Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2018"

Moscow, May 30—June 2, 2018

CREATING A CORPUS OF SYNTACTIC CO-OCCURRENCES FOR RUSSIAN[1]

Klyshinsky E. S. (klyshinsky@mail.ru)
Keldysh IAM RAS, Moscow, Russia

Lukashevich N. Y. (natalukashevich@mail.ru),
Kobozeva I. M. (kobozeva@list.ru)
Moscow State University, Moscow, Russia

In the paper we discuss methods used to create CoSyCo, a corpus of syntactic co-occurrences, which provides information on syntactically related words in Russian. We describe a list of shallow parsing templates, which were used to collect data for CoSyCo. The paper includes an overview of the corpora collected for CoSyCo creation and an outline of how the noun 'virus' is used in its subcorpora as an example of the information which can be obtained from this online resource.

Keywords: corpora creation, shallow parsing, grammatically ambiguous text, words combinations, the Russian language

ОПЫТ СОЗДАНИЯ КОРПУСА СИНТАКСИЧЕСКИХ КОМБИНАЦИЙ РУССКОГО ЯЗЫКА

Клышинский Э. С. (klyshinsky@mail.ru)
ИПМ им. М. В. Келдыша РАН, Москва, Россия

Лукашевич Н. Ю. (natalukashevich@mail.ru),
Кобозева И. М. (kobozeva@list.ru)
МГУ им. М. В. Ломоносова, Москва, Россия

[1] The project is partially funded by RFH grant 15-04-12019.

В данной статье дается основная информация о составе нового открытого ресурса — КоСиКо, корпуса синтаксических комбинаций, содержащего синтаксически связанные группы слов русского языка. Описывается состав текстовых корпусов, использованных для формирования КоСиКо, дается информация о шаблонах анализа текстов, использованных для извлечения информации. На примере употребления слова «вирус» с прилагательными показано, какого рода информацию можно получить из корпуса.

Ключевые слова: текст без снятой омонимии, поверхностный синтаксический анализ, создание корпуса, лексическая сочетаемость, русский язык

1. CoSyCo: a Corpus of Syntactic Co-occurrences

In this paper we continue a series of works introducing a Corpus of Syntactic CoOccurrences (CoSyCo)[2]—a new resource providing information on word combinations in Russian.

It allows to get lists of word combinations together with examples of sentences in which they are used in real texts in the Internet giving information on word's co-occurrences, on syntactic relations between words.

We have already briefly outlined in (Klyshinsky, Lukashevich, 2017) the current state of affairs with regard to online resources providing similar information for Russian: one cannot say that there is a total lack of them. However, we believe that a freely accessible database which would be of a size big enough for various kinds of tasks, collected over huge untagged corpora with simple methods, offer a convenient interface and certain other important features is still to be designed [Klyshinsky, Lukashevich, 2017].

In this work we would like to focus on the structure and contents of CoSyCo database and to discuss methods used to create it.

2. CoSyCo structure

For this project we gathered data from open sources which we grouped in the following five subcorpora[3].

[2] http://cosyco.ru/

[3] The fact that Librusec fiction collection by far outweighs all other subcorpora is to a great extent a result of technical issues (i.e. texts from which sites we managed to collect). Our intention was primarily to make a variety of text styles and genres available to a user. The importance of including texts which differ in style and genre into a corpus has been widely discussed in [Belikov et al, 2012], [Belikov et al, 2013], [Lukashevich et al, 2016]. Besides, which particular subcorpora size combination would make the corpus 'balanced' is not a trivial issue either. We definitely plan to increase the size of smaller subcorpora, but we believe that at the moment they can still be of help in a research as they are.

Table 1. CoSyCo subcorpora

	CoSyCo subcorpora	mln words	%
1.	News sites	1,400.9	8.07%
2.	IT news sites	142.4	0.82%
3.	Lib.rus.ec fiction collection	~15,000.0	86.38%
4.	Science sites	102.2	0.59%
5.	Wikipedia.ru texts (dump 01/05/2016)	~401.0	2.31%
6.	Russian Patents (http://www1.fips.ru/)	317.8	1.83%
	Total	~17,364.0	100.0%

News sites included the following sources:

Table 2. News subcorpus in CoSyCo

News sites:	1,400.9 mln words	100%
lenta.ru	89.0	6.35%
RBK	66.0	4.71%
RIA Novosti	473.0	33.76%
Nezavisimaya gazeta	56.3	4.02%
Vzglyad	72.0	5.14%
Rossiyskaya gazeta	88.5	6.32%
Commersant	158.0	11.28%
Polit.ru	81.6	5.82%
Utro.ru	47.5	3.39%
Ibusiness	10.5	0.75%
Championat.com	1.2	0.09%
Moskovsky Komsomolets	72.1	5.15%
Gazeta.ru	78.4	5.60%
Komsomol'skaya Pravda	106.8	7.62%

IT news were taken from the following sources:

Table 3. IT news subcorpus in CoSyCo

IT news	142.4 mln words	100%
Membrana.ru	7.7	5.41%
CNews	43.6	30.62%
Computerra.ru	28.0	19.66%
Compulenta.ru	16.0	11.24%
PCWeek	23.3	16.36%
OSPNews	9.0	6.32%
Popular Mechanics	3.7	2.60%
NPlus1.ru	11.1	7.80%

Science sites covered a wide range of spheres and topics:

Table 4. Science subcorpus in CoSyCo

Science sites	102.2 mln words	100%
Childpsy.ru (dissertations)	12.6	12.33%
Civil Service Journal	1.8	1.76%
Delist.ru	8.9	8.71%
Dialogue conference	2.0	1.96%
Discollection.ru	11.3	11.06%
disser.aspirantura.spb.ru	1.6	1.57%
Geographic journals and books collection	4.3	4.21%
Musical journals	0.9	0.88%
Programming books collection	19.8	19.37%
Pu7.ru	18.5	18.10%
CAD and Graphics Journal	5.4	5.28%
Scientific Visualization Journal	0.2	0.20%
Information Security Journal	1.0	0.98%
Software Systems Journal	2.8	2.74%
Tomsk State University (TSU) Journal of Biology	1.0	0.98%
TSU Journal. Control, Computers and Informatics	0.6	0.59%
TSU Journal. Applied Discrete Mathematics	0.3	0.29%
TSU Journal	9.2	9.00%

3. Syntactic patterns used for data extraction

We have already partly described what methods and software were used to create CoSyCo database in [Klyshinsky et al., 2011], [Klyshinsky et al., 2016] and [Klyshinsky, Lukashevich, 2017]. One of the essential parts of CoSyCo project is a software tool for the extraction of syntactically connected words. This tool is written as a data-driven system that takes as input a template and a corpus and extracts combinations of a given format. In this part, we will discuss these templates in more detail.

It is known that in Russian some sequences of PoS-unambiguous words[4] can be considered as syntactically unambiguous without grammatical disambiguation. The structure of such sequences can be represented in the form of templates, which will help to identify whether the words in a phrase with the given structure are syntactically connected or not.

Experiments previously conducted on news corpora for seven European languages as described in [Klyshinsky et al, 2015] demonstrate that there are significant

[4] We understand a PoS-unambiguous word as a word with the same part of speech identified for every possible grammatical analysis.

differences in the structure of homonymy/ambiguity: in Russian up to 50 % of words are unambiguous, and almost 80% of words are PoS-unambiguous (as compared to about 40% in English).

Another research in [Klyshinsky, 2017] focused on syntactic inversion in 33 languages. The study compared the number of left- and right-branching sentences for different types of syntactic links. The resulting figures demonstrated that the syntax of the Russian language is not so free as it may seem: Russian did not make it to the top ten languages with the free word order, coming in the middle of the rating.

These two points brought us to the general idea of the current research: if we take into account only very simple cases where it is easy to identify a syntactic relation between words (with no mistakes or with a negligible amount of them), and apply the corresponding templates to a very large corpus, with a comparatively high rate of PoS-unambiguous words and a relatively strict word order it should be possible to find most of possible combinations for a representative amount of words.

Certain points should be explained here. We did not plan to use various available tools for homonymy disambiguation in our work, because we wanted to avoid mistakes which they inevitably add. As for the existing search tools in tagged corpora (like Sketch Engine), they do not allow to work with non-disambiguated texts. Since we were particularly interested in this task, we needed to develop our own tool for it.

The next step was to formulate the templates for extracting syntactically connected words and check their work manually (for details see [Klyshinsky, Lukashevich 2017]).

Below, we will describe several templates of this kind which were used on the initial stages of the project.

I. Under a noun phrase (NP) in our work we understand a sequence of adjectives (possibly combined with optional adverbs) and a noun[5] which agree in gender, number and case. Obviously a noun phrase in Russian may contain various other elements, but we take into account only those which have such clear structure and, moreover, which contain only PoS-unambiguous words. A prepositional phrase (PP) is a noun phrase with a preceding preposition (as in (1)).

(1) PP = Prep + NP = Prep + (Adv) + (Adj$^+$) + N

Вероника повернулась, чтобы встретиться

'Veronika turned to look

с	***мягкими***	***зелеными***	***глазами***.
into	***soft***	***green***	***eyes***.'

The group of templates below (II–VII) helps to establish whether there is a syntactic connection between words in certain positions in a sentence. For this group of templates it is important that all words should be PoS-unambiguous, and that the NPs and PPs mentioned in the templates should be clearly separable from the context before and after them (e.g., by the beginning or the end of the sentence, the use of a PoS-unambiguous verb, a preposition, etc). The second condition is true not only for templates II–VII, but for VIII–X as well.

[5] Adjectives here should not be in a short or comparative form.

II. If a sentence starts with a single NP (as in (2a)) or a PP (as in (2b)) and such a phrase is followed by a single verb, then the noun in such a phrase and the verb are syntactically connected[6].

(2) a. ***Российские*** ***аналитики*** ***соглашаются*** *с тем, что …*
Russian ***analysts*** ***agree*** *that…*

b. ***На*** ***севере*** ***граничит*** *с Латвией.*
In the ***North*** *(it)* ***borders*** *Latvia.*

III. The noun in the first NP or PP[7] which is used after a single verb is syntactically linked with this verb.

(3) *Новая технология* ***предоставляет*** ***опытным*** ***пользователям*** *расширенный набор возможностей печати.*
The new technology ***offers*** ***experienced*** ***users*** *a broader range of printing options.*

IV. The same conclusion as in II that a noun and a verb are syntactically linked can be made if the NP or PP is placed at the beginning of a subordinate clause which starts with a connector after a comma and if this NP or PP is followed by a single verb.

(4) *Блатт хотел, чтобы* ***сезон*** ***завершился*** *в начале мая.*
Blatt wanted that the ***season*** ***be over*** *in the beginning of May.*

V. An adverb placed between a preposition, noun, conjunction, or personal pronoun and an adjective is syntactically connected to this adjective.

(5) *Знаменитые эльфийские лучники* ***практически*** ***беспомощны*** *при такой погоде.*
The famous elven archers are ***virtually*** ***helpless*** *in this kind of weather.*

VI. If a participle is used before a noun in NP or PP (i.e. the position of the participle is typical for an adjective), then it is syntactically connected to the noun.

(6) ***Рассматриваемая*** ***проблема*** *находится на стыке дисциплин.*
The investigated ***problem*** *is at the intersection of several domains.*

[6] It is important to note that we are not concerned about the direction of the connection here.

[7] There may be several noun or prepositional phrases after a verb, we are talking about the first of them.

VII. If a participle is used after NP or PP, is separated from it by a comma and agrees with the noun in this preceding NP or PP in gender, number and case, then the participle is syntactically connected with the preceding noun.

(7) *Системная* ***интеграция,*** ***проводимая*** *на заводе компании, …*
The system ***integration*** ***performed*** *at the plant of the company…*

In all templates above it was necessary that all words should be PoS-unambiguous. However, certain cases of PoS-ambiguity can be successfully resolved during analysis[8]. Templates VIII, IX and X show examples of this.

VIII. If NP or PP includes a word, which is ambiguous between an adjective and a participle (as in (8a) or it is ambiguous between an adjective and a noun (as in (8b))[9], and if there is a PoS-unambiguous adjective in the same phrase then the ambiguous word should be considered an adjective.

(8) a. (Prep)NP = (Prep) + ?**Adj**/ Part + Adj + Noun
В Москве прошло вручение премии имени Елены Мухиной,
которой награждаются люди ***с*** ***ограниченными***
физическими ***способностями****.*
limited -ADJ/PART
The ceremony of Elena Mukhina's award,
which is granted to people ***with*** ***limited***
physical ***abilities,*** *took place in Moscow.*

b. (Prep)NP = (Prep) + ?**Adj**/ Noun + Adj + Noun
Прямая ***длинная*** ***линия*** *лезвия была скошена к концу.*
direct-ADJ/a line-NOUN
The ***direct*** ***long*** ***line*** *of the blade was slanted towards its end.*

IX. If NP or PP is at the end of the sentence and its last word is ambiguous between a noun and a verb (9a) or a noun and an adjective or participle (9b), then this last word in the phrase should be considered N. (The sequence should also meet the necessary criterion that in the resulting phrase the noun agrees with the preceding adjective(s) in its gender, number and case. The same applies to (9b).)

(9) IX a. …(Prep)NP = (Prep) + (Adj) + ?**Noun**/Verb.
a. *Он уставился* ***на*** ***лобовое*** ***стекло****.*
glass-NOUN / flow down—PAST-SG-N
He stared at the ***front*** ***window****.*
IX. b. …(Prep)NP = (Prep) + (Adj) + **?Noun**/Adj.
b. *Предстоит долгий путь до* ***финишной*** ***прямой****.*
It is still a long way to the ***home*** ***straight****.*
direct-ADJ/a line-NOUN

[8] This is especially important for Russian, where a lot of nouns are derived from adjectives, so that they are ambiguous in every form (e.g. больной 'ill / an ill person').

[9] which are the most typical ambiguity cases for adjectives

X. If NP or PP is followed by a(nother) PP and the last word of the first phrase is ambiguous between a noun and a verb (10a) or a noun and an adjective or participle (10b), this last word should be considered a noun. (Here the first NP or PP should be preceded by a verb, a punctuation mark, or the beginning of the sentence.)

(10) X a. [(Prep) +…+?Noun/V]PP/NP + PP/NP
a. *Он разбил* ***оконное*** ***стекло*** *в школьном коридоре.*
glass-NOUN / flow down -PAST -SG-N
He broke a ***window*** *pane* *in the school's passage way.*

X b. [(Prep) +…+?Noun/Adj/Part] PP/NP + PP/NP
b. ***Сводные*** ***данные*** *о значениях параметров …*
data-N / give-ADJ/Part-PL
The ***integrated*** ***data*** *on the parameters …*

4. Improving the results

In this section we will discuss the results obtained with the initial set of templates, and what steps had to be taken to improve them. (It was briefly mentioned in [Klyshinsky, Lukashevich, 2017]), here we will try to go into more detail.)

When we assessed how complete the database of combinations was, we found that a certain part of vocabulary was missing. While checking why this happened, we saw that at least one reason was that words which are grammatically ambiguous in all their forms in Russian (e.g. ученый is ambiguous between a noun ‹a scientist› and an adjective 'learned, academic' in every form) were disregarded during processing, They proved to be so frequent, that this dropped the amount of identified nouns and adjectives down.

To avoid this, we had to lift certain restrictions in several templates—we had to allow words ambiguous between a noun and an adjective in templates I and IX[10]. (The resulting templates are I*, IXa*, and IXb* respectively).

I* PP = Prep + NP = Prep + (Adv) + (Adj^{+}) + **?Adj/ Noun**

IXa* …(Prep)NP = (Prep) + **?Adj/ Noun** + ?Noun/Verb

IXb* …(Prep)NP = (Prep)+ **?Adj/ Noun** + ?Noun/Adj.

We also added a new template which identified verbs from short forms of participles and established a link between such a verb and a noun in the noun phrase.

[10] This technically meant that we had to "soften" our initial position that only PoS-unambiguous words should be taken into account. The table below shows that these changes significantly improved the figures in CoSyCo database. This increase in figures also allows to indirectly assess the relative percentage of words ambiguous between a noun and an adjective. We believe all of this to be important, that is why we deliberately give a detailed account of the course of work instead of simply showing the current set of templates.

IX. If a participle in a short form is followed by NP or PP, then it is syntactically linked with the noun in NP or PP, and the same holds true for its producing verb.

Similarly, if NP or PP at the beginning of the sentence is followed by a participle in a short form, the same conclusions can be made.

(11) a. *Вырезки не были* ***разложены*** ***в хронологическом*** ***порядке***.
The cuttings were not ***placed*** ***in chronological*** ***order***.

b. ***Личный*** ***состав*** ***размещен*** *в закрытом городке.*
The military ***personnel*** ***was placed*** *in a restricted-access town.*

Another change was an additional set of conditions in several templates. Template II, for example, will also hold true if the requested group is used after a punctuation mark, as in (2c):

(2) c. *Необходимо тестирование 60% программ,*
считают ***эксперты*** *Ассоциации.*
It is necessary to test 60% of software,
believe *Association* ***experts***.

The table below shows the effect such amendments had on the figures in the database.

Combination	Lemma combinations, mln		Token combinations, mln		Total occurences, mln	
	old	new	old	new	old	new
noun+adj	12.1	18.3	25.5	39.8	383	746
verb+prep+noun	29.2	33.4	53.5	60.3	349	412
participle+noun	3.1		5.1		28.1	
participle+prep+noun	1.2		1.8		4.3	

Table 5. CoSyCo database before and after amending the templates

Combination	nouns		adjective		verbs	
	old	new	old	new	old	new
noun+adj	67,000	71,000	41,000	42,000		
verb+prep+noun	73,000	73,000			28,000	28,000
participle+noun	52,000				20,000	
participle+prep+noun	40,000				15,000	

We compared the vocabulary extracted from CoSyCo database with the one obtained from SynTagRus (for details see [Klyshinky, Lukashevich, 2017]) and with the dictionary of I-RU bigrams from the database of Collocations Colligations Corpora [Kormacheva et al, 2016]. We found out that the vocabularies of the latter two resources differ significantly. For most frequent words (with frequencies over 1,000) the differences were between 1% and 4%. However, for words with lower occurrence figures (over 10) the differences were between 30% and 70%, with CoSyCo vocabulary being more complete. Comparison results are shown in Table 6.

Table 6. Comparison of I-RU and CoSyCo vocabularies

		I-Ru		CoSyCo	
Part of Speech	Frequency	Not found in CoSyCo	Total	Not found in I-Ru	Total
Noun	>1,000	226 (4,3%)	5,229	523 (3,1%)	16,881
	>500	534 (6,4%)	8,376	1,333 (6,0%)	22,209
	>100	3,887 (18,4%)	21,122	7,987 (21,7%)	36,866
	>10	30,298 (49,1%)	61,720	27,102 (46,3%)	58,524
Adjective	>1,000	10 (0,6%)	1,677	4,138 (30,3%)	13,635
	>500	22 (0,8%)	2,728	6,890 (41,2%)	16,705
	>100	405 (6,4%)	6,312	14,390 (58,8%)	24,481
	>10	4,014 (28,3%)	14,192	23,026 (69,4%)	33,194
Verb	>1,000	21 (0,9%)	2,291	197 (2,0%)	9,975
	>500	56 (1,6%)	3,601	725 (6,1%)	11,975
	>100	426 (5,2%)	8,153	4,073 (24,6%)	16,561
	>10	3,670 (22,5%)	16,291	10,598 (45,6%)	23,219

To assess the recall for identified word combinations we did the following. We applied a "weak" template according to which any two words which are Adj + Noun are syntactically linked. Such a template will definitely bring many false connections, but the most frequent ones should presumably be correct. We analyzed 10,000 most frequent combinations obtained with the help of this "weak" templates. We found only 159 nouns (about 1.5%) for which in CoSyCo database there were less than 75% adjectives identified with the "weak" pattern. The links were mostly missing when there was a mistake (e.g. one of the words did not belong to the requested part of speech in the context). We checked adjectives with frequencies higher than 10, and for about 2,400 nouns we did not manage to find only 1% of such adjectives, whereas for 23 nouns over 5% of such adjectives were missing. This usually meant that the omitted links were in the least frequent part of the list.

5. Word combination types on the site

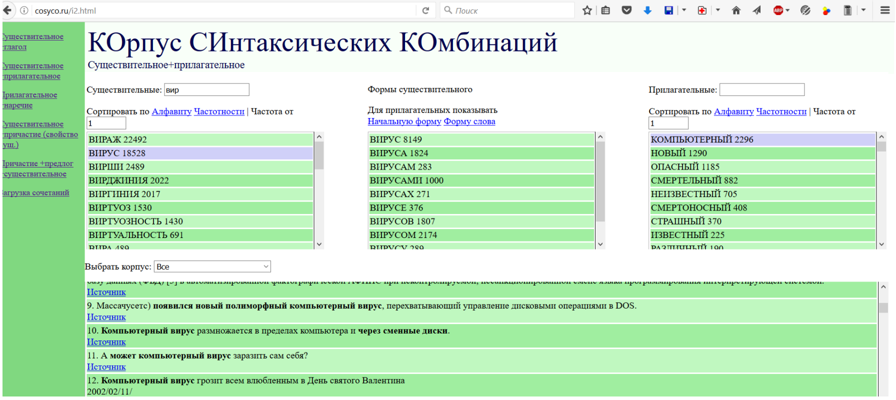
At the moment a user of the site can find data on the following types of combinations (the relevant type can be selected from the left side menu of the screen):

- **verb**+preposition+noun (***арендовать*** *у компании 'rent from a company'*)[11],
- **noun**+adjective (*компьютерный* ***вирус*** *'computer virus'*),

[11] In the title, the head constituent for the combination is highlighted in bold. The order of the elements in the title does not necessarily coincide with the typical word order in sentences with such word combinations. This was done on purpose to keep the logic so that it helps to find the relevant section for a combination regardless of the real word order in the sentence.

- **noun**+participle (*созданный **имидж** ‹created image›*),
- **participle**+preposition+noun (***арендованный** у компании ‘rented from a company’*),
- **adjective**+adverb (*очень **амбициозный** ‘very ambitious’*).

Fig. 1 shows what the search page of the site looks like.


cosyco.ru/i2.html
Поиск
уществительное +глагол
уществительное +прилагательное
Ірилагательное +наречие
уществительное +причастие (свойство ущ.)
Іричастие +предлог +существительное
агрузка сочетаний

КОрпус СИнтаксических КОмбинаций

Существительное+прилагательное

Существительные: вир

Сортировать по Алфавиту Частотности | Частота от 1

ВИРАЖ 22492
ВИРУС 18528
ВИРШИ 2489
ВИРДЖИНИЯ 2022
ВИРГИНИЯ 2017
ВИРТУОЗ 1530
ВИРТУОЗНОСТЬ 1430
ВИРТУАЛЬНОСТЬ 691
ВИРА 489

Формы существительного

Для прилагательных показывать Начальную форму Форму слова

ВИРУС 8149
ВИРУСА 1824
ВИРУСАМ 283
ВИРУСАМИ 1000
ВИРУСАХ 271
ВИРУСЕ 376
ВИРУСОВ 1807
ВИРУСОМ 2174
ВИРУСУ 289

Прилагательные:

Сортировать по Алфавиту Частотности | Частота от 1

КОМПЬЮТЕРНЫЙ 2296
НОВЫЙ 1290
ОПАСНЫЙ 1185
СМЕРТЕЛЬНЫЙ 882
НЕИЗВЕСТНЫЙ 705
СМЕРТОНОСНЫЙ 408
СТРАШНЫЙ 370
ИЗВЕСТНЫЙ 225
РАЗЛИЧНЫЙ 190

Выбрать корпус: Все

Источник

9. Массачусетс) **появился новый полиморфный компьютерный вирус**, перехватывающий управление дисковыми операциями в DOS.
Источник

10. **Компьютерный вирус** размножается в пределах компьютера и **через сменные диски**.
Источник

11. А **может компьютерный вирус** заразить сам себя?
Источник

12. **Компьютерный вирус** грозит всем влюбленным в День святого Валентина
2002/02/11/

Fig. 1. CoSyCo search page showing examples for the word combination *компьютерный вирус* ‘computer virus’

On the search page for the verb+preposition+noun and participle+preposition+noun sections, the left column lists verbs, the head constituent for the verb+preposition+noun combination and a derivational basis for the participle in participle+preposition+noun combination. In the middle column, a user can choose a preposition from the list of options that combine with the chosen verb, and in the right column s/he immediately gets a list of nouns which were found in the texts (=can be combined) with this verb+preposition and a list of sentences containing this expression in the lower part of the screen[12]. The resulting lists of words can be sorted by frequency or alphabetically.

In a similar way for the noun+adjective and noun+participle sections the left column gives a list of nouns. A user can choose a word form of the selected noun in the middle column and in the left column s/he sees a list of adjectives or participles (respectively) which were found in real texts as modifiers of the selected noun.

An important feature of CoSyCo is that it is possible to choose the source of examples with the expression in question to be shown on the screen. A user can leave the default “All” option on or can select one of the five subcorpora from a drop-down list; then the list of example sentences shown contains only those from the selected subcorpora. Although the example usually includes one sentence, it is possible to have a look at a broader context with the help of a link to the source text placed after the example.

[12] The figure after the verb shows how often the word is found in the whole corpus. The figures after words in the middle and right columns show absolute frequencies of respective word combinations in the whole corpus. Unfortunately, the problem of duplicates in the corpus is still being resolved, so the figures currently on the screen are not accurate.

6. Adjective + NOUN combinations across CoSyCo subcorpora and other resources

To try comparing the output of CoSyCo with that of existing resources of similar size we took the lists of most frequent adjectives used with the noun *вирус*‘virus’ in CoSyCo, RuTenTen and GICR.

The table below shows top ten most frequent adjectives used with this noun in the average CoSyCo collection with the figures for the same words in RuTenTen and GICR[13]. For each word “a” column contains absolute co-occurrence figures for the combination of this adjective with *вирус*‘virus’ in this corpus; “b” column shows how often this pair is found as compared to the total number of any adjective+ *вирус* ‘virus’ combinations in the corpus.

Table 7. Adjectives+ *вирус* ‘virus’ in CoSyCo, RuTenTen and GICR

	CoSyCo		RuTenTen		GICR	
	A	b	a	B	a	b
	42,321		95,040		25,267	
КОМПЬЮТЕРНЫЙ ‘computer’	5,331	0.125965	12,257	0.128967	2,612	0.1033759
НОВЫЙ ‘new’	3,030	0.071595	9,483	0.099779	2,346	0.0928483
ОПАСНЫЙ ‘dangerous’	2,628	0.062096	5,237	0.055103	1,676	0.0663316
СМЕРТЕЛЬНЫЙ ‘deadly’	1,899	0.044871	1,934	0.020349	823	0.0325721
НЕИЗВЕСТНЫЙ ‘unknown’	1,195	0.028236	2,054	0.021612	442	0.0174931
СМЕРТОНОСНЫЙ ‘lethal’	893	0.021100	718	0.007555	236	0.0093403
СТРАШНЫЙ ‘dreadful’	589	0.013917	1,350	0.014205	709	0.0280603
ИЗВЕСТНЫЙ ‘known’	396	0.009350	1,717	0.018066	143	0.0056596
ОБЫЧНЫЙ ‘ordinary’	329	0.007773	576	0.006061	121	0.0047889

The figures show that in general the lists of adjectives are rather similar, and the variations in their frequencies may be explained by differences in the corpus structure, the style and genre differences of texts constituting them.

We also analyzed lists of top 100 most frequent adjectives used with this noun from the point of view of semantic classes which could be identified there. The tables

[13] For GICR we analyzed only data from three out of four available segments—news, Zhurnalny zal and LiveJournal. It was not clear beforehand what picture VKontakte texts would give, so we decided not to include them without prior research.

below show data for two groups which proved to be most frequent in virtually all segments in CoSyCo and GICR.

The first group can be identified as describing various features of a computer virus. Table 8 shows that adjectives from this group penetrate most of modern text styles and genres, with the adjective *komp'uterny*'computer' outweighing all other words in this group.

Table 8. 'COMPUTER' group of adjectives used with *вирус* 'virus'

	Komp'uterny 'computer'		*pochtovy* 'mail'		*mobilny* 'mobile'		total for group	
Source	a	b	a	b	a	b	a	b
CoSyCo news	1,121	13.18%	7	0.08%	115	1.35%	1,365	16.05%
CoSyCo compnews	990	24.93%	130	3.27%	98	2.47%	1,441	36.29%
CoSyCo Librusec	2,869	10.99%	(24)	0.09%	(29)	0.11%	3,538	20.05%
CoSyCo Wiki	276	25.72%	4	0.37%	7	0.65%	334	31.13%
CoSyCo science	50	11.99%	1	0.24%	5	1.20%	73	17.50%
GICR news	432	11.11%	24	0.62%	18	0.46%	517	13.30%
GICR zhurzal	45	11.57%	2	0.51%	—	—	54	13.88%
GICR Livejournal	2,135	10.31%	34	0.16%	78	0.37%	2,687	12.80%

The second group includes various adjectives united by the component 'know' in their lexical meaning. These words are also widely used to characterize a virus in every segment.

Table 9. 'KNOWN' group of adjectives used with *вирус* 'virus'

	novy 'new'		*neizvestny* 'unknown'		*izvestny* 'known'		total for group	
Source	a	b	a	b	a	b	a	b
CoSyCo news	949	11.16%	187	2.19%	57	0.67%	1,245	14.64%
CoSyCo compnews	621	15.64%	130	3.27%	117	2.95%	891	22.44%
CoSyCo Libr	1,417	5.43%	839	3.21%	183	0.70%	2,883	11.04%
CoSyCo Wiki	34	3.17%	37	3.45%	23	2.14%	107	9.97%
CoSyCo science	2	0.48%	1	0.24%	5	1.20%	9	2.16%
GICR news	692	17.80%	66	1.70%	32	0.82%	809	20.81%
GICR zhurzal	18	4.63%	14	3.60%	—	—	39	10.03%
GICR ljournal	1,636	7.79%	362	1.72%	111	0.53%	2,296	10.94%

Tables 8 and 9 show that results are quite comparable and in general both groups demonstrate similar tendencies in respective segments of different sources.

7. Conclusion

In the paper we have described the text corpora and methods used to create CoSyCo, a corpus of syntactic co-occurrences which provides information on syntactically related words in Russian. Currently there is a lot of room for improvement: there is a need to address deduplication issues, to increase the number of word combinations available on the site, as well as to improve its interface and the quality of output. We also understand the necessity to conduct a thorough comparison of the output of CoSyCo with that of the existing resources of similar size, and intend to do this in the nearest future.

References

1. *Belikov V., Selegey V., Sharoff S.* (2012), Preliminary considerations towards developing the General Internet Corpus of Russian [Prolegomeny k proektu General'nogo internet-korpusa russkogo yazyka], Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog" 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoii Konferentsii "Dialog 2012"], Bekasovo, vol. 1, pp. 37–49.
2. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Corpus as language: from scalability to register variation [Korpus kak yazyk: ot masshtabiruemosti k differentsialnoi polnote] Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog" (2013) [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoii Konferentsii "Dialog" (2013)], Bekasovo, vol. 1, pp. 83–96.
3. *Klyshinsky E., Kochetkova N., Litvinov M., Maximov V.* (2011), Method of POS-disambiguation using information about words co-occurrence (for Russian), Proceedings of the annual meeting of the Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL), Hamburg, pp. 191–195.
4. *Klyshinsky E., Ermakov P., Lukashevich N., Karpik O.* (2016) The Corpus of Syntactic Co-occurences: the First Glance, in Proc. of the Fifth International Conference on Analysis of Images, Social Networks and Texts (AIST 2016), pp. 85–90.
5. *Klyshinsky E.*(2017) The Freedom of the Russian Syntax is Slightly Exaggerated, in Proc. of Workshop on New Information Technologies in Automated Systems, pp. 112–116.
6. *Klyshinsky E., Lukashevich N.* (2017) Corpus of Syntactic Co-Occurrences: A Delayed Promise, Artificial Intelligence and Natural Language: 6th Conference, AINL 2017, St. Petersburg, pp. 121–131.
7. *Kormacheva D., Pivovarova L., Kopotev M.* (2014), Automatic Collocation Extraction and Classification of Automatically Obtained Bigrams, in Proceedings of Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014), pp. 27–33 .
8. *Lukashevich N., Klyshinky E., Kobozeva I.* (2016), Lexical research in Russian: are modern corpora flexible enough?, Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialog" (2016) [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi Konferentsii "Dialog" (2016)], Moscow, pp. 385–397.