Applying Word Sketches to Russian*

Maria Khokhlova[1,2]

[1] Faculty of Philology and Arts, St. Petersburg University, Russia
[2] Institute for Linguistic Studies, St. Petersburg, Russia

Abstract. The paper describes work on writing a Russian Sketch grammar for the system Sketch Engine. The objective of such a system is to provide lexicographers with sufficient lexical material and tools for getting information about a word's collocability and to generate lists of the most frequent phrases for a given word, and then to classify them for appropriate syntactic models. The system will give information about a word's collocability on concrete dependency models, and will generate lists of the most frequent phrases for a given word for various grammatical models.

Key words: Word Sketches; Russian

1 Introduction

The system known as Sketch Engine was developed by British and Czech scholars (A. Kilgarriff, P. Rychlý, H. Pomikálek; [1]). The Sketch Engine combines approaches of both traditional linguistics (e.g. syntactic models) and statistics. It is widely used by scholars when compiling grammars and dictionaries (Oxford University Press, Cambridge University Press, Collins, Macmillan etc.). It was developed for a number of languages (English, Irish, Spanish, Italian, German, Portuguese, Slovene, French, Czech, Chinese, Japanese). However, there is no such a system for the Russian language. Sketch Engine is a corpus tool which takes as input a corpus of any language and corresponding grammar patterns and which generates word sketches for words of that language. Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour [2,3]. One can understand word sketches as typical phrases determined on the one hand by syntax that restricts words' collocability in a given language and on the other hand by probability closely related to word usage.

2 Methods of Corpus Linguistics and Collocations

Corpora are vital tools for linguistic studies and solution for applied tasks. The application of corpora methods to the analysis of lexical collocability enables

* This project was supported by the Russian Humanitarian Scientific Fund (grant number # 07-04-00161a).

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, pp. 91–99, 2009. © Masaryk University, Brno 2009

to write grammars and compile dictionaries of a new type, dictionaries of collocations, idioms etc. The issue of collocability is highly important in modern linguistics. The investigation of collocability is closely connected to the study of syntagmatics as a deeper level of lexical relations. With arrival of text corpora and corpus linguistics lexicographers and other linguists have gained an opportunity to look at big collections of word usage. Corpora not only help to study lexical units in context but also to get data on word frequency, frequency of lexemes, grammatical categories, their collocability etc.

Although the above mentioned corpora opportunities are very useful, there is a need of another kind of software for further improvement of linguistic research as it is impossible to process huge amount of linguistic data manually. It can be described as an additional system between a corpus and its users (linguists) which can process significant language data.

The problem of syntagmatic relations is one of the most notorious in linguistics. There are various concepts of collocation and ways of how to extract collocations. Statistical methods for data treatment are widely used in corpus linguistics. Our intention is to study statistical methods of collocation extraction in comparison with the traditional (semantic) methods. That's why we chose the system Sketch Engine as a platform for implementing this task. Other software for processing corpus data (various corpus managers etc.) does not provide such features.

Nowadays there are several ways in statistics to calculate coherence of collocation parts, to highlight the most important ones. There are different measures based on calculation of words' "closeness" in a text, namely, MI (mutual information), t-score, log-likelihood, z-score, chi-square. They are based on comparison of frequencies registered for pairs of words in a real corpus material with independent (relative) frequencies. And statistically significant deviations of real frequencies from hypothetical probabilities are being searched. But formulas for different measures more often than not produce elevated numbers for word frequency, length of word window etc. As a result, they extract not only set phrases but free phrases as well as lexical items of the same semantic fields. The association measures do not take into account grammatical relations between tokens either. Besides, the statistical methods give significant results when they are based on representative corpora. Thus it is a need in such corpora that often lack.

3 Building Syntactic Models of Phrases in Russian

3.1 Corpus Building

The first preparatory stage of the project consisted in collecting texts to build a corpus of Russian. Originally we had a test corpus of letters of N.V. Gogol' [4], a famous Russian writer (1809–1852). This corpus contained about 0.5 mln tokens. As far as we know there isn't any work on extracting collocations on such a material (Russian texts of the XIXth century). The Russian language of the XIXth

century is notable for syntactic constructions that are different from modern ones. During this work (described in [5]) we have shown that methods presented can be effectively used for studying the authors' language and writing authors' dictionaries, for revealing collocability of words in different styles or within the given time period.

Afterwards we decided to make a number of corpora that reflect various language styles. They are fiction (about 10 mln tokens), scientific texts (about 0.5 mln tokens), news (about 5 mln tokens; journalistic genre), and texts of "common" style from the Internet (subcorpus of 10 mln tokens, this only corpus was compiled by S.A. Sharoff). This proportion can be seen as a strange one but we speak only about first steps in this project. Further work will be done on increasing corpora (their volume and number). This choice was motivated by a number of reasons. First of all to obtain better results we need to have quite similar texts (time period, genre etc.). Secondly, texts should be homogeneous (inside one corpus), have similar structure to give more statistical "weight" to its phrases (as their probability will be higher). The issue of corpus composition is a crucial one in linguistics, but we do not intend to discuss it here for lack of space and it wasn't our goal to compile corpora in this "narrow" scientific sense.

Then these texts were uploaded to the Sketch Engine where they were automatically processed and morphologically lemmatized and annotated by the program TreeTagger [6]. The Sketch Engine input format, often called "vertical" or "word-per-line", is as defined at the University of Stuttgart in the 1990s and widely used in the corpus linguistics community. Each token (e.g., word or punctuation mark) is on a separate line and where there are associated fields of information, typically the lemma and a POS-tag; they are included in tab-separated fields. Structural information, such as document beginnings and ends, sentence and paragraph mark-up, and meta-information such as the author, title and date of the document, its region and its text type, are presented in XML-like form on separate lines [7].

3.2 Word Sketch Grammar

The Sketch Engine needs to know how to select words that are connected by grammatical relations, i.e. that can be possibly collocations. That's why a scholar has to write a set of rules that describe grammatical relations that exist between words (word pairs). Strictly speaking, grammatical relations are defined as regular expressions over part-of-speech tagging.

During the second stage we investigated various sets of rules for different languages (English, Czech, Slovak etc), made a comparison of differences in the Russian and Czech syntax relevant to word sketches and then wrote grammatical rules that take into account syntactic constructions of the Russian language based on the morphologically tagged corpus in terms of grammar of Sketch Engine. This grammar represents itself a collection of definitions that allow the system to automatically identify possible relations of words to the keyword. On the basis of these rules and statistical measures it generates tables with word sketches for a keyword.

While writing rules we used regular expressions and query language IMS Corpus Workbench. The system searches for tags which correspond to word forms. For example, tag *Ncfpnn* means common noun *(Nc)* female gender *(f)* plural *(p)* noun case *(n)*: «Эти /P---pn/ этот перспективы /Ncfpnn/ перспектива и /C/ исвязаны/Afp-p-s/ связанный». After slashes there are a POS-tag and lemma. Below there is an example of grammatical rules for the phrases *"adjective+noun"*:

```
*DUAL
=a_modifier/modifies
  2:"A....n." (([word=","]|[word="и"]|[word="или"]) [tag="A....n."])0,3 1:"N...n."
  2:"A....g." (([word=","]|[word="и"]|[word="или"]) [tag="A....g."])0,3 1:"N...g."
  2:"A....d." (([word=","]|[word="и"]|[word="или"]) [tag="A....d."])0,3 1:"N...d."
  2:"A....a." (([word=","]|[word="и"]|[word="или"]) [tag="A....a."])0,3 1:"N...a."
  2:"A....i." (([word=","]|[word="и"]|[word="или"]) [tag="A....i."])0,3 1:"N...i."
  2:"A....l." (([word=","]|[word="и"]|[word="или"]) [tag="A....j."])0,3 1:"N...l."
```

Above mentioned rules take into account all such phrases, e.g. nouns and adjectives in the same case with conjunctions «и» ("and"), «или» ("or"), comma or adjectives between them within the distance of 3 words. The numeral 1 stands for a keyword (for instance, *1:"N...n."*) and the numeral 2 indicates a collocate (for instance, *2:"A...n."*). For example, «лучшие /Afp-pnf/ хороший помощники /Ncmpny/ помощник», «печатный /Afpmsnf/ печатный текст /Ncmsnn/ текст», «яркие /Afp-pnf/ яркий мысли/Ncfpnn/ мысль», «сегодняшний /Afpmsaf/ сегодняшний день /Ncmsan/ день», «благоприятные /Afp-paf/ благоприятный условия /Ncnpan/ условие», «потенциальным /Afp-pdf/ потенциальный возможностям /Ncfpdn/ возможность», «стандартным/Afp-pdf/ стандартный кредитам /Ncmpdn/ кредит». Here are several examples of relations between words: =subject/subject_of («собака лает» / "the dog is barking") =object/object_of («принять решение» / "make a decision") =a_modifier/modifies («крепкий чай» / "strong tea") Originally these rules were written on the basis of existing rules for English and Czech [3]. Then we have written the second variant of word sketches rules within the approach of Vladimir Benko (oral paper presented at Mondilex workshop in Bratislava, April 2009) [8] for the Slovak National Corpus [9]. Its distinctive feature is that these rules describe all phrases found in a corpus. For example, "verb + any word" (see below):

```
=Verb X/X Verb
    2:[tag="V.*"] 1:[tag!="SENT"]
    1:[tag!="SENT"] 2:[tag="V.*"]
```

The second line means that there will be found all phrases for any word (if it isn't a punctuation mark that has its own tag in the corpus) with a verb. The rule in the third line describes the same phrases but a verb is to the right of a keyword.

It should be remarked that this approach has its advantage as word sketches are generated for any word (because very often morphological ambiguity or mistakes of automatic tagging prevent from giving objective results).

In the theory of information retrieval there are two notions – "precision" and "recall". Precision means the percentage of documents returned that are relevant, i.e. in case of words it's the percentage of correct collocations compared to all phrases given. Recall is the fraction of the documents that are relevant to the query (that are successfully retrieved), i.e. the fraction correct collocations between all the collocations. Let's consider the following example. If our word sketch for *"tea"* contains only *"strong"* and *"green"*, it has 100% precision, since all the collocates given are correct, but low recall, since there are many other collocates it does not give. Using these terms we can say that the first approach (the first variant of rules) gives higher precision while the second one higher recall.

3.3 Word Sketch Tables

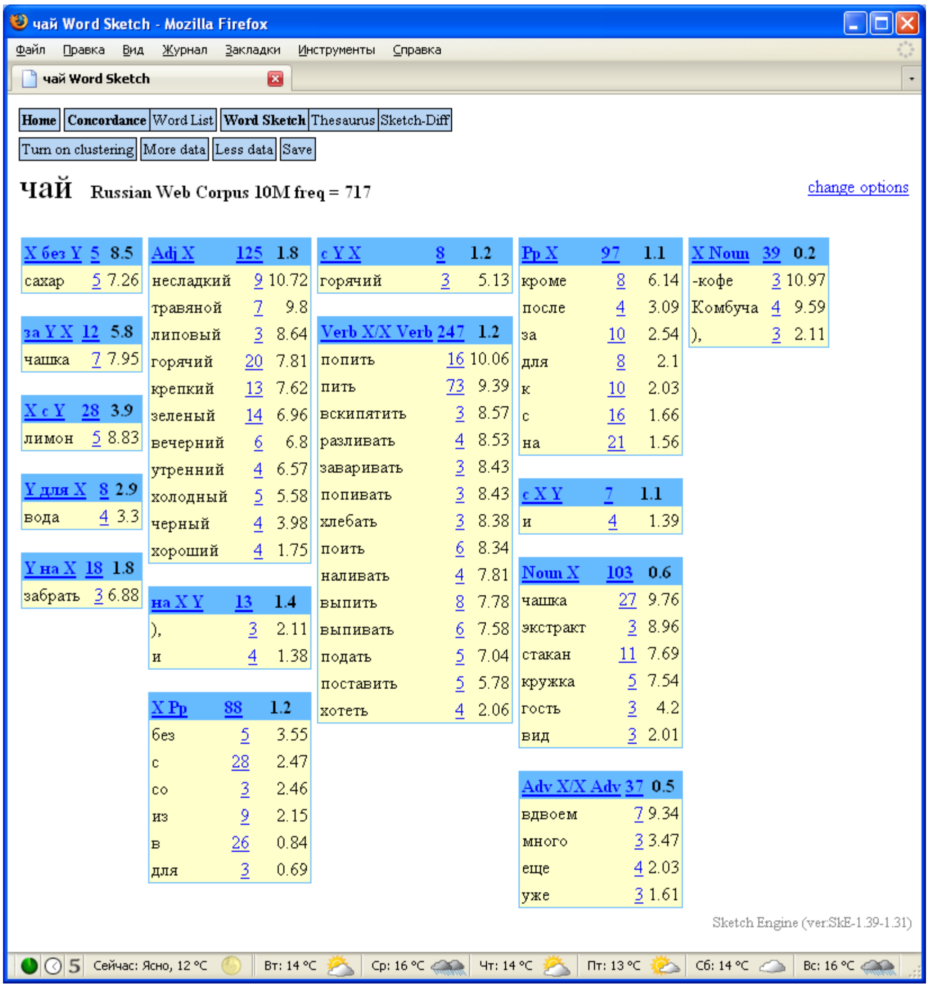
Table 1 shows word sketch for the Russian word «чай» ("tea"). The blue heading of each small table has the name of the grammatical relation between words. X stands for the keyword, whereas Y signifies a collocate. In the column *"Adj X"* (the model "adjective + keyword") we find typical qualifying adjectives (that can be applied to other nouns too), several set phrases, and also terms (they are true for English too): «несладкий» ("non-sweet"), «травяной» ("herbal"), «липовый» ("lime leaf"), «горячий» ("hot"), «крепкий» ("hot"), «зеленый» ("green"), «вечерний» ("evening"), «утренний» ("morning"), «холодный» ("cold"), «черный» ("black"), «хороший» ("good"). As for the column "Verb X/X Verb" (the model "verb + keyword / keyword + verb") here we also find collocates that are inherent for the word "tea" in Russian. They are «пить» or «попить» ("to drink"), «вскипятить» ("to boil"), «разливать» ("to pour"), «заваривать» ("to brew"), «хлебать» ("to gulp"), «подать» ("to serve"). The user can choose various options for the display of the word sketches. Collocates can be ranked according to the raw frequency of the collocation, or according to its salience score [10]. The user can set a frequency threshold so low-frequency collocations are not shown, or click a button for "more data" or "less data". They can go to the related concordance by clicking on the hit-count for a collocation.

3.4 Word Sketch Differences

Once the word sketch grammar is written this information is used in other Sketch Engine feature, namely, Word Sketch Differences. This feature shows for two semantically related words their behaviour (what they do have in common and in what differ). This information is presented in the form of multicolored diagrams. Such summary offers both common collocates that share the comparing pair and also collocates that are inherent only for one word in this pair. Synonymous words tend to share some of the collocates but not all.

Table 2 shows word sketch differences for the Russian words «большой» ("big") and «крупный» ("large"); the number of tokens for «большой» is 7593, for «крупный» is 1997. The compared two words are on each end of the multicolored scale. The yellow color shows common collocates (as we can see this part is the biggest one), the green one denotes collocates for the word «большой», and the pink one indicates collocates for the word «крупный». Each table has

Table 1. Word sketch for the Russian word «чай» ("tea")


чай Word Sketch - Mozilla Firefox
Файл Правка Вид Журнал Закладки Инструменты Справка
чай Word Sketch
Home Concordance Word List Word Sketch Thesaurus Sketch-Diff
Turn on clustering More data Less data Save
change options

чай Russian Web Corpus 10M freq = 717

X без Y	5	8.5
сахар	5	7.26

за Y X	12	5.8
чашка	7	7.95

X с Y	28	3.9
лимон	5	8.83

Y для X	8	2.9
вода	4	3.3

Y на X	18	1.8
забрать	3	6.88

Adj X	125	1.8
несладкий	9	10.72
травяной	7	9.8
липовый	3	8.64
горячий	20	7.81
крепкий	13	7.62
зеленый	14	6.96
вечерний	6	6.8
утренний	4	6.57
холодный	5	5.58
черный	4	3.98
хороший	4	1.75

на X Y	13	1.4
),	3	2.11
и	4	1.38

X Pp	88	1.2
без	5	3.55
с	28	2.47
со	3	2.46
из	9	2.15
в	26	0.84
для	3	0.69

с Y X	8	1.2
горячий	3	5.13

Verb X/X Verb	247	1.2
попить	16	10.06
пить	73	9.39
вскипятить	3	8.57
разливать	4	8.53
заваривать	3	8.43
попивать	3	8.43
хлебать	3	8.38
поить	6	8.34
наливать	4	7.81
выпить	8	7.78
выпивать	6	7.58
подать	5	7.04
поставить	5	5.78
хотеть	4	2.06

Pp X	97	1.1
кроме	8	6.14
после	4	3.09
за	10	2.54
для	8	2.1
к	10	2.03
с	16	1.66
на	21	1.56

с X Y	7	1.1
и	4	1.39

Noun X	103	0.6
чашка	27	9.76
экстракт	3	8.96
стакан	11	7.69
кружка	5	7.54
гость	3	4.2
вид	3	2.01

Adv X/X Adv	37	0.5
вдвоем	7	9.34
много	3	3.47
еще	4	2.03
уже	3	1.61

X Noun	39	0.2
-кофе	3	10.97
Комбуча	4	9.59
),	3	2.11

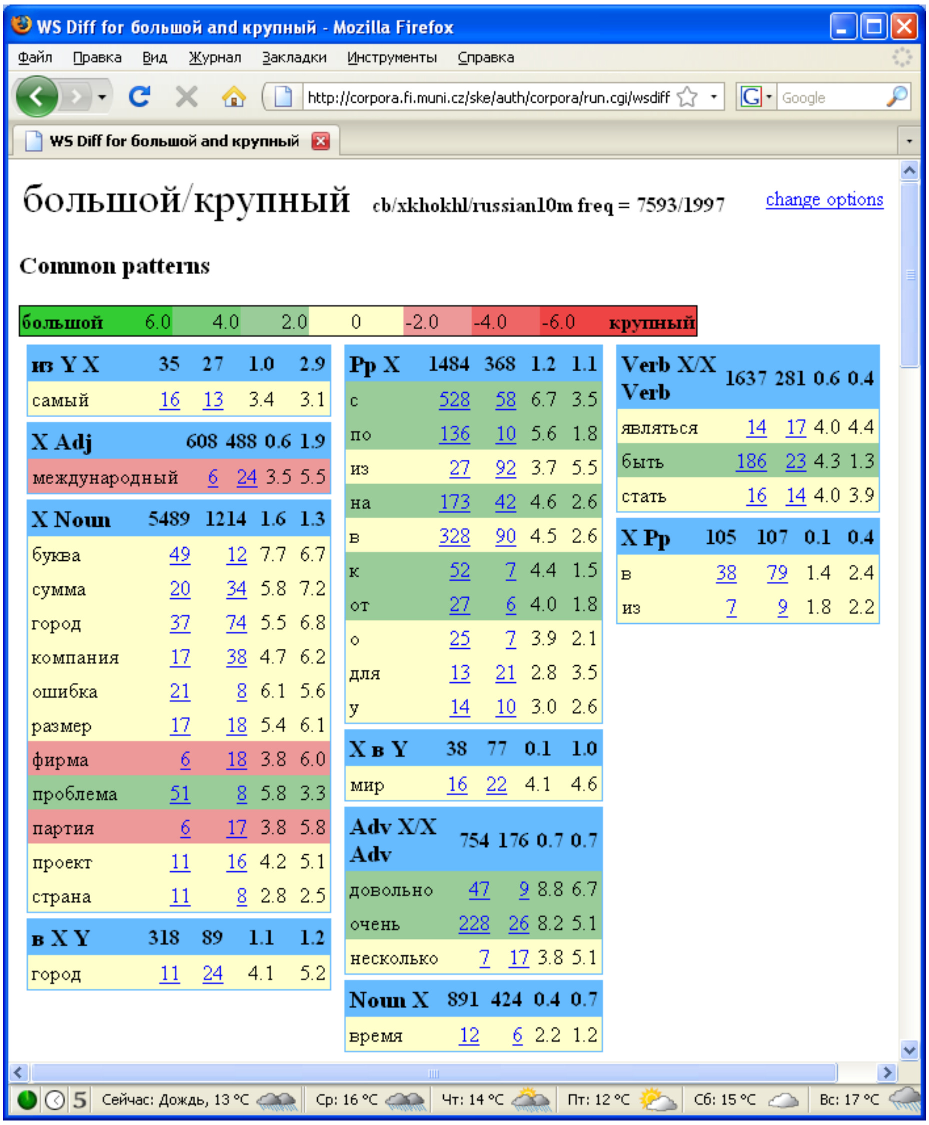
Sketch Engine (ver:SkE-1.39-1.31)

five columns: a collocate, a collocate's frequency for the first word, a collocate's frequency for the second word, and statistical measures (in this case it's salience, computed for the collocate and the word).

4 Results

There is a question of corpus volume. For example, we know that different association measures extract different collocations but here one can't see differences between results obtained by a number of statistical measures, it means that collocates will be quite the same. This problem arises from low

Table 2. Word sketch differences for the Russian words «большой» (“big”) and «крупный» (“large”)


WS Diff for большой and крупный - Mozilla Firefox
Файл Правка Вид Журнал Закладки Инструменты Справка
http://corpora.fi.muni.cz/ske/auth/corpora/run.cgi/wsdiff
Google
WS Diff for большой and крупный
5 Сейчас: Дождь, 13 °C Ср: 16 °C Чт: 14 °C Пт: 12 °C Сб: 15 °C Вс: 17 °C

большой/крупный cb/xkhokhl/russian10m freq = 7593/1997 change options

Common patterns

большой	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	крупный

из Y X	35	27	1.0	2.9
самый	16	13	3.4	3.1

X Adj	608	488	0.6	1.9
международный	6	24	3.5	5.5

X Noun	5489	1214	1.6	1.3
буква	49	12	7.7	6.7
сумма	20	34	5.8	7.2
город	37	74	5.5	6.8
компания	17	38	4.7	6.2
ошибка	21	8	6.1	5.6
размер	17	18	5.4	6.1
фирма	6	18	3.8	6.0
проблема	51	8	5.8	3.3
партия	6	17	3.8	5.8
проект	11	16	4.2	5.1
страна	11	8	2.8	2.5

в X Y	318	89	1.1	1.2
город	11	24	4.1	5.2

Pp X	1484	368	1.2	1.1
с	528	58	6.7	3.5
по	136	10	5.6	1.8
из	27	92	3.7	5.5
на	173	42	4.6	2.6
в	328	90	4.5	2.6
к	52	7	4.4	1.5
от	27	6	4.0	1.8
о	25	7	3.9	2.1
для	13	21	2.8	3.5
у	14	10	3.0	2.6

X в Y	38	77	0.1	1.0
мир	16	22	4.1	4.6

Adv X/X Adv	754	176	0.7	0.7
довольно	47	9	8.8	6.7
очень	228	26	8.2	5.1
несколько	7	17	3.8	5.1

Noun X	891	424	0.4	0.7
время	12	6	2.2	1.2

Verb X/X Verb	1637	281	0.6	0.4
являться	14	17	4.0	4.4
быть	186	23	4.3	1.3
стать	16	14	4.0	3.9

X Pp	105	107	0.1	0.4
в	38	79	1.4	2.4
из	7	9	1.8	2.2

frequencies of words and phrases. As was pointed above we are going to work on further corpus data increase.

A number of problems arise from errors in morphological annotation as: 1) every punctuation mark has its own tag (so it should be excluded in the sketch grammar); 2) parts of compound nouns also have different lemmata that is why in sketch tables we can find only one part of such words as a collocate; 3) usual mistakes of annotation, e.g. homonyms or homographs, mistakes in assigning the correct case or number; 4) mistakes in assigning correct lemmata (it is especially the case while annotating texts of the last centuries or, vice versa, of modern period with lots of neologisms).

The evaluation of the results obtained suggests that the word sketch mechanism is a useful tool for selecting the most significant collocations that are often not presented in dictionaries.

5 Conclusion

We believe that the present project may contribute to the theoretical studies of the Russian language (at the borderland between lexicography and syntax) as well as to the solution of a number of practical issues.

Further development of this mechanism of collocation extraction is closely related to writing more exact grammatical rules (that will be based on syntactically parsed corpus), more corpus data etc. Most errors in the word sketches result from errors in lemmatisation and POS-tagging. We are currently explore alternative tools for automatic morphological annotation. Manual morphological disambiguation can be seen as a possible solution for the problem of reducing errors of annotation. But this work is labour- and time-consuming and unfortunately can be applied only to a small part of a corpus.

Also there is a question of further sketch grammar improvement. New variant of the sketch grammar should be based on compilation of various grammars of the Russian language (Russian Academy Grammar [11] etc.).

The results of the research project are of practical value, as the information about a word's collocability is not often reflected in dictionaries and other reference books. The data about words' syntagmatic behaviour may find an extensive use in various fields of linguistics, such as in: dictionary compiling, language learning and teaching, translation (including machine translation), phraseology, information retrieval etc.

Acknowledgement I cordially thank my friends and colleagues from the NLP laboratory of the Masaryk University (Brno, Czech Republic) for their help and support.

References

1. Sketch Engine project: `http://www.sketchengine.co.uk`
2. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D. (2004). The Sketch Engine. In: Proceedings of EURALEX-2004, 105–116.
3. Rychlý, P., Smrž, P. Manatee, Bonito and Word Sketches for Czech. (2004). In: Trudy mezhdunarodnoy konferentsii "Korpusnaja lingvistika-2004": Sbornik dokladov. St.-Petersburg, 324–334.

4. Gogol, N.V. Polnoye sobraniye sochineniy: [V 14 t.]. (1937–1952). Moscow – St.-Petersburg, 1937–1952. T. X-XIV.
5. Khokhlova, M., Zakharov, V. Corpus-based analysis of lexico-grammatical patterns (on the corpus of letters of N.V. Gogol). (2009). In: Proceedings of the Fifth International Conference "Computer Treatment of Slavic and East European Languages", Bratislava, Slovakia, 25–27 November 2009. Bratislava. (in print)
6. TreeTagger: `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`
7. Documentation to the Sketch Engine: `http://trac.sketchengine.co.uk`
8. Benko, V. Word Sketches for the Slovak National Corpus [Oral presentation at Mondilex workshop]:
   `http://korpus.juls.savba.sk/~mondilex/programme3.pdf`
9. Slovak National Corpus: `http://korpus.sk`
10. Rychlý, P. A Lexicographer-Friendly Association Score. (2008) In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008. Brno, 6–9.
11. Russkaja grammatika. (1980) Toma I, II. Moscow. (AG-80).