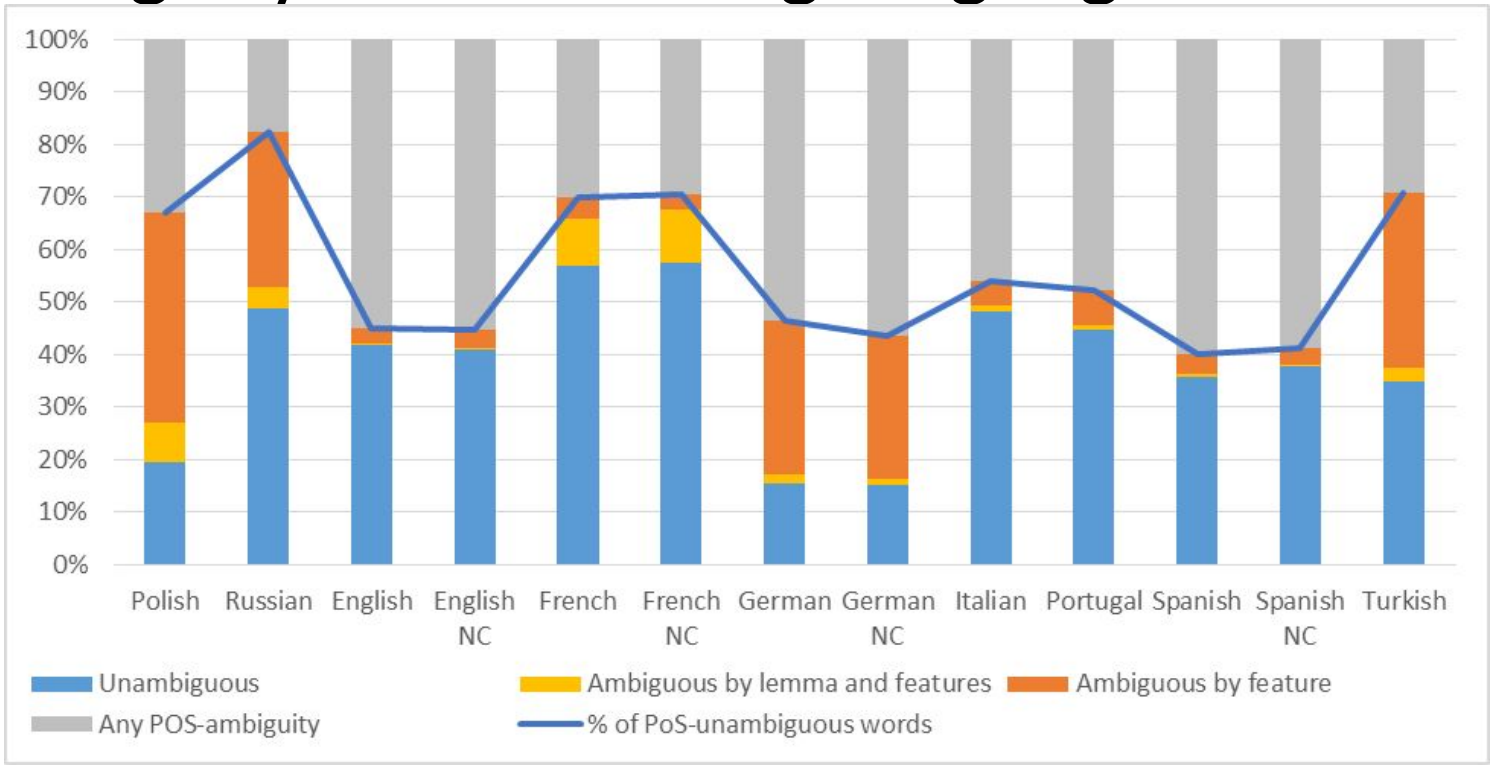
CREATING A CORPUS OF SYNTACTIC CO-OCCURRENCES FOR RUSSIAN

Moscow, Dialog'2018
June 1, 2018
Klyshinsky E.S., Lukashevich N.Yu, Kobozeva I.M.
eklyshinsky@hse.ru, natalukashevich@mail.ru, kobozeva@list.ru

Agenda

1. Motivation
2. Templates
3. Corpora and Results
4. Numerical Evaluation
5. Qualitative Evaluation

Motivation:
PoS-ambiguity differs among languages



Hypothesis

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.

Templates: Noun→Adj

I. Prep + (Adv) + (Adj^{+}) + N

(1) *Вероника повернулась, чтобы встретиться*
Veronika turned to look
с мягкими зелеными глазами.
into soft green eyes.

Templates: Adv→Adj

V. Prep/Noun/Conj + Adv + Adj

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

Templates: Verb→Noun

II. $|sign (Prep) + (Adv) + (Adj^{+}) + N + V

(a) ***Российские аналитики соглашаются,*** *что*
Russian analysts agree *that …*

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

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

Templates: Verb→Noun

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

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

IV. 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.*

Templates: Participle↔Noun

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***
The investigated problem
находится на стыке дисциплин.
is at the intersection of several domains.

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*** компанией, …
The system ***integration performed*** *by the company…*

Templates with PoS-ambiguity

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

(a) (Prep)NP = (Prep) + ?**Adj**/ Part + Adj + Noun

(8a) В Москве прошло вручение премии имени Елены Мухиной, которой награждаются люди

The ceremony of Elena Mukhina's award, which is granted to people

***с ограниченными физическими способностями**.*

***with limited physical abilities**, took place in Moscow.*

(b) (Prep)NP = (Prep) + ?**Adj**/ Noun + Adj + Noun

(8b) ***Прямая длинная линия*** *лезвия была скошена к концу.*

The ***direct long line*** *of the blade was slanted towards its end.*

Templates with PoS-ambiguity

IX. If NP or PP is at the end of the sentence and its last word is ambiguous between a noun and a verb or a noun and an adjective or participle, 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.

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

(9a) Он уставился ***на лобовое*** ***стекло***.
He stared at the ***front*** ***window***.

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

(9b) Предстоит долгий путь до ***финишной*** ***прямой***.
It is still a long way to the ***home*** ***straight***.

Templates with PoS-ambiguity

X. If NP or PP is followed by another PP and the last word of the first phrase is ambiguous between a noun and a verb or a noun and an adjective or participle, this last word should be considered a noun.

X (a) $[(\text{Prep}) + \ldots + ?\text{Noun/V}]_{\text{PP/NP}} + \text{PP/NP}$

(10a) Он разбил ***оконное*** ***<u>стекло</u>*** *в школьном коридоре.*

He broke a ***window*** **pane** in the school's passage way.

X (b) $[(\text{Prep}) + \ldots + ?\text{Noun/Adj/Part}]_{\text{PP/NP}} + \text{PP/NP}$

(10b) ***Сводные*** ***<u>данные</u>*** *о значениях параметров …*

The ***integrated*** ***<u>data</u>*** on the parameters …

Improving results

When we assessed how complete the database of combinations was, we found that a certain part of vocabulary was missing, because words which are grammatically ambiguous in all their forms in Russian were disregarded during processing.

Ученый — 'a scientist'_N / ‘learned, academic’_Adj

To avoid this, we had to lift certain restrictions in several templates.

Templates with PoS-ambiguity

XI. 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.

(11a) Вырезки не были ***разложены*** ***в*** ***хронологическом порядке*****.**

The cuttings were not ***placed*** ***in chronological*** ***order*****.**

(11b) ***Личный*** ***состав*** ***размещен*** *в закрытом городке.*

The military personnel was placed *in a restricted-access town.*

Templates with PoS-ambiguity

We had to allow words ambiguous between a noun and an adjective in templates I and IX.

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.

Used corpora

	CoSyCo subcorpora	mln words	%
1.	News sites (14 sources)	1 400.9	8.07%
2.	Popular science and IT news sites (8 sources)	142.4	0.82%
3.	Lib.rus.ec fiction collection	~15 000.0	86.38%
4.	Science sites (18 sources)	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%

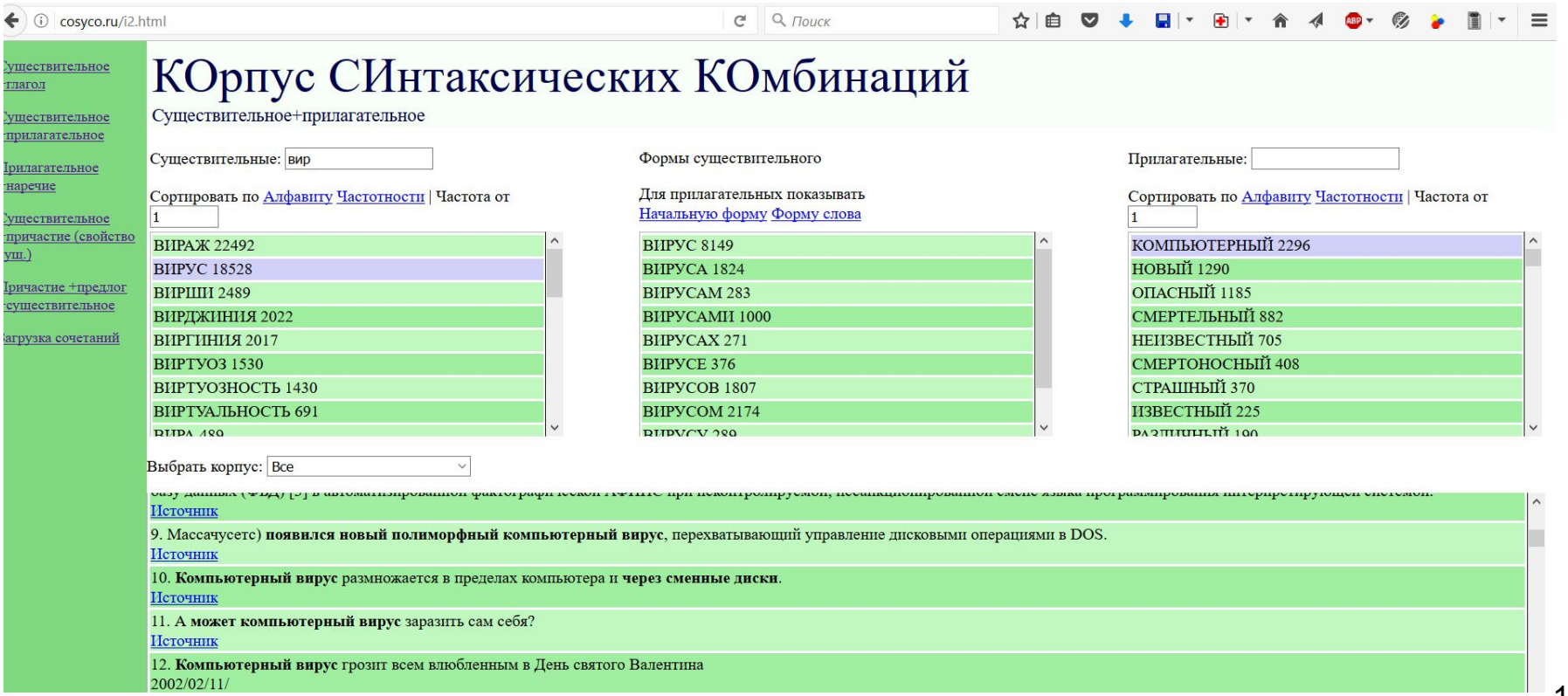
Results (combinations)

Combination	**Lemma combinations, mln**		**Token combinations, mln**		**Total occurrences, 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	

Results (dictionary size)

Combination	nouns		adjectives		verbs	
	old	new	old	new	old	new
noun+adj	67000	71000	41000	42000		
verb+prep+noun	73000	73000			28000	28000
participle+noun	52000				20000	
participle+prep+noun	40000				15000	

http://cosyco.ru/



Known Problems

- Initial form selection
  - *В ПОЛЕ (ПОЛА, ПОЛЕ, ПОЛЯ)*
- Deduplication
- Number of syntactical combinations

Evaluation

Combinations from SynTagRus found in CoSyCo

Combination	Found in SynTagRus	Among them in CoSyCo	% Found
Verb+Noun(+Prep)	100 125	81685	81.5
Noun+Adj	60485	58077	96.0

Evaluation

Comparison of I-RU and CoSyCo vocabularies

PoS	Frequency	I-Ru		CoSyCo	
		Not found in CoSyCo	Total	Not found in I-Ru	Total
Noun	>1000	226 (4,3%)	5229	523 (3,1%)	16881
	>500	534 (6,4%)	8376	1333 (6,0%)	22209
	>100	3887 (18,4%)	21122	7987 (21,7%)	36866
	>10	30298 (49,1%)	61720	27102 (46,3%)	58524

Evaluation

Comparison of I-RU and CoSyCo vocabularies

PoS	Frequency	I-Ru		CoSyCo	
		Not found in CoSyCo	Total	Not found in I-Ru	Total
Adjective	>1000	10 (0,6%)	1677	523 (3,1%)	13635
	>500	22 (0,8%)	2728	1333 (6,0%)	16705
	>100	405 (6,4%)	6312	7987 (21,7%)	24481
	>10	4014 (28,3%)	14192	27102 (46,3%)	33194

Evaluation

Comparison of I-RU and CoSyCo vocabularies

PoS	Frequency	I-Ru		CoSyCo	
		Not found in CoSyCo	Total	Not found in I-Ru	Total
Verb	>1000	21 (0,9%)	2291	197 (2,0%)	9975
	>500	56 (1,6%)	3601	725 (6,1%)	11975
	>100	426 (5,2%)	8153	4073 (24,6%)	16561
	>10	3670 (22,5%)	16291	10598 (45,6%)	23219

Quantitative Comparison

In order to compare the output of CoSyCo with that of existing resources of similar size we analyzed lists of most frequent adjectives used with the noun *вирус ‘virus’* in CoSyCo, RuTenTen and GICR.

Top 10 most frequent adjectives + *вирус* ‘virus’

Adj+Virus_N	Cosyco		RuTenTen		GICR	
	42321		95040		25267	
КОМПЬЮТЕРНЫЙ ‘computer’	5331	0.126	12257	0.129	2612	0.103
НОВЫЙ ‘new’	3030	0.072	9483	0.099	2346	0.093
ОПАСНЫЙ ‘dangerous’	2628	0.062	5237	0.055	1676	0.066
СМЕРТЕЛЬНЫЙ ‘lethal’	1899	0.045	1934	0.020	823	0.033
НЕИЗВЕСТНЫЙ ‘unknown’	1195	0.028	2054	0.021	442	0.018
СМЕРТОНОСНЫЙ ‘lethal’	893	0.021	718	0.008	236	0.009
СТРАШНЫЙ ‘dreadful’	589	0.014	1350	0.014	709	0.028
ИЗВЕСТНЫЙ ‘known’	396	0.009	1717	0.018	143	0.006
ОБЫЧНЫЙ ‘ordinary’	329	0.008	576	0.006	121	0.005

‘COMPUTER’ group of adjectives used with *вирус* ‘virus’

Source	***компьютерный ‘computer’***		***почтовый ‘mail’***		***мобильный ‘mobile’***		**total**	
CoSyCo news	1121	13.18%	7	0.08%	115	1.35%	1365	16.05%
CoSyCo compnews	990	24.93%	130	3.27%	98	2.47%	1441	36.29%
CoSyCo Librusec	2869	10.99%	24	0.09%	29	0.11%	3538	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	2135	10.31%	34	0.16%	78	0.37%	2687	12.80%

‘KNOWN’ group of adjectives used with *вирус* ‘virus’

Source	***новый ‘new’***		***неизвестный ‘unknown’***		***известный ‘known’***		**total**	
CoSyCo news	949	11.16%	187	2.19%	57	0.67%	1245	14.64%
CoSyCo compnews	621	15.64%	130	3.27%	117	2.95%	891	22.44%
CoSyCo Librusec	1417	5.43%	839	3.21%	183	0.70%	2883	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	0.83%	809	20.81%
GICR zhurzal	18	4.63%	14	3.60%	-	-	39	10.03%
GICR Livejournal	1636	7.79%	362	1.72%	111	0.53%	2296	10.94%

http://cosyco.ru/