# The Corpus of Syntactic Co-occurences: the First Glance

Edward S. Klyshinsky[1,2], Petr D. Ermakov[1], Natalia Yu. Lukashevich[3], and Olesya V. Karpik[2]

[1] NRU Higher School of Economics, Moscow, Russia,
{eklyshinsky,permakov}@hse.ru,
[2] Keldysh IAM, Moscow, Russia,
parlak@mail.ru
[3] Moscow State University, Moscow, Russia,
natalukashevich@mail.ru

**Abstract.** Modern corpora provide suitable access to the stored data. However, they are convenient rather for researchers than for students learning a foreign language and not familiar with the corpus linguistics. Therefore, we set the task of creating a corpus, which contains information on words co-occurrence, their syntactical relations and their government for the Russian language.

**Keywords:** words co-occurences, web-site, educational corpus.

## 1 Introduction

Modern corpora offer various opportunities both for research and education. However, as they differ in size, functionality and methods used, some of them are more suited for solving particular tasks than others. If we consider an educational corpus, learners of a language need such a corpus that will help them check whether a word combination is correct or not, whether there are any mistakes in the government, whether the constructed word combination can be used in the given context, etc. Though many corpora (e.g. Sketch Engine [1], Russian National Corpus [2], [3], and some others) offer examples of correctly written text containing given words, they are only able to search for contact groups or co-occurrences but not syntactically related words. Several treebanks like Pen Treebank [4], Tiger [5], and SynTagRus [6] can provide such information, but their size is enough just for searching the most frequent phenomena. These corpora are big enough to train a parser, but they do not contain information about a big part of the vocabulary for the target language, which is a crucial requirement for an educational corpus. It should contain examples that illustrate at least the most frequent word combinations. Moreover, learners need a specialized interface to these corpora to be created.

Unlike a corpus for researchers, an educational corpus can have less recall but higher precision rate. Such a corpus should be tagged using the information from a vocabulary sampled by linguists but not automatically generated from texts using forecasting technique. Types and initial forms should not be predicted stochastically because this procedure has a relatively higher error rate in comparison with manual tagging. The corpus interface should provide convenient access to lists of words syntactically connected with the selected one, as well as statistical information on co-occurrences' frequencies. It should preferably allow switching between co-occurrences and word forms, change the focus from the head word to the dependent one and vice versa.

Even the most developed tool, the Sketch Engine, does not provide all this functionality. Such operations as ordering by frequency of occurrence need extra calculations, and in a big corpus they take a long time. The list of dependent words is previously shortened and cannot be ordered by frequency. Predicted words can be crucial in a research project, but they may mislead even an intermediate-level learner. Because of these features the Sketch Engine should rather be classified as a corpus for researchers than one for a student learning a foreign language and not familiar with the corpus linguistics.

The quality of modern parsers is enough for machine translation, however it is not enough to gather a database of word co-occurrences that can be used as a learner's corpus. Some postprocessing brings us to results demonstrated by RNC Sketches that can be found at http://ling.go.mail.ru/synt/. This site provides information on words co-occurrences, syntactical roles of words, frequencies of their co-occurrences. However, the size of analyzed corpora does not allow to find all possible combinations, and this project inherits all properties of Sketch Engine (i.e. like Sketch Engine, RNC Sketches is rather a corpus for researchers than a corpus for students learning Russian). Therefore, we set the task of creating a corpus, which contains information on words co-occurrence, their syntactical relations and their government for Russian. Having access to such a corpus, a student could consult if the constructed phrase is correct or not according to the mutual words' occurence.

We analyzed several corpora using shallow parsing technique. As result, we sampled a database for the most part of Russian vocabulary. Thus, we believe that our corpus of syntactically connected words is not complete but big enough for educational purposes.

## 2 Co-occurence extraction

As it was claimed in [7], a regular automaton allows extracting the subcategorization frame of English verbs for such phrases, which have clear syntactical structure. However, direct implementation of this method for Russian texts leads us to a high rate of mistakes because of homonymy. It should be noted here that Russian homonymy is very different from the homonymy in English, as it was shown in previous studies in [8]. Unlike English, Russian words are more often ambiguous due to confusion of forms within one part of speech than to confusion of different parts of speech, In fact, Russian texts contain about 80% of POS-unambiguous words. Thus, if we gather phrases that contain unambiguous words only, it reduces the number of considered co-occurrences but also reduces the percentage of mistakes.

We used the method introduced in [9]. The constructed regular automaton gathers such phrases like a noun or a prepositional phrase following the only verb in the sentence, a sole noun or a prepositional phrase followed by a verb, and some other special cases of Russian sentences. We gathered information about such combinations as verb+noun(+preposition), verb+adverb, noun+adjective, adjective+adverb. (In our experiments, we considered participles and gerundial adverbs as a form of the corresponding verb.)

The main difficulty here was to develop a new language to describe regular expressions used over untagged text to extract syntactically related words. This type of regular expression should allow to find subordinated or governed words. Such language should also allow to study not only the initial form of the word or its type but a set of types obtained by tagging.

In our experiments we used the following untagged Russian corpora: fiction taken from LibRusEc.ru (ca 15 bln word tokens), news wire downloaded from different web sites (about 1 bln word tokens), popular science news wires (about 150 mln word tokens), and scientific texts (PhD thesis and its summaries, conference proceedings, scientific journals in different domains –about 100 mln word tokens in total). Thus, the size of our corpora is comparable with the size of Sketch Engine RuTenTen corpus, which is 14.5 bln of tokens, but not so good balanced.

On the first step, we applied selected regular expressions to these corpora. We applied such phrases with clear syntactical structure as the following: an only noun phrase in the beginning of the sentence and followed by a sole verb. In this case the noun is obviously connected to the verb, all adjectives are connected to the noun, and the adverb on the first position can be connected to the first adjective. The complete set of applied rules is described in [9].

The collected results were as big as ca 79 mln verb+prep+noun, ca 134 mln verb+noun, ca 35 mln adj+noun co-occurrences and some other combinations. We failed to gather information for a variety of connections, e.g., on noun+noun because of a very high rate of mistakes in the output (in about 50% of extracted combinations both nouns are connected to the same verb). On the second step, we filtered obtained results in order to reduce the error rate (e.g. filter out all combinations that are not included in verbal subcategorization frame [9]). Finally, we calculated the frequency of occurrence for obtained combinations reduced to initial forms.

The constructed corpus contains 7.5 mln of verb+noun(+preposition) unique combinations and 2.3 mln of noun+adjective. Finally, we gathered information on combinations of 23000 verbs and 57000 nouns, and combinations of 39000 nouns and 31000 adjectives. We have calculated the frequency of occurrence for every word combination: every co-occurrence is connected with a set of examples selected from the text.

Some of the recent tools, such as Sketch Engine [1] and CoCoCo (http://cococo.cs.helsinki.fi/), are fetching co-occurences in on-line mode; a user sends his request, and the server calculates set the co-occurences according to this request. This approach economises hard disk space, however processing of a user's request can take a long time. Following such corpora as RNC [2], we are precalculating any possible output; as result a user is able to navigate through the list of words and receive their connection in a short time. Among other, we do not use word prediction during the analysis of our corpora. Thus, the list of words is shorter than in Sketch Engine or RNC, however our databse has less rate of mistakes.

For example, if a student is looking for information on adjectives combined with the word "администрация (administration)" this student will get combinations as it shown in Table. 1. Selecting a word, user will get a set of examples.

Table 1. Words co-occured with word ADMINISTRATION an their frequencies

| **Frequency** | **Word (Translation)** |
|---:|---|
| 1726 | местный (local) |
| 1666 | городской (city) |
| 1278 | американский (american) |
| 967 | новый (new) |
| ... | ... |

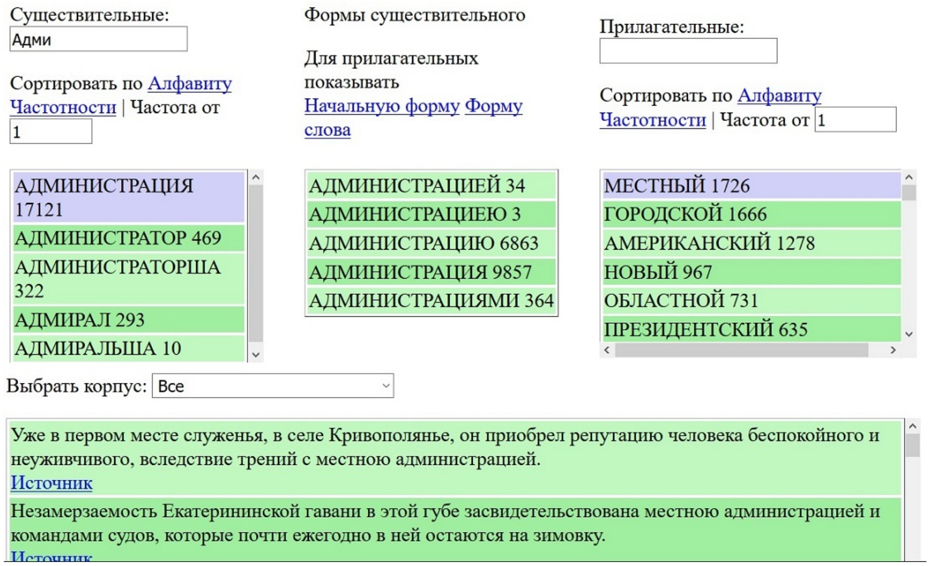
Our corpus contains connections between word forms and their initial forms, thus, a user is able to see a word form and its connections.

Note that our corpus contains rather usage than syntactically correct occurrences. Sometimes both incorrect usage and academic norm are equally frequent. This situation reflects the fact that this phenomenon should be investigated and learned "in vivo" but not "in vitro". Our corpus contains rare syntactically but not semantically correct combinations. In this case, a student can consult frequencies of combinations and make a proper choice.

The gathered corpus of syntactic co-occurrences is available through web site www.cosyco.ru. A student can select a type of syntactic connection between words.

Unlike in such corpora as Sketch Engine, we organized the structure of the corpus as a tree. A user enters a word he or she is interested in, and the site demonstrates the list of words connected with this one by the selected type of connection. In the case of a noun, it could be a list of governed adjectives or governing verbs (and prepositions) depending on the user choice; in the case of an adjective, it could be a list of governing nouns or connected adverbs. When selecting a combination, the user receives a list of sentences containing this co-occurrence. The source of the selected sentence can be found using Yandex search engine by clicking the proper link. In case of noun+adjective combinations, the user is also provided by forms of adjectives (see Fig. 1).

Both nouns and adjectives can be filtered out by their frequency. We did not provide any other measures such as MI, log-likelihood and some others because of they are usefull for a comparison of different words combinations - but this is not the case. The main point here is the head word and all its dependent words, thus the dependent words can be arranged using its frequencies only.



Существительные: Адми

Сортировать по Алфавиту Частотности | Частота от 1

АДМИНИСТРАЦИЯ 17121
АДМИНИСТРАТОР 469
АДМИНИСТРАТОРША 322
АДМИРАЛ 293
АДМИРАЛЬША 10

Формы существительного

Для прилагательных показывать Начальную форму Форму слова

АДМИНИСТРАЦИЕЙ 34
АДМИНИСТРАЦИЕЮ 3
АДМИНИСТРАЦИЮ 6863
АДМИНИСТРАЦИЯ 9857
АДМИНИСТРАЦИЯМИ 364

Прилагательные:

Сортировать по Алфавиту Частотности | Частота от 1

МЕСТНЫЙ 1726
ГОРОДСКОЙ 1666
АМЕРИКАНСКИЙ 1278
НОВЫЙ 967
ОБЛАСТНОЙ 731
ПРЕЗИДЕНТСКИЙ 635

Выбрать корпус: Все

Уже в первом месте служенья, в селе Кривополянье, он приобрел репутацию человека беспокойного и неуживчивого, вследствие трений с местною администрацией.
Источник

Незамерзаемость Екатерининской гавани в этой губе засвидетельствована местною администрацией и командами судов, которые почти ежегодно в ней остаются на зимовку.
Источник

Fig. 1. Example: combinations of word ADMINISTRATION with adjectives

Using a new kind of regular automaton, we gathered a big corpus of syntactically connected word combinations for Russian. The corpus contains 7.5 mln of verb+noun(+preposition) and 2.3 mln of noun+adjective unique combinations extracted from news wire and fiction texts. We have made a first step to create an interface suitable for a student learning Russian.

In our future work, we are going to provide the filtering by subcorpora in addition to the information on the subcorpora genres. We plan to increase the size of the used corpora and attach some new regular expressions. This will allow us to broaden the amount of gathered information. Another problem is the purity of corpus. We have found that LibRusEc corpus contains texts in Ukranian, Belorussian, Bulgarian and Serbian languages. Both news wire and fiction subcorpora are containing text duplicates, thus multiple versions of the same sentence should be eliminated. Such corpora as GICR [10] are processing this information but in case of an educational corpus it hinders the student's perception. Finally, we used the simplest tokenizer and parser in our experiments; thus, we have to completely remaster our corpus and extract new database of co-occurences. The current version can be used as a learner's corpus just after all these changes; however, it can be used as a researcher corpus in the current state.

**Acknowledgement.** This project is supported by RFH grant 15-02-12019.

## References

1. *Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D.* The Sketch Engine, *In Proc. of EURALEX 2004*, pp. 105–116.
2. *Lashevskaja, O., Plungian V.* Morphological annotation in Russian National Corpus: a theoretical feedback // Proc. Of 5th International Conference on Formal Description of Slavic Languages (FDSL-5). Nov. 2003, pp. 26-28.
3. *Savchuk, S. O., Sichinava, D. V.* Obuchayuschiy korpus russkogo yazyka i ego ispol'zovanie v prepodavatel'skoy praktike [In Russian] // Russian National Corpus 2006-2008: New Results and Perspectives. 2009, pp. 317-334.
4. *Marcus, P. M., Santorini, B., Marcinkiewicz, M. A.* Building a Large Annotated Corpus of English: The Penn Treebank // Journal of Computational Linguistics, Vol. 19, Iss. 2, 1993. Pp. 313-330
5. *Brants, S., Dipper, S., Eisenberg, P. et al.* TIGER: Linguistic Interpretation of a German Corpus // Journal of Language and Computation, 2004 (2), 597-620.
6. *Frolova, T. I., Podlesskaya, O. Yu.* Tagging lexical functions in Russian texts of SynTagRus // In Proc. of «Dialog 2011», pp. 207-218
7. *Manning, D.* Automatic Acquisition of a Large Subcategorization Dictionary from Corpora *In Proc. of the 31st Meeting of ACL*, pp. 235–242.
8. *Klyshinsky, E. S., Logacheva, V. K.,Nivre, J.* The Distribution of Ambiguous Words in European Languages [In Russian] // In Proc. of Corpus Linguistics 2015, pp.270-277.
9. *Klyshinsky, E., Kochetkova, N., Litvinov, M., Maximov, V.* Method of POS-disambiguation Using Information about Words Co-occurrence (For Russian) // In Proc. of GSCL'2011, pp. 191-195
10. *Belikov, V., Selegey, V., Sharoff, S.* Preliminary considerations towards developing the General Internet Corpus of Russian. [Prolegomeny k proektu General'nogo internet-korpusa russkogo yazyka] // In Proc. Int. Conf. on Computational Linguistics "Dialog", 2012, vol.1, pp. 37—49.