Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

LEXICAL RESEARCH IN RUSSIAN: ARE MODERN CORPORA FLEXIBLE ENOUGH?

Lukashevich N. Y. (natalukashevich@mail.ru)
Moscow State University, Moscow, Russia

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

Kobozeva I. M. (kobozeva@list.ru)
Moscow State University, Moscow, Russia

The article discusses what modern tools offer for a corpus-based lexical research in Russian. As an example we analyzed how the adjective *gordy* 'proud' is used in modern news texts. We studied data from such resources as two general Russian language corpora (RNC, GICR) and a corpus of syntactic co-occurrences containing information on syntactic relations of words for Russian (CoSyCo[1]). If a corpus includes a variety of genres and allows to make fine-grained distinctions between text sources, it helps to highlight important style- and genre-dependent differences. Our comparison has demonstrated that there are quite significant differences in the usage of *gordy* which become clear when we study general news and IT news corpora separately, however, in general they show certain similar tendencies. It is also shown that when more varied genres are taken into account it may make more visible such style and genre features which it is not so easy to notice otherwise.

Key words: corpus-based research, flexibility, words co-occurrence, Russian lexis, lexical semantics

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

НАСКОЛЬКО ГИБКИ КОРПУСА ДЛЯ ЦЕЛЕЙ ИССЛЕДОВАНИЯ РУССКОЙ ЛЕКСИКИ

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

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

Кобозева И. М. (kobozeva@list.ru)
МГУ им. М. В. Ломоносова, Москва, Россия

В работе обсуждаются возможности, предлагаемые современными средствами для корпусных исследований лексики русского языка. В качестве примера анализируется употребление прилагательного *гордый* в современных новостных текстах на материале данных двух общих корпусов русского языка (НКРЯ, ГИКРЯ) и корпуса синтаксической сочетаемости, содержащего информацию о синтаксических связях слов в русском языке (КОСИКО). Сравнение показывает, что, несмотря на наличие общих тенденций, есть значительные различия в употреблении слова *гордый* в текстах компьютерных новостей и новостей общей тематики. Делается вывод, что наличие текстов разных жанров и возможность проводить разграничения между ними с точностью до источника позволяет увидеть существенные жанровые и стилистические различия, не столь заметные при рассмотрении материала в общем.

Ключевые слова: корпусные исследования, гибкость, совместная встречаемость слов, русская лексика, лексическая семантика

Much has been said in the ongoing discussion what kind of corpus a linguist/lexicographer needs. Such features of a corpus as its representativeness, volume, accessibility, etc. have been widely discussed. Recently the concept of register variation (Belikov et al, 2013) came into focus. It has been mentioned that this approach which allows to account for heterogeneity of language data is relevant not only for sociolinguistic studies, but also for a whole lot of linguistic tasks including comparative studies of texts from different genres (Belikov et al, 2012).

The issue of how to apply this approach to the sphere of genre and style is not that simple. For different research we may need different degrees of what might be called genre and style granularity in a corpus. Besides, a researcher may not be aware in advance of where and what kind of differences the study will reveal, so ideally (s)he needs a possibility to tune this granularity in accordance with the current demands.

In this paper we show that the ability to make fine-grained distinctions between sources in a corpus and to compare texts of similar but not identical genres may be of crucial importance.

1. Existing corpora and flexibility

It is a fact that modern linguistic and lexicographic research is mostly conducted on corpus data. Such studies depend much of the quality of corpora, the type(s) of phenomena covered by these resources, their structure, their limitations, etc.

It is also generally understood that the more a corpus allows tailoring the initial data and search in accordance with the needs of a particular study, the more varied tasks can be solved with its help.

Sites of modern corpora offer a variety of tools aimed to help researchers. Quite a lot of resources allow to refine selection features of words in the search string. These include Russian National Corpus (RNC) (Lashevskaja, Plungian, 2003), the General Internet Corpus of Russian (GICR) (Belikov et al, 2012), Sketch Engine (Kilgariff et al., 2004), etc.

Sketch Engine provides word sketches—corpus-based summaries of the grammatical and collocational behaviour of a selected word.

RNC Sketches[2] project also generates sketches using syntactically tagged texts from RNC; its resulting output contains information on syntactical relations between words.

Until quite recently the existing corpora for Russian did not offer a way to create a subcorpus of your choice. However, at the moment several corpora provide this opportunity to a researcher.

One of them is GICR, which was developed as a resource that allows to apply the method of segmental statistics in a wide range of linguistic tasks (Belikov et al, 2012). Its interface makes it possible to select not only the corpus segment, but also the type of text sources according to the list of segment-specific attributes (such as the author's year of birth, place of birth, gender for blogs, the name of the source for news, etc).

Another one, RNC has previously only allowed to select text sources from the main subcorpus on the basis of certain features. However, quite recently a similar possibility appeared in other subcorpora including its newspaper subcorpus[3].

In this paper we will focus on how using this option affects the results of a lexical study.

2. Data collection

Our aim was to study the usage of the word *gordy* 'proud' in such sphere as news texts. We have analysed how the word is used in the first 1,000 contexts of the main subcorpus of RNC, its subcorpus of newspaper texts, and the news segment in GICR.

RNC's newspaper subcorpus (as of 12/12/2015) included texts of the following 7 sources:

[2] (http://ling.go.mail.ru/synt/)

[3] Unfortunately, this happened after the data for this research was collected and analysed, so we were not able to include it in this paper.

Total	**173.5 mln (words)**	**100%**
Izvestia	9,282,250	5.35%
Komsomolskaya Pravda	44,867,100	25.86%
Novy region	25,174,850	14.51%
RBK Daily	26,424,050	15.23%
RIA Novosti	15,545,600	8.96%
Sovetsky sport	12,977,800	7.48%
Trud	39,228,350	22.61%

GICR news segment (as of 03/02/2016) included texts from the four sources as follows:

Source	**851 mln (words)**	**100%**
lenta	110,418,290	12.97%
regnum	218,025,910	25.60%
ria	337,669,976	39.65%
rosbalt	185,456,412	21.78%

For both corpora we obtained contexts with *gordy* followed directly by an (in) animate noun. For GICR we used search queries with the switched on by default option of deleting duplicates in the results.

Besides these two, we used a self-designed collection of texts, which is flexible to the highest degree as a result. We took as such a collection of the news subcorpus of CoSyCo—a corpus of syntactic co-occurrences containing information on syntactic relations of words for Russian, which is currently being developed for the purposes of teaching Russian as a foreign language.

CoSyCo news subcorpus contains texts from the following sites:

News sites:	**982.2 mln (words)**	**100%**
lenta.ru (lenta)	71,300,115	7.26
RBK (rbc)	61,933,721	6.31
RIA Novosti (ria)	409,971,920	41.74
Nezavisimaya gazeta (ng)	48,923,879	4.98
Vzglyad (vz)	72,370,767	7.37
Rossiyskaya gazeta(rg)	71,467,194	7.28
Commersant (commersant)	140,585,843	14.31
Polit.ru (polit)	49,697,364	5.06
Utro.ru(utro)	45,770,623	4.66
Ibusiness.ru (ibusiness)	10,131,894	1.03
IT news	**113 mln (words)**	**100%**
Мембрана (membrana)	7,391,018	6.40
CNews (cnews)	35,830,813	31.04
Компьютерра (computerra)	28,068,619	24.32
Компьюлента (compulenta)	16,204,248	14.04
PCWeek (pcweek)	27,924,230	24.19

Since our newspaper collection was not tagged, we used the software tool we created during the development of CoSyCo database[4]. The site cosyco.ru provides an easy access to a database of such syntactically connected word combinations as adjective+noun and verb+preposition+noun. The paper (Klyshinsky et al., 2011) describes the method that was used to create this database. In the current project, we have created a tagger for a more flexible clause extraction. The developed software tool processes more than 15 mln word tokens per minute, thus it is very convenient for small collections (e.g. texts of one newspaper for one year). Processing large corpora (e.g. the whole Librusec collection[5]) takes about 12–16 hours. The tool works as a Windows command-line local application taking input in an XML-file, which contains a query and a list of input and output files.

Following such tools as NLTK and Stanford Parser, our software allows writing regular-expression-like queries including an ambiguity analysis. Words in input queries can have not just one but several initial forms or parts of speech. Like the Universal Dependencies[6] initiative, we separate feature name and its value. Such separation of name and value of features helps to make the denotation of words coordination or its absence relatively simpler. For example, the query

```
(in;prep;)(;adv;)(;adj;gender+, number=pl, case+ & ;noun;
 case-)(;adj;case-  & ;noun;gender+,number=pl, case+)
```

matches a clause that includes the preposition *'in'*, an adverb followed by two words that are ambiguous for part of speech (between adjective and noun). Two last words are in plural form **`(number=pl)`**; the first adjective and second noun are coordinated by gender and case **`(gender+, case+)`**, while the first noun and the second adjective **`(case-)`** are not coordinated. The main features of the developed query language are described in (Vlasova et al, 2016), however, in this project we used a slightly different notation.

We ran our tool over the selected collection using a simple query

```
(ГОРДЫЙ;adj; case+, gender+, number+)(;adj; case+, gender+,
number+)*(;noun; case+, gender+, number+).
```
[7]

This means the word *gordy* in all its forms that is possibly followed by an iteration of adjectives; the clause is finished by a noun; all words have the same values of case, gender, and number. Our tests showed that this query has high recall and precision. We selected the first 1,000 of sentences from the RNC's main subcorpus output for the word gordy. Our tool

[4] http://cosyco.ru/

[5] http://lib.rus.ec/

[6] http://universaldependencies.org/

[7] As for duplicates (which are a problem for some of the sources used), they had to be deleted manually, as currently there is no automatic deduplication in CoSyCo. (By duplicates here we mean exact copies of sentences, as there were also cases when only a part of the sentence was repeated (mostly when official political comments were reported)—such cases were counted as separate instances.)

returned 364 sentences with just 8 mistakes found by the assessor and 100% recall. Mistake rate on our newspaper collection varies from 1% up to 13% with the average at 5%.

Out of the relevant contexts for each resource we compiled a list of nouns, which co-occurred with *gordy* in its texts.

These lists of nouns were analyzed from the point of view of semantic classes which could be identified there. At the same time the correlations between the semantic class of a noun and the semantic interpretation (sense) of its adjectival modifier *gordy* were studied.

3. Meanings of gordy and semantic classes of co-occurring nouns

During the analysis of 1,000 contexts from RNC's main subcorpus we came to distinguish 5 senses of *gordy*[8]:

1) *gordy X* ≈ a person X whose behavior shows that (s)he has a sense of dignity and self-respect:

(1) *(Katya) suddenly felt doubly happy: her beloved was not an ordinary man, no, he was tough, proud and pure.*[9] [RNC, E. Kazakevich, Zvezda][10]

2) *gordy X* ≈ a person X who is feeling pleased with the fact that smth that (s)he (or someone associated with him/her) owns or smth (s)he (or the associates) achieved should make other people think better of him/her or rank him/her higher:

(2) *Alevtina is proud that she earns her living herself and does not depend on anybody…*[11] [RNC, V. Makanin, Otdushina]*[*

3) *gordy X* ≈ a person X who thinks of oneself as being better than other people and treats them with contempt because of that:

(3) *You had better go and stay with the guests, or they will think you are too proud*[12]. [RNC, A. Chekhov, V rodnom uglu]

[8] These senses correspond to slightly modified three senses of this word given in MAS. Unlike MAS, which unites all figurative meanings under *gordy2* sense, we actually singled them out as separate ones (namely, *gordy* 4 and 5) and added them to the three MAS senses related to person.

[9] (Катя) вдруг почувствовала себя вдвойне счастливой: ее любимый был не обычный человек, нет, он суровый, гордый и чистый. [Э. Г. Казакевич. Звезда (1946)]

[10] Authors' translation—here and below, except for (5).

[11] Алевтина горда тем, что зарабатывает на жизнь сама и ни от кого не зависит …. [В. Маканин, Отдушина (1977)]

[12] Ты бы посидела с гостями, а то подумают, что ты гордая. [А. П. Чехов. В родном углу (1897)]

4) (figurative) majestic, stately

> (4) *High, proud celestial mountain peaks were glimmering golden in the sunset sky.*[13] [RNC, V. Skripkin, Tinga]

5) (figurative) sublime, lofty, elevated:

> (5) *Fyodor thought…with proud, joyous energy, with passionate impatience, he was already looking for the creation of something…* [RNC, V. Nabokov. The Gift (M. Scammel, V. Nabokov, 1962)][14]

The resulting list of semantic classes of nouns modified by *gordy* (with minimal examples) is presented below. It is divided into four groups on the basis of combinability with the 5 senses of the adjective.

1. Nouns denoting persons:
   - a person in general or a male/female person:
     gordaya devushka 'a proud girl'
   - a person according to family status
     gordy otets 'a proud father'
   - a person according to their nationality/ethnicity:
     gordy amerikanets 'a proud American'
   - a person according to their social status
     gordy korol' 'a proud king'
   - a big group of people (community)
     gordy narod 'a proud people'

etc.

2. Nouns related to the situation of a person being proud (in sense 2):
   - nouns denoting a person as a possessor of smth or an agent of some deed:
     gordy pobeditel' 'a proud winner'
   - nouns denoting emotions experienced by a person who feels proud (in the sense 2):
     gordaya radost' 'proud joy'

3. Nouns referring to an object which is similar to a proud person in some respect:
   - an inanimate object:
     gordaya bashnya 'a proud tower'
   - an animal, a bird:
     gordy oryol 'a proud eagle'

[13] Высокие, гордые вершины небесных гор румянились в закатном небе. [В. Скрипкин. Тинга // «Октябрь», 2002]

[14] Федор Константинович …с какой-то радостной, гордой энергией… уже искал создания чего-то нового… [В. В. Набоков, Дар (1935–1937)]

4. Nouns denoting features of a person or results of human activity:

- general characteristics of a person (as a subject of mental and social activity):
  gordiy um 'a proud mind'
- emotional states, feelings, personality traits:
  gordoe spokoistvie 'proud tranquility'
  gordoe prezrenie 'proud contempt'
- features and qualities of appearance and movement:
  gordaya osanka 'proud demeanor'
- psychological states, processes, and 'products' of mental work
  gordaya mechta 'a proud dream'
- representational objects (linguistic units and expressions)
  gordoye imya 'a proud name'
- *gordaya nadpis*' 'a proud inscription'

etc.

Nouns in the first group above are supposedly used with the first or the third meaning of *gordy* (depending on whether the speaker assesses the given instance of behaviour positively or negatively), and the second group combines with the second meaning of *gordy*[15].

The third group of nouns covers all cases of metaphoric transfer when an object is compared to a proud person (realization of *gordy4*); the comparison is often based on the look typical for a proud person (i.e. holding one's head high, not bending, etc).

The fourth group includes cases (linked with *gordy5*) when the characteristic is transferred by metonymy from a proud person (1, 2 or 3) to something which can express pride (as a personality trait or emotion). Here we find not only "inherent" features of a person (related to appearance, mind, character, etc), but also "results" of social, mental, emotional activity of a person.

4. Results

To evaluate the variance for the word *gordy*, we calculated a feature vector including all nouns which co-occurred with *gordy*. In corpus-based research, the feature vector usually contains instances per million (ipm) value (Lyashevskaya and Sharoff, 2009). However, here we are interested in the changes in the frequency distribution. That is why we took the feature vector containing absolute values of words co-occurrences and normalized it on the sum of frequencies, i.e. calculated the conditional probability of meeting a noun in a context with the word *gordy*.

[15] Strictly speaking, nouns in the first group may also be used with *gordy* in the second meaning (*gordy*2): e.g. a musician may be called proud not only because of something done out of pride as a character trait, but also because of the emotion felt after the performance. It is also possible (but less typical) to think of an owner of X who treats others with contempt. So it will be more precise to say that nouns in the second group **tend** to be used with *gordy*2 and are separated from all the rest on this basis.

We started from the idea that by normalizing data across different sources (or groups of sources) we lose information on their genre and style. To show this we combined all CoSyCo collections of general news into one and calculated all co-occurrence frequencies for the resulting “average” collection. The same was done for IT news. This gave us a chance to compare frequencies in the “average” collection and by each source separately.

In the tables below we show data for several of the semantic classes with the highest scores (which also differed the most).

In each table we included several most frequent nouns and also figures for the whole class in the column Total.

For each word and group “a” columns contain absolute co-occurrence figures for the combination of the adjective *gordy* with this noun or this semantic class of nouns; “b” columns show how often this pair is found as compared to the total number of such pairs in the collection for this source; “c” columns contain relative frequency of the pair occurrence in the “average” collection (for CoSyCo data).

Of the 17 nouns denoting various representational objects (denoting the kinds of names) the following three were most frequent:

	imya **‘name’**			***zvanie*** **‘rank’**			***nazvanie*** **‘name’**			**total**		
	a	**b, %**	**c, %**	**a**	**b, %**	**c, %**	**a**	**b, %**	**c, %**	**a**	**b, %**	**c, %**
RNC 1000	14	3.91		5	1.40		10	2.79		31	8.7	
RNC news	**60**	**5.89**		**42**	**4.12**		**32**	**3.14**		**149**	**14.7**	
GICR news	**39**	**4.63**		**47**	**5.58**		**27**	**3.21**		**126**	**15**	
lenta	1	1.18	0.12	4	4.71	0.47	5	5.88	0.59	15	17.65	1.76
rian	13	4.71	1.53	13	4.71	1.53	11	3.99	1.29	39	14.13	4.58
regnum	18	6.72	2.11	12	4.48	1.41	3	1.12	0.35	37	13.81	4.34
rosbalt	7	3.14	0.82	18	8.07	2.11	8	3.59	0.94	35	15.70	4.11
CoSyCo news												
commersant	14	5.49	0.70	19	7.45	0.95	9	3.53	0.45	47	18.4	2.35
rian	28	4.03	1.40	18	2.59	0.90	21	3.03	1.05	73	10.5	3.64
rg	14	6.06	0.70	11	4.76	0.55	8	3.46	0.40	34	14.7	1.70
utro	3	1.96	0.15	9	5.88	0.45	13	8.50	0.65	26	17	1.30
vz	5	2.75	0.25	14	7.69	0.70	10	5.49	0.50	33	18.1	1.65
nezavisimaya	14	6.36	0.70	8	3.64	0.40	6	2.73	0.30	29	13.5	1.45
lenta	1	1.89	0.05	—	—	—	1	1.89	0.05	2	3.8	0.10
polit	1	1.28	0.05	—	—	—	2	2.56	0.10	4	5.2	0.20
rbc	2	2.13	0.10	6	6.38	0.30	10	10.64	0.50	18	20.2	0.90
ibusiness	3	6.82	0.15	1	2.27	0.05	5	11.36	0.25	9	22.5	0.45
AVG CoSyCo NEWS	**85**		**4.24**	**86**		**4.29**	**85**		**4.24**	**275**		**13.72**
IT news												
cnews	2	9.09	0.48	—	—	—				2	9.09	0.48
compulenta	6	13.95	1.45	1	2.33	0.24	3	6.98	0.72	11	23.26	2.66
computerra	39	14.18	9.42	13	4.73	3.14	21	7.64	5.07	87	26.55	21.01
membrana	3	9.68	0.72	1	3.23	0.24	1	3.23	0.24	7	16.14	1.69
Pc week	2	4.65	0.48	1	2.33	0.24	3	6.98	0.72	6	13.96	1.45
AVG IT NEWS	**52**		**12.56**	**16**		**3.86**	**28**		**6.76**	**113**		**27.29**

It is clear that the weight of this group varies a lot: from 3.8% for Lenta to 26.55% for Computerra. RNC and GICR as non-segregated sources also show rather high scores of 14.7% and 15% (as compared to scores below 5% for most other semantic groups and classes). GICR figures by source vary from 13.81% up to 17.65%. The difference in two Lenta.ru collections stands out. GICR contains a bigger version of this site including both news and analytics, while CoSyCo version contains news wire only. Thus, the difference between figures could be explained by the difference in the collections' style. It is also evident that IT news tend to take the higher end of the range, whereas sources with more varied topics score lower. The reason behind this may be that in the first group the necessity to name new objects (projects, organizations, etc) is higher. It should be noted that in such cases the word tends to be used in a humourous or ironic way[16].

Of the 24 nouns from another subgroup of representational objects (referring to a way of showing the name/class of the object) the word *nadpis'* 'inscription' stood out from the rest. For this subgroup the figures are much lower[17] than for the *imya* group, but the tendency still remains for IT news to occupy the upper end of the diapason.

Of the 33 words grouped as names of (emotional) states, feelings, personality traits two words show high frequencies, far outstepping all the rest:

	odinochestvo **'loneliness'**			***molchanie*** **'silence'**			**total**		
	a	**b**	**c**	**a**	**b**	**c**	**a**	**b**	**C**
RNC 1000	32	8.94		5	1.40		51	14.2	
RNC news	**174**	**17.08**		**12**	**1.18**		**195**	**19.1**	
GICR news	**137**	**16.27**		**22**	**2.61**		**168**	**20**	
lenta	12	14.12	1.41	2	2.35	0.23	14	16.47	1.64
ria	52	18.84	6.10	1	0.36	0.12	57	20.65	6.69
regnum	42	15.67	4.93	14	5.22	1.64	59	22.01	6.92
rosbalt	32	14.35	3.76	5	2.24	0.59	38	17.04	4.46
CoSyCo news									
commersant	35	13.73	1.75	2	0.78	0.10	41	16.1	2.05
rian	73	10.52	3.64	3	0.43	0.15	81	11.7	4.04
rg	43	18.61	2.15	3	1.30	0.15	56	24.2	2.79
utro	38	24.84	1.90	2	1.31	0.10	40	26.1	2.00

[16] Irony and humour are considered examples of the so-called *non bona fide* modus of discourse (Shilikhina 2014). They appear when there is intended incoherence in the utterance (i.e. a disruption in semantic cohesion within the utterance or an incongruity between the utterance and the situation described), which signals the presence of implicit meanings. Irony presupposes implicit negative deontic assessment.

IT news attracted our attention in this respect: when we checked the proportion of humourous and ironic contexts in IT news, we managed to find about one or two "serious" (bona fide) uses of *gordy* per 100 sentences. This figure for general news texts is lower (varies in the range of 10–30%), but is still presumably much higher than in fiction (this requires further research).

[17] For this reason we will not include these data here.

	odinochestvo 'loneliness'			*molchanie* 'silence'			total		
	a	**b**	**c**	**a**	**b**	**c**	**a**	**b**	**C**
vz	28	15.38	1.40	1	0.55	0.05	30	16.5	1.50
nezavisimaya	34	15.45	1.70	4	1.82	0.20	44	20.5	2.20
lenta	—	—	—	—	—	—	—	—	0.00
polit	8	10.26	0.40	1	1.28	0.05	11	14.1	0.55
rbc	24	25.53	1.20	—	—	—	24	25.5	1.20
ibusiness	3	6.82	0.15	—	—	—	3	7.5	0.15
AVG CoSyCo NEWS	**286**		**14.27**	**16**		**0.80**	**330**		**16.47**
IT news									
cnews	2	9.09	0.48	1	4.55	0.24	3	13.6	0.72
compulenta	14	32.56	3.38	—	—	—	14	32.6	3.38
computerra	29	10.55	7.00	6	2.18	1.45	35	12.73	8.45
membrana	7	22.58	1.69	—	—	—	7	22.6	1.69
Pc week	13	30.23	3.14	1	2.33	0.24	14	32.6	3.38
AVG IT NEWS	**65**		**15.70**	**8**		**1.93**	**73**		**17.63**

Odinochestvo (which accounts for 60–90% of the class weight) also shows remarkable variation from 7.5% (ibusiness) up to 32.56% (computerra), with IT news once again taking the higher end of the range. *Molchanie* shows the same tendency, though with lower figures. In polythematic sources *gordy* is more often used with other nouns from this group. Both these words show such high frequencies because they are used in clichéd expressions as desemantised phrasemes marking humour or irony in the utterance. Here the difference between the two versions of Lenta. ru becomes even more obvious: if these markers of humour and irony are not found in "normal" news at all, it is definitely not so for analytics (the figures are on a par with other polythematic sources).

The group of nouns naming a person according to nationality/ethnicity is remarkable in the sense that of the 86 nouns belonging to the group it is hard to name any which would be used more often than the rest (let alone do it consistently through several sources). The total class figures (which will be not given here for lack of space) show that the class accounts for more than 5% of *gordy* usage for many sources, and predictably the scores are higher for polythematic sources.

Of the 24 nouns referring to a big group of people (community) the following three were more frequent:

	narod 'a people'			*strana* 'a country'			*gosudarstvo* 'a state'			total		
	a	**b, %**	**c, %**	**a**	**b, %**	**c**	**a**	**b, %**	**c, %**	**a**	**b, %**	**c, %**
RNC 1000	5	1.40		4	1.12		—	—		9	2.5	
RNC news	**15**	**1.47**		**16**	**1.57**		**5**	**0.49**		**57**	**5.6**	
GICR news	**61**	**7.24**		**21**	**2.49**		**15**	**1.78**		**127**	**15.1**	
lenta	3	3.53	0.35	1	1.18	0.12	—	—	—	9	10.59	1.06
ria	17	6.16	2.00	3	1.09	0.35	2	0.72	0.23	26	9.42	3.05

	narod **'a people'**			***strana*** **'a country'**			***gosudarstvo*** **'a state'**			**total**		
	a	**b, %**	**c, %**	**a**	**b, %**	**c**	**a**	**b, %**	**c, %**	**a**	**b, %**	**c, %**
regnum	26	9.70	3.05	11	4.10	1.29	7	2.61	0.82	59	22.01	6.92
rosbalt	15	6.73	1.76	6	2.69	0.70	6	2.69	0.70	33	14.80	3.87
CoSyCo news												
commersant	5	1.96	0.25	9	3.53	0.45	1	0.39	0.05	23	9.1	1.15
rian	25	3.60	1.25	11	1.59	0.55	2	0.29	0.10	48	6.9	2.40
rg	6	2.60	0.30	2	0.87	0.10	1	0.43	0.05	13	5.6	0.65
utro	10	6.54	0.50	3	1.96	0.15	1	0.65	0.05	18	11.8	0.90
vz	2	1.10	0.10	2	1.10	0.10	1	0.55	0.05	9	4.9	0.45
nezavisimaya	5	2.27	0.25	1	0.45	0.05	—	—	—	13	6	0.65
lenta	1	1.89	0.05	—	—	—	—	—	—	2	3.8	0.10
polit	5	6.41	0.25	3	3.85	0.15	1	1.28	0.05	12	15.4	0.60
rbc	1	1.06	0.05	4	4.26	0.20	—	—	—	5	5.3	0.25
ibusiness	—	—	—	1	2.27	0.05	—	—	—	2	5	0.10
AVG CoSyCo NEWS	**60**		**2.99**	**36**		**1.80**	**7**		**0.35**	**145**		**7.24**
IT news												
cnews	—	—	—	4	18.18	0.97	—	—	—	4	18.18	0.97
compulenta	—	—	—	—	—	—	—	—	—	—	—	—
computerra	2	0.73	0.48	2	0.73	0.48	—	—	—	8	2.92	1.92
membrana	—	—	—	—	—	—	—	—	—	—	—	—
Pc week	—	—	—	—	—	—	—	—	—	—	—	—
AVG IT NEWS	**2**		**0.48**	**6**		**1.45**	**—**		**—**	**12**		**1.93**

Results predictably show that scores are higher for polythematic sources, especially for those more focused on politics.

5. Conclusion

As we can see, figures for "generalized" corpora change more or less in the same way, accurate to the selection of chosen sources. However, splitting the corpus into subcorpora leads to significant changes in the word usage distribution. Moreover, stylistically different parts of the same corpus show dramatic differences. IT news sources show a similar tendency.

Results demonstrate that at the first glance in general the usage of *gordy* in news and newspaper texts can be as varied as in fiction: most of the classes identified in fiction are present in many sources. However, it is clear that *gordy* tends to appear more frequently with nouns in several particular zones from the whole list of possible combinations. Such classes as nouns naming a person according to nationality/ethnicity, big groups of people or representational objects and desemantised phrasemes are prominent for polythematic news sources. For IT news names of representational objects and desemantised phrasemes are the most frequently used, scoring higher than the same classes in general news.

The choice of such zones predictably depends on the topics covered by the source. Another crucial factor is the style of the source. The more markers of humour and irony (Shilikhina 2014) in the source, the more probable it is that *gordy* is used not seriously and expresses negative attitude towards the object or event characterized as such.

In sum, it may seem obvious that if we combine together texts of different styles and genres we need to be able to study them separately. Otherwise if we study them as if they were a homogeneous text, we get results which conceal the existing genre and style features. Existing corpora are gradually becoming more flexible in this respect, as they start to allow separating the data from different sources. It remains an open question what particular styles and genres should be included in a corpus which is intended as suitable for various kinds of research. What degree of granularity it would be reasonable to ensure in such a corpus is also a matter of further studies.

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. *Kilgarriff A., Rychly P., Smrz P., Tugwell D.* (2004), The Sketch Engine, Proceedings of the XI Euralex International Congress, Lorient, France, pp. 105–116.
4. *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.
5. *Lashevskaja, O., Plungian V.* (2003), Morphological annotation in Russian National Corpus: a theoretical feedback, Proceedings of the 5th International Conference on Formal Description of Slavic Languages (FDSL-5), Leipzig, pp. 26–28.
6. *Shilikhina K. M.* (2014), Semantics and Pragmatics of Verbal Irony [Semantika i pragmatika verbal'noi ironii], NAUKA-UNIPRESS, Voronezh.
7. *Vlasova A. A., Korolyov D. V., Klyshinsky E. S.* (2016), Development of the software tool for searching of clauses in untagged texts [Razrabotka instrumental'nogo sredstva dlia poiska sintaksicheskikh konstruktsij v nerazmechennoj kollektsii tekstov], Proceedings of New Information Technologies in Automated Systems, Ekaterinburg, pp. 81–84.