** Precision and recall.

| Headword | | Precision (dictionary, logDice) | Precision (expert, logDice) | Precision (dictionary, freq) | Precision (expert, freq) | Recall (dictionary, logDice) | Recall (dictionary, freq) |
|---|---|---|---|---|---|---|---|
| bolezn' | 'illness' | 0.48 | 0.88 | 0.52 | 0.84 | 0.57 | 0.62 |
| bor'ba | 'fight' | 0.38 | 0.54 | 0.42 | 0.54 | 0.35 | 0.38 |
| chuvstvo | 'sense' | 0.16 | 0.24 | 0.28 | 0.30 | 0.22 | 0.39 |
| delo | 'case' | 0.18 | 0.36 | 0.22 | 0.36 | 0.21 | 0.26 |
| interes | 'interest' | 0.28 | 0.46 | 0.24 | 0.36 | 0.30 | 0.26 |
| kholod | 'cold' | 0.42 | 0.50 | 0.36 | 0.42 | 0.64 | 0.55 |
| krasota | 'beauty' | 0.36 | 0.56 | 0.36 | 0.44 | 0.47 | 0.47 |
| lyubov' | 'love' | 0.32 | 0.66 | 0.30 | 0.56 | 0.33 | 0.30 |
| nenavist' | 'hate' | 0.32 | 0.38 | 0.38 | 0.44 | 0.44 | 0.53 |
| pamyat' | 'memory' | 0.20 | 0.58 | 0.18 | 0.42 | 0.25 | 0.23 |
| radost' | 'joy' | 0.38 | 0.46 | 0.36 | 0.38 | 0.48 | 0.45 |
| rol' | 'role' | 0.48 | 0.56 | 0.44 | 0.50 | 0.71 | 0.65 |
| sila | 'force' | 0.46 | 0.84 | 0.44 | 0.70 | 0.24 | 0.23 |
| sistema | 'system' | 0.08 | 0.40 | 0.08 | 0.36 | 0.11 | 0.11 |
| strast' | 'passion' | 0.32 | 0.42 | 0.36 | 0.42 | 0.47 | 0.53 |
| toska | 'boredom' | 0.40 | 0.44 | 0.50 | 0.54 | 0.37 | 0.46 |
| um | 'intellect' | 0.38 | 0.52 | 0.34 | 0.38 | 0.51 | 0.49 |
| usiliye | 'effort' | 0.18 | 0.34 | 0.24 | 0.28 | 0.28 | 0.38 |
| uspekh | 'success' | 0.56 | 0.64 | 0.40 | 0.52 | 0.47 | 0.34 |
| znacheniye | 'meaning' | 0.28 | 0.40 | 0.36 | 0.02 | 0.38 | 0.49 |

The mean precision computed against gold standard was quite low and equal to 0.33 and 0.34 for logDice and joint frequency respectively. The expert analysis revealed fascinating collocates among word sketches and hence raised the mean precision to 0.51 and 0.44 respectively. LogDice showed again more interesting results according to human assessment compared to joint frequency (e.g. *kriticheskoye znacheniye* 'critical value' or *shkurnyy interes* 'selfish interest').

## 5 Conclusion

We examined word sketches for 20 nouns that form the largest number of collocations according to six Russian dictionaries. Thus, this formal evaluation was based on a comparison between corpus and lexicographic data. In total 1,000 word sketches per measure (logDice and joint frequency) were analyzed. The analysis showed that the precision of the word sketches output is a bit low with regard to the data extracted from different Russian dictionaries while they show higher and more promising results assessed by expert evaluation. At least half of the produced word sketches can be called "true collocations" and can be included into dictionaries (that do not list them yet) and here we can foresee broad perspectives.

Although logDice measure shows quite similar quantitative results with joint frequency, however, it turns out to be much more successful for extracting and ranking word sketches according to the expert assessment. This confirms the choice of this measure as the default one in Sketch Engine. These results can be also relevant for further evaluation of statistical measures used for collocation extraction. In future we plan to evaluate other models described in the word sketch grammar and analyse more headwords.

**Acknowledgments.** This work was supported by the grant of the Russian Science Foundation (Project No. 19-78-00091).

## References

1. Borisova, E.: A Word in a Text. A Dictionary of Russian Collocations with English-Russian Dictionary of Keywords [Slovo v tekste. Slovar' kollokatsiy (ustoychivykh sochetaniy) russkogo yazyka s anglo-russkim slovarem klyuchevykh slov]. Filologiya: Moscow (1995).
2. The Dictionary of the Russian Language [Slovar' russkogo jazyka v 4 tomakh]. Yevgen'yeva A. P. (ed.-in-chief). Vol. 1–4, 2nd edition, revised and supplemented. Russkij jazyk: Moscow (1981–1984).
3. Evert, S.: Corpora and collocations. In: Corpus Linguistics. An International Handbook 2. pp. 1212-–1248. (2008)
4. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. In: Proceedings of the 7th International Corpus Linguistics Conference CL 2013, the United Kingdom, July 2013, pp. 125–127 (2013).

5. Khokhlova, M.: Applying Word Sketches to Russian. In: Proceedings of Raslan 2009. Recent Advances in Slavonic Natural Language Processing, pp. 91–99. Masaryk University: Brno (2009)
6. Khokhlova, M.: Building a Gold Standard for a Russian Collocations Database. In: Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, pp. 863—869. Ljubljana (2018)
7. Khokhlova, M.: Collocations in Russian Lexicography and Russian Collocations Database. In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France, pp. 3191–3199. European Language Resources Association (2020)
8. Kilgarriff, A., Kovar, V., Krek, S., Srdanovic, I., and Tiberius, C.: A quantitative evaluation of word sketches. In: Proceedings of the XIV Euralex International Congress, pp. 372–379. Fryske Academy: Leeuwarden (2010)
9. Kustova, G.: Dictionary of Russian Idiomatic Expressions [Slovar′ russkoyj idiomatiki. Sochetaniya slov so znacheniyem vysokoy stepeni] (2008), `http://dict.ruslang.ru`. Last accessed 14 Nov 2020.
10. The Large Explanatory Dictionary of the Russian Language [Bol′shoy tolkovyy slovar′ russkogo yazyka]. S.A. Kuznetsov (ed.). Norint: St. Petersburg (1998).
11. Oubine, I.: Dictionary of Russian and English Lexical Intensifiers [Slovar′ usilitel′nykh slovoso-chetaniy russkogo I angliyskogo yazykov]. Russian Language: Moscow (1987).
12. Reginina, K., Tjurina, G., Shirokova, L.: Set Expressions of the Russian Language. A Reference Book for Foreign Students [Ustoychivye slovosochetaniya russkogo yazyka: Uchebnoye posobiye dlya studentov-inostrantsev]. Shirokova, L. I. (ed.). Moscow (1980).
13. Sketch Engine Homepage, `http://www.sketchengine.eu`. Last accessed 14 Nov 2020