Studying Word Sketches for Russian

Maria Khokhlova[1,2], Victor Zakharov[1,2]
[1] St. Petersburg State University
Universitetskaya nab., 11, 199056 St. Petersburg, Russia
[2] Institute for Linguistic Studies
Tuchkov per., 9, 199053 St. Petersburg, Russia
khokhlova.marie@gmail.com, vz1311@yandex.ru

Abstract

Without any doubt corpora are vital tools for linguistic studies and solution for applied tasks. Although 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. The Sketch Engine representing itself a corpus tool which takes as input a corpus of any language and corresponding grammar patterns. The paper describes the writing of Sketch grammar for the Russian language as a part of the Sketch Engine system. The system gives information about a word's collocability on concrete dependency models, and generates lists of the most frequent phrases for a given word based on appropriate models. The paper deals with two different approaches to writing rules for the grammar, based on morphological information, and also with applying word sketches to the Russian language. The data evidences that such results may find an extensive use in various fields of linguistics, such as dictionary compiling, language learning and teaching, translation (including machine translation), phraseology, information retrieval etc.

1. Introduction

The present paper describes our work on developing a word sketch grammar for the Russian language. Its purpose is to work out a system of statistical and syntactic patterns (models of phrases or sketches) for the Russian language based on a morphologically annotated corpus.
The objective of such a system is to provide lexicographers with sufficient lexical material and tools for getting information about a word's collocability. The system will generate lists of the most frequent phrases for a given word for various grammatical models.

2. Methods of Corpus Linguistics and Collocations

Corpora are vital tools for linguistic studies and solution for applied tasks. The application of methods of corpus linguistics to the analysis of lexical collocability enables to write grammars and compile common and specialized dictionaries of a new type. Statistical and probabilistic approach plays a significant role in this methodology.
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.
Although corpora opportunities are very useful, there is a need of another kind of software along with corpus managers 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, filter the results, and facilitate the postprocessing of received output data. 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.

3. Sketch Engine

Such a system known as Sketch Engine was developed by British and Czech scholars (A. Kilgarriff, P. Rychly, H. Pomikalek). 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 has not been such a system for the Russian language until recently (strictly speaking the system itself processed Russian texts but without the word sketch module).
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 corpus-based summaries of a word's grammatical and collocational behaviour (Kilgarriff et al., 2004; Rychly et al., 2004). One can understand word sketches as typical phrases determined on the one hand by syntax that restricts words' combinations in a given language and on the other hand by probability closely related to semantics and/or word usage.
Our task was to analyze so-called word sketch grammars for other languages and to develop the similar rules for Russian.

4. Word Sketches for Russian

4.1 Corpus Building

We have built 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 55 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). Then these texts were automatically processed and morphologically lemmatized and annotated by the program TreeTagger.

But we realize, and previous experiments confirm this, that to reach reliable and authentic results the volume of corpora has to be bigger. The choice of different types of texts was motivated by a number of reasons. First of all, we must have texts of different genres and subject areas in order to study different levels of lexis and word usage. For example scientific texts show quite a strict word order and a set of grammatical patterns (cf. *the paper deals with*, *the evidence shows* etc.) Secondly, to obtain better results we need to have quite a large amount of similar texts (time period, genre etc.). Thus, texts should be homogeneous (inside a corpus), have similar structure to give more statistical "weight" to its set phrases (as their probability will be higher). For the time being we don't intend to build a representative corpus of Russian as it's a task in itself. Further work will be done on increasing corpora (their volume and number). In this paper we discuss results of applying word sketch rules to the corpus of newspapers (journalistic genre).

4.2 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) in a language. Strictly speaking, here grammatical relations are defined as regular expressions over part-of-speech tagging.

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/и связаны /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....l."]){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."*).

Here are several examples of relations between words:

=symmetric (*«старики, дети и инвалиды»* / "old people, children and disabled people")

=subject/subject_of (*«собака лает»* / "the dog is barking");

=object/object_of (*«принять решение»* / "make a decision");

=a_modifier/modifies (*«крепкий чай»* / "strong tea");

=inst_modifier/inst_modifies (*«убил ножом»* / "killed with a knife")

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 word sketch grammar. This grammar itself represents a collection of definitions that allow the system to automatically identify possible relations of words to the keyword. Taking into account these rules the system selects predefined types of phrases and then on the basis of statistical measures it generates tables with word sketches for a keyword sorted according to the selected associative measure.

Originally these rules were written on the basis of existing rules for English and Czech (Rychly et al., 2004).

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) (Benko, 2009) for the Slovak National Corpus.

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.

4.3 Word Sketch Tables

The user can choose various options to display of the word sketches. Collocates can be ranked according to the raw frequency of the collocation, or according to its salience score (Rychly, 2008). 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.

Fig. 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. A table has two columns of numbers – the former indicates the frequency of the given collocation while the latter means the score of a statistical measure for this collocation.

In the column "*Adj X"* (the model "adjective + keyword") we find typical qualifying adjectives (that can be applied to other nouns too): *«галлюциногенный»* ("hallucinogenic"), *«сладкий»* ("sweet"), *«горячий»* ("hot"), *«холодный»* ("cold"), *«качественный»* ("high quality"), «тёплый» ("warm"), «традиционный» ("traditional"); set phrases: *«крепкий»* ("strong"), and terms: *«цейлонский»* ("Ceylon"), *«травяной»* ("herbal"), *«зелёный»* ("green"), *«вьетнамский»* ("Vietnamese"), *«майский»* ("May tea" is a trademark in Russia), *«чёрный»* ("black"), *«английский»* ("English"), *«индийский»* ("Indian"), *«китайский»* ("Chinese").

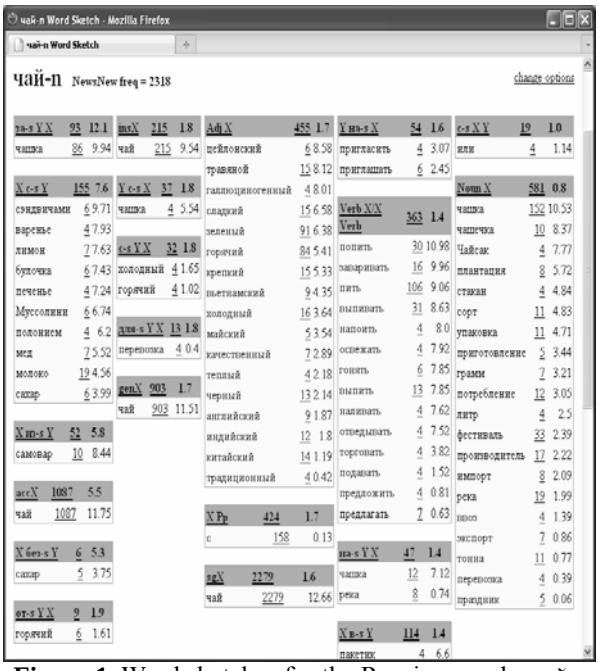

Figure 1. Word sketches for the Russian word *«чай»* ("tea")

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»/«попить»* or *«выпивать»/«выпить»* ("to drink", "to drink up"), *«заваривать»* ("to brew"), *«напоить»* ("to give to drink"), *«освежать»* ("refresh") *«наливать»* ("to pour"), *«подавать»* ("to serve") etc.

Besides collocations and terms between word sketches we also find words that belong to the same lexico-semantic classes. For example, the Russian word *«рука»* ("hand" or "arm") has the following collocates for the model "noun + keyword" (see Fig. 2): *«перелом»* ("fracture"), *«ожог»* ("burn"), *«порез»* ("cut"), *«онемение»* ("numbness"), *«обморожение»* ("chilblain"), *«ушиб»* ("contusion"), *«рана»* ("wound"). All these instances share the common sememe related to injury or some kind of disability.

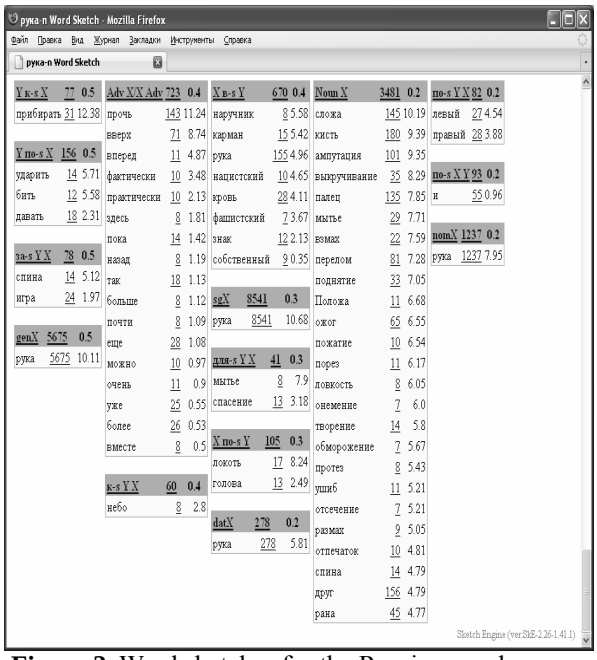

Figure 2. Word sketches for the Russian word *«рука»* ("hand" or "arm")

Apart from traditional corpus managers the Sketch Engine allows to get trigrams, e.g. with prepositions (see Fig. 3).

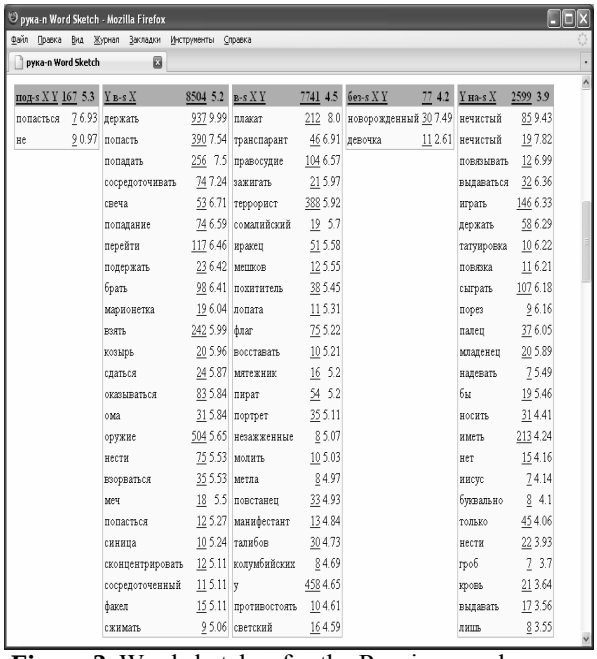

Figure 3. Word sketches for the Russian word *«рука»* ("hand" or "arm") with prepositions

Here we see a part of the output for the word *«рука»* ("hand" or "arm") that shows five tables corresponding to the following models: *«под XY»*, *«Y в X»*, *«в X Y»*, *«без X Y»*, *«Y из X»* (Y can be any word). E.g.: *«под руку попасться»* ("caught by the arm"), *«держать в руках»* ("hold in one's hands"), *«брать в руки»* ("to take into one's hands"), *«марионетка в руках»* ("marionette in one's hands"), *«сдаться в руки»* ("to yield"), *«синица в руках»* ("a bird in the hand", the beginning of a famous proverb), *«в руках правосудия»* ("in the hands of justice"), *«в руках террористов»* ("in the hands of terrorists"), *«в руках похитителей»* ("in the hands of kidnappers"), *«в руках мятежников»* ("in the hands of rebels"), *«в руках пиратов»* ("in the hands of pirates") etc. The last examples are clichés, widely used in newspapers, so this fact explains their high frequency in the given corpus. Quite often such collocations (containing prepositions) are not given in dictionaries or only a few phrases are listed in the entries, as there is no consistent approach to describe this kind of lexis. The tables in question suggest interesting hypotheses as for instance what data should be present in dictionaries. They also may help lexicographers to collect examples that are not usually found through a simple search in corpora or it's time-consuming.

5. Conclusion and Further Work

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 o 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 Sketch Engine outputs an acceptable number of collocations that can be looked over (as compared to hundreds of examples in "classic" concordance lists). Moreover there is a "cluster" function.

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.

Further development of this mechanism of collocation extraction is closely related to writing more exact grammatical rules (that will be based on additional annotation), more corpus data etc. The process of writing a word sketch grammar for any language is an iterative one: lexicographers give feedback on word sketches, discussing the word sketch rules the grammar lacks and hence get updated word sketches (at the same time they can add new corpora or necessary texts).

Also there is a question of further sketch grammar improvement as Russian has quite a free word order that requires further elaborating word sketch grammar. New variant of the sketch grammar should be based on compilation of various grammars of the Russian language (Russian Academy Grammar etc.).

As further development of this system we'd like to proceed with evaluation of various statistic measures (see above) and their application to word sketch.

The Sketch Engine system allows to collect data for "narrower" linguistics purposes. So, it can be used by scholars while studying the verbal frames (valencies) and for other similar investigations taking into account syntactic models based on the real corpus examples.

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

6. References

Sketch Engine project: http://www.sketchengine.co.uk/

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. In: *Proceedings of EURALEX-2004*, pp. 105--116.

Rychly, P., Smrz, P. (2004). Manatee, Bonito and Word Sketches for Czech. In *Trudy mezhdunarodnoy konferentsii "Korpusnaja lingvistika-2004"*: Sbornik dokladov. St.-Petersburg, pp. 324--334.

Gogol, N.V. (1937–1952). *Polnoye sobraniye sochineniy*: [V 14 volumes.]. Moscow – St.-Petersburg, 1937–1952, v. X--XIV.

Khokhlova, M., Zakharov, V. (2009). Corpus-based analysis of lexico-grammatical patterns (on the corpus of letters of N.V. Gogol). In *Proceedings of the Fifth International Conference "Computer Treatment of Slavic and East European Languages"*, Bratislava, Slovakia, 25–27 November 2009. Bratislava, pp. 211--216.

TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Documentation to the Sketch Engine: http://trac.sketchengine.co.uk/

Benko, V. Word Sketches for the Slovak National Corpus [*Oral presentation at Mondilex workshop*]: http://korpus.juls.savba.sk/~mondilex/programme3.pdf

Slovak National Corpus: http://korpus.sk

Rychly, P. (2008). A Lexicographer-Friendly Association Score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing*, RASLAN 2008. Brno, pp. 6--9.

Russkaja grammatika. (1980). Volumes I, II. Moscow. (AG-80).