Corpus-based Investigation of Language Changes: the Case of the RNC

Dr Svetlana Savchuk
Institute of Russian Language
Russian Academy of Science
savsvetlana@mail.ru

1. Introduction

Corpus-based approach is extremely essential and may be fruitful in research of language variation and correlation between norm and real usage. Variation is a fundamental property of a language that imparts to it flexibility, redundancy, possibility to express one meaning in different ways. There are orthoepic and accentologic variations on a phonetic level, orthographic and punctuation variations – on a graphic one, word-formative variants and different variants of forms of words and syntactic constructions on a grammatical level, variants of words, synonyms – on a semantic level.

There are no variants that are entirely identical; as a rule there is some specialization between them, some semantic or stylistic difference.

Real usage is an inexhaustible source of variations: a new phenomenon appears, comes up against the existing one, prevails and replaces it or gets out of use. The first period of the process is characterized by constantly changing correlation between old and new.

Thus if we are interested in problem of variation on a synchronic level, we study co-existing variants, their distribution among different spheres of functioning, social, professional variants of a language, *etc*.

If we occupy with the problem in diachronic aspect, we deal with changes in usage, with appearance and disappearance of variants, changing of correlation between two variants, with strengthening and slackening of trends, *etc.*

When several variants exist, they are usually differently estimated with regard to standard. A variant can be codified in the literary language or can stay aside the norm. At that there is quite a natural discrepancy: variants can fluctuate over a long period of time, while estimation of these variants in relation to the norms of the standard language changes continually.

It is quite obvious that estimation and recommendations are based on the scientific conception of language norm that meets the following requirements:

1. Norm must correspond with the system of language;

2. Norm is determined not only by system regularities of language but also by social estimation, thus it is, in a sense, a result of interaction between public choice and individual habits;

3. Codified norm, fixed in grammars and dictionaries, must correlate with usage (i.e. the spontaneous language norm);

4. Statistical values, such as frequency of occurrence, should be accepted as objective indicators of currency of language phenomena (Graudina, 1980: 63).

Thus a corpus of texts seems to be a reliable tool for research in linguistic norms and variation, because representing a state of a certain language of a certain period of time, it provides a linguist with material for a quantitative and qualitative analyses of this state of language. Naturally reliability of this research depends on parameters of the corpus: its capacity, coverage, linguistic information represented in it. For investigations concerned with a wide range of linguistic phenomena only modern large corpora are suitable. The paper shows that the National Corpus of the Russian Language has all necessary quantitative and qualitative characteristics, so that it is able to provide an adequate set of examples for various types of linguistic researches. The investigation of variants of verb government of several synonymous verbs exemplifies the potentiality of the RNC.

2. The Russian National Corpus

The Russian National Corpus (RNC) is being created within the program of the Russian Academy of Science by a large group of specialists from Moscow, St. Petersburg, Voronezh and other Russian university centers since 2003. Although works under the project are still going on, the corpus can be already used (and it is used) for research and educational purposes.

The RNC meets all requirements that for large contemporary corpora of texts (EAGLES 1996, Sinclair, 2003, Butler, 2004, Reference Guide, 2007). First of all it is of great size. At present it contains about 140 million tokens. Great size of the corpus allows posing and solving with its help different linguistic problems being sure in reliability of results.

Besides that the RNC is a representative corpus that may be regarded as a reflection of language usage for a certain epoch. The corpus contains only entire texts of different forms of speech - spoken, written, electronic - and of different functional spheres: fiction, journalism, memoirs, academic prose, administrative documents, religious texts, everyday life dialogues, poetry, etc.

The RNC is an annotated corpus: all texts are supplied with metainformation about author, creation date, functional sphere, text type, domain, *etc*., so that users can form their own subcorpora. Every word has morphological and semantic annotation. Texts in the spoken subcorpus are supplied extra by sociological annotation that shows sex, age, and occupation of a speaker.

The corpus is available on open access on the site ruscorpora.ru. Search on the corpus is provided by the Yandex.Server system.

Having projected the corpus the creators were guided by the world experience of corpora making and used methods that are accepted by the world practice and that were adopted for the Russian language and linguistic traditions of its research. (Sharoff, 2004).

The Russian National Corpus consists of the following subcorpora:

- Corpus of modern written texts, 1950–2006 – a core corpus of 100 million tokens (94 million is available)
- Corpus of written texts, XVIII – the first half of XX (44,5 million tokens)
- Corpus of Spoken Language (more than 4,5 million tokens)
- Poetic corpus
- Dialect corpus
- Parallel aligned corpus

As may be seen, there are two main chronological components in the RNC: that is corpus of modern texts (of the period from the second half of the XX century up to our days) and the diachronic one that includes texts of XVIII - first half of XX. Such corpus has independent value for both historians of the language and researchers of its modern state because represents any language phenomenon in its evolution, in diachronic aspect.

3. Corpus of modern texts: the Russian language today

3.1. Principles of selection of texts

To meet the modern requirements texts for this corpus were selected according to external criteria. In (Sharoff, Savchuk 2004) there was indicated that the parametric classification of Sinclair and the stylistic one, based on the traditional definition of text types, do not come into conflict with each other and can be integrated in one system of texts classification. That's why the final system of texts annotating, accepted in the RNC, contains parameters such as sphere of functioning, type, chronotop.

Besides that, the core of the parametric approach recommended by EAGLES and followed by all European corpora engineers, is to build first of all a theoretic model of the corpus. It is built by listing of all possible valid attribute combinations, excluding the cases of low probability; and so all possible variants of texts came of this enumeration. Then this model is filled with real texts in definite proportions, so that every attribute combination was represented by at least several texts. (Butler, 2004: 152).

The RNC is supposed to reflect the real usage of the language of a certain period, so the compilers decided to minimize the interference in real balance of texts functioning in different spheres of discourse. Toward this end a ratio of texts of different functional spheres was specified on the basis of preliminary analysis of existing corpora, sociological researches, monitoring of the book market, press, electronic resources of the Runet, *etc*., forecasting of research interests of users.

Compiling the corpus content the researchers endeavor to restrict their interference only by selection of editions. Public significance, estimation of specialists and critics, readers' demand - that were the main factors taken into account. Newspapers and magazines, including literary ones, represented in wide political, thematic, regional range, are included in the corpus entirely. Thus we can assume that the RNC represents a real thematic and genre ratio of published written texts of the period from the end of XX century to the beginning of XXI century. As regards unpublished texts (manuscripts, *etc*.) and periodicals before the 1990s, preparation of these texts are very laborious and expensive; at the moment selection of these texts is provided according to genre and thematic variety within every sphere of functioning.

3.2. Composition

The Corpus of modern texts is balanced by spheres of functioning, genre and thematic structure of texts, dating.

Spheres of functioning	**Tokens**	**Percentage**
Fiction	34190863	36
Journalism	39144005	43
Teaching and scientific	10694746	11
Official	1736929	2
Advertisement	515469	0,5

Church and theological	1336039	1
Everyday life	435807	0,5
eCommunication	1275240	1
Spoken	4634611	5
Total	93963709	100

Table 1: The distribution of modern texts according to spheres of functioning

According to date of creation the texts are distributed as follows:

- 1950-1986 - 18 percent
- 1987-1992 - 4 percent
- 1993-2000 - 26 percent
- 2001-2005 - 52 percent

4. Diachronic corpus: the Russian language yesterday

4.1. Principles of text selection

Building of the diachronic corpus has several specific characters. First of all it is necessary to take into account that most part of texts of this period are not available in electronic form. This concerns private letters, journalism, scientific articles, advertisement, *etc*. These texts are not republished, but editions that remain are in very bad state, so that it is quite impossible even to scan and recognize them. There are, certainly, published letters written by famous writers, significant scientists, political figures, but these letters are stylistically and thematically closer to journalism or to fiction than to private letters.

Oral speech of that period is obviously quite inaccessible. Although there are several interviews and transcripts of political speeches, court proceedings, these texts can not be considered as spontaneous speech, they are prepared texts. But there is one source of oral speech of the first half of the XX century. That is text content of Soviet films of the 1930-1950s that will form a part of a multimedia subcorpus of the RNC.

Besides that one should consider that the period the XIX - first half of the XX is extremely heterogeneous linguistically. In the XIX century standards of the modern language (that is defined as *from Pushkin to our days*) were developed, modern system of styles and genres were formed. That was in many respects a natural result of conscious efforts of writers, literary critics and other personalities of those times, that's why quota of fiction and critics for that period is to be higher than for corpus of modern texts.

The first half of the XX century is a period of complicated political and ideological reorganizations in Russia that certainly was reflected in language of that time. From the sociolinguistic point of view period before October 1917 can be considered as sequential to the XIX century. At the same time this period is characterized by keen interest to ideology, philosophy, religion, psychology, problems of creation that resulted in forming of various trends and schools, i.e. symbolism, formalism, futurism. At the same time journalism, socialistic and agitation literature were developed, the process of democratization of the language accelerated.

Period after October 1917 (namely 1920-1940s) is characterized by strengthening of state influence everywhere, including language; a definite linguistic policy is provided by the government. The so-called *likbez* (campaign against illiteracy), spreading of culture among masses resulted in expansion in the number of speakers of standard language. As a consequence of these processes standards of the

literary language shattered, it acquired multiple dialectic words, ones from industrial vocabulary, neologisms, abbreviations, rethinking of word meaning. In the 1930-1940s a normative tendency strengthened, the struggle for purity of the Russian language began, that found expression in a work on codification of standards of the literary language, creating of vocabularies.

All these peculiarities of historic and, as a sequence, of linguistic situation were taken into account by gathering texts for the corpus. First of all a list of the most prominent and significant authors of that period was formed. Several functional spheres were preliminarily defined: fiction, science and philosophy, journalism, criticism, memoirs. Besides that the corpus contains official documents, newspapers, private letters, advertisements and agitation texts.

One should consider that structure and ratio of functional spheres, text types and genres can vary for different periods. For example, in the XIX fiction was opposed to non-fiction as a whole, that's why the difference between scientific literature and journalism was weak. Novel was a very young genre; tales, essays or stories of that time are quite different things that they are today. But the adopted system and principles of text annotation enable us to solve these problems, its efficacy was proved by making a subcorpus of the XVIII century.

4.2. Pilot version of the corpus of the XVIII century

Widening of the RNC with texts of the XVIII century became a natural step in the work on making a diachronic corpus. Working on the subcorpus of the XVIII century began in 2006. Now it can be regarded as a pilot projects that was aimed first of all at proving of the adopted system of annotation, testing its flexibility and efficacy for describing of old texts.

In the XVIII century Russian literary standards were not stable yet, that's a period of transition from literary language based on the Church Slavonic language to literary language of a new type, based on proper Russian language system.

History of the Russian literary language of the XVIII century is not studied (at least from the linguistic point of view) well. Studying of a literary language is often substituted for studying of a language of literature, i.e. language of several significant writers. But peculiarities of evolution determine standards of genre system of this epoch: language of official documents, journalism, sermons, private letters, *etc*. The corpus, including texts of different types and genres, is designed to help future scientists in their researches of language of that period.

At present it is accepted to distinguish two or three periods in the history of Russian literary language of the XVIII century:

1) Times of Peter the Great (the end of the XVII — the beginning of the XVIII) – that's a period of confusion and merging – quite mechanical sometimes – of natural spoken language, Slavonicisms and Europeisms on the basis of state official language; of forming of new styles of *dialects of common citizens* and literary styles that hold a position in the middle, between high Slavonic style and everyday speech.

2) Period of Lomonosov (the 1740-1750s — the end of the XVIII) is a period of stylistic reglamentation and standardization of new literary language in terms of the theory of tree styles.

3) Period of Karamzin (the end of the XVIII — XIX) is characterized with reorganization of the literary language, which reflected in abolition of genre restricts,

in creating of new style of Russian language that became average literary standard, close to spoken language of educated citizens (Vinogradov, 1978).

The pilot corpus of texts of XVIII century contains prose of the second and third periods that represents wide range of text types in different spheres of language functioning. ***Fiction*** is represented by authors that exerted great influence on forming of the literary language: N.M. Karamzin, I.A. Krylov, N.I. Novikov, A.N. Radishchev, D.I. Fonvizin, M.D. Chulkov. ***Journalism*** is represented first of all by satiric articles by N.I. Novikov (published in magazines "Truten" - lit. "Drone", "Pustomelja" - "Twaddler", "Koshelek" - "Wallet", "Zhivopisec" - "Painter"), disputes between N.I. Novikov and Catherine the Great, social and political articles and essays by D.N. Fonvizin, A.N. Radishchev, a philosophic treatise by G. Skovoroda, a lampoon of M.M. Shcherbatov and memoirs of A.T. Bolotov. ***Educational and scientific sphere*** That is works by A.N. Radishchev on economy, law, history, politics, philological writings by D.N. Fonvizin and N.I. Novikov. ***Official sphere*** is represented by different memos, petitions, testaments, projects, edicts, army regulations. ***Everyday life sphere*** — private letters by N.M. Karamzin, A.N. Radishchev, D.N. Fonvizin, I.F. Bogdanovich, A.A. Boratynsky (a father of the famous poet), G. Skovoroda. ***Church and theological sphere*** that is works by a brilliant representative of Russian religious eloquence Platon (Levshin) and Feofan Prokopovich.

4.3. Composition of the Diachronic corpus

The diachronic corpus contains the following functional types: fiction of different kinds, criticism, journalism, including newspapers and magazines, scientific and philosophic works, documents, public, agitation texts, memoirs, diaries, and text made not for publishing: private diaries, personal letters. At the moment the content is still gathering and it is too early to talk about balanced corpus, nevertheless we can assume that the variety of genres and text types will be represented to the full extent in the final version of the corpus.

Sphere of functioning	**Tokens**	**Percentage**	**Planned proportion**
Fiction	25636036	58	40
Journalism	10418397	23	35
Educational and scientific	5704410	13	15
Official	208229	0,5	5
Church and theological	1141902	2,5	2
Everyday life	1228918	3	2,5
Advertisement	-	-	0,5
Total	44337892	100	100

Table 2: The distribution of texts according to spheres of functioning

According to date of creation the texts are distributed as follows:

XVIII	1700–1730 – 0,04 percent 1731–1779 – 2 percent 1780–1799 – 1,96 percent
XIX	1800–1830 – 5 percent 1831–1860 – 16 percent 1861–1899 – 31 percent
XX	1900–1920 – 14 percent 1921–1949 – 30 percent

4.4. The problem of orthographic variants

One of the most difficult problems that arose during preparation of texts for the diachronic corpus is preparation of texts written in old (pre-revolutionary) orthography. Many texts are scanned and recognized from old editions where old spelling is used. This problem is especially relevant on preparing texts of the XVIII century, as there were no strict rules that regulated spelling. That's why when these texts were prepared for new publication they were edited according to rules efficacious at the moment of the editing. Sometimes when a text becomes popular and is often published, this modernization of orthography goes very far, so that modern popular editions of N.M. Karamzin, I.A. Krylov, D.I. Fonvizin are in full agreement with modern spelling rules and standards. There are certainly special scientific editions that treat more cautiously with spelling of the original text, correcting only spellings that can be restored automatically (e.g., ъ after a hard consonant in the end of a word, *i* before vowels and *j*, *etc*.). If the text is reprinted from the manuscript for the first time, the academic edition tends to keep all individual orthographical peculiarities of the original text.

The analogue strategy is adopted by preparation of texts for the RNC: the electronic version should be close to the edition as far as possible. That's why if we take modern edition of texts of the XVIII century, orthography in it will be consistent with modern orthographical rules (accepted in 1956), if we reproduce a pre-revolutionary edition, then all characteristic features of its spelling are to be kept (except for script changes that were introduced during the reform in 1918).

Variety spellings of one word can be interesting for researchers, who study history and modern state of orthographic standards. At the moment an old variant of spelling can be found in the RNC only with the help of exact search. The problem is similar for texts of the XIX century. It can be solved by widening of the vocabulary by insertion in it of different spelling variants. That is supposed to be done; now a glossary of such variants is formed. These changes will allow to provide a morphological search of old spelling of words equally with new ones.

5. Corpus of spoken Russian: the Russian language tomorrow?

5.1. General Characteristics

Some specialist claim that to reflect a real ratio of texts a corpus should contain 95% of spoken texts, because it is just oral speech that forms a core of linguistic performance. In practice all modern corpora have a subcorpus of oral texts but it forms no more that 10% of the whole corpora. That can be explained by complexity and expensiveness of their gathering: recording, transcription of the records, and the whole process of text preparation and annotation.

But there are no doubts that oral speech is a material of great value, because it can reflect changes and tendencies that have just appeared in the language and that will then spread in all other communicative spheres. That is especially relevant for the Russian language because the main trend of development of the Russian literary language was always approaching of literary written and spoken language. This process is especially intensive during revolutions when spoken elements penetrate in all spheres of usage and can be adopted by the literary standards.

Researches on the spoken Russian are provided since the 1960s in many scientific centres in Moscow, St. Petersburg, Saratov, Perm, Yekaterinburg, Omsk, Krasnoyarsk, Ulyanovsk, *etc*. There are well-known works of E.A. Zemskaya,

O.A. Lapteva, M.V. Kitajgorodskaya, N.N. Rozanova, O.B. Sirotinina, V.E. Goldin, G.G. Infantova, *etc*. It is significant that the investigations in this area are usually based on quite restricted material, gathered by the researcher himself. Corpus of spoken texts enhances greatly possibilities of researchers.

Corpus of spoken texts – the Corpus of Natural Russian Speech – represents oral speech in its functional variety.

1) The Corpus contains whole original texts (not separated notes), that allows to discover that can escape during selective records by acoustic perception.

2) The Corpus contains large amount of texts that allows estimate frequency or randomness of a phenomenon, reveal regularities, make some statistical conclusions - to do all, that is impossible with such volume of information which researchers usually dispose of.

3) It contains texts that vary in respect of sociological, temporal, geographic parameters.

4) Text in the corpus covers rather a wide time span - about 50 years. First records are dated from 1956; the last ones are made in winter 2006. That allows to track changes that take place in spoken language (although these changes are very rapid), note new trends, *etc*.

5) The Corpus contains (and that is a significant distinction from usual collections of records of spoken texts) texts that concern different spheres of communication, various situations. We can't share an opinion that real spoken language is only spontaneous speech of townspeople in direct contact environment. Spoken text in the RNC can be both a dialogue in a shop, conversation during a dinner in the bosom of a family, and also a report, lecture, meeting of a writer with his audience, interview, talk-show, sport commentary and many other kinds of texts.

It seems that problem of distinction of boundaries of real oral speech and its separation from texts written-to-be-spoken concerns the problem of definition of spontaneous and prepared speech. Spoken text can be prepared to a greater or lesser extent; texts may be situated on the scale of spontaneity in the following way (descending) (Galyashina, 2002): 1) spontaneous speech 2) quasi-spontaneous speech (partly prepared) 3) beforehand speech prepared

Spontaneous speech	• Spontaneous dialogue • Spontaneous monologue
Quasi-spontaneous speech	• Interview (answers on questions) • Monologue on a given theme • Reproducing of somebody else's speech • Thought out speech on a prepared plan • Stereotype speech according to some pattern text • Repeating of a prompter's speech
Beforehand prepared speech	• Retelling of a written text • Summary of a written text • Reproduction of a text learned by heart • Reading aloud of a known text • Reading aloud of an unknown text

Table 3: The scale of spontaniety of oral speech

The corpus of spoken Russian doesn't contain beforehand prepared texts, but it contains the so-called quasi-spontaneous texts, that is first of all records of public speeches and content of a multimedia subcorpus.

6) Multimedia subcorpus is a unique part of the corpus of spoken texts within the RNC. It contains records of speech content of fiction films and cartoons (and documentary films and advertisement in the project). Earlier this sphere of functioning of a language escaped from attention of researchers of the spoken language and creators of large texts corpora.

Thus, the corpus of spoken language is really a representative collection of texts that reflects functioning of the contemporary spoken Russian language. Let's demonstrate how it influences on the structure and content of the corpus.

5.2. Composition of the corpus of spoken Russian

Total capacity of the corpus is about 4,4 million tokens. Texts are distributed according to spheres of spoken communication in the following way:

Sphere of functioning	**Tokens**	**Percentage**
Public speech	3542822	80
Private speech	308570	7
Speech of cinema	533993	12

Within each sphere texts are classified by the main text types.

Sphere of functioning	**Text types**	**Tokens**	**Percentage**
Public speech	talk	929498	26,20
	interview	349956	7,1
	discussion	1917469	54,1
	lecture	81880	2,3
	parliamentary hearings	87881	2,5
	conference	41602	1,1
	round table	49331	1,4
	narration	32353	1,0
	other	53998	4,3
Private speech	conversation	241643	78,3
	telephone conversation	25736	8,4
	tale	10326	4,7
	retelling	3702	1,2
	microdialogue	22929	7,4
Speech of cinema	drama	154294	29
	comedy	268109	50
	action	31613	6
	detective	10697	2
	films for children	23404	4,3
	other	45876	8,7

Table 4: The distribution of spoken texts according to text types

There are texts of various subject fields in the corpus. Texts that are marked as *private life* are most frequent (more than 50 percent), then there are texts on political themes, ones that concern public life, art and culture, science, leisure and entertainments, sport.

As for dating of texts, most of them are of 2003-2006, then follows texts dating from the 1990s (over 400 thousand tokens), then from the 1970s (260 thousand), from the 1980s (160 thousand), before 1970 - 160 thousand.

Date of recording of a text		**Place where a text was recored**
Before 1970 –	4 percent	Moscow and Moscow region
1971–1979 –	6 percent	Voronezh
1980–1989 –	4 percent	Novosibirsk
1990–1999 –	9 percent	St. Petersburg
2000–2006 –	77 percent	Samara
		Saratov
		Taganrog

Sources of the texts:

- transcriptions of spoken texts published by specialists in the 1970-1990s;
- collections of unpublished transcriptions prepared in different scientific centers: in the Institute of Russian Language, in MSU and SpbSU, in Universities of Saratov and Ulyanovsk;
- transcriptions of conversations of sociologists in focus-groups on various themes of public significance rendered by the Public Opinion Foundation;
- records of spoken texts made by the members of the Corpus team or under their direction.

6. How the corpus can be used for research of standards and variation

As it was shown above, the RNC being a representative collection of texts, reflects Russian usage in two dimensions: in horizontal – in all possible functional varieties – and in historic perspective.

So a linguist who is interested in language standards has two opportunities: 1) to set correlation of variants in modern usage and 2) to study an evolution that occurred during a certain period. These findings can be compared with recommendations of vocabularies in order to reveal its accordance to real usage.

The main body of the RNC contains texts of the last two centuries, that's why it is suitable for studying of short- and medium-term language changes. The total capacity of the corpus allows to study quite frequent phenomena. That's why that one can receive rather reliable results in studying of such issues like the following:

- lexical variations, including changes in content and semantic relations in rows of synonyms and thematic groups;
- morphological variants of nouns, verb and their evolution;
- variants of government, agreements and other syntactic constructions;
- productivity of word-formative models and means of word-forming, *etc*.

We will demonstrate how the RNC can be used for describing models of government of verbs of one synonymic row.

6.1. Problem definition

Verb government is one of the flexible points of the norm where changes can be observed during quite a short period of time and even more flexible and unsteady are stylistic estimation of variants and recommendations on its usage.

The goal of the research is to 1) prove on material of the RNC how these recommendations correlate the modern real usage; 2) to estimate potentialities of the RNC as a tool for diachronic studies.

Material of the investigation is the verbs *беспокоиться*, *тревожиться*, *волноваться*, *переживать*, that form one synonymic row with common meaning ‘worry, be anxious/uneasy (about smb, smth)’.

5.2. Descriptions in vocabularies and manuals

Semantic correlation of these verbs is described in The New Explanatory Dictionary of Russian Synonyms (NEDRS), founded on the database of texts of the second half of XX century; from the classics of XIX – the beginning of XX century there were taken only examples that agree with the modern standards.

Verb	**Semantic similarity**	**Semantic differs**	**Example**
Беспокоиться Bespokoitsa	‘to have an unpleasant feeling that exists when a person doesn’t know anything significant about the situation that concerns him and when he is afraid that the situation has changed or can change for the worse’	Intellectual estimation of the situation prevails. External manifestation is motion activity.	Он беспокоится, если кого-нибудь нет дома On bespokoitsa, esli kogo-nibud’ net doma (He worries when somebody is not at home)
Тревожиться Trevozhitsa		It describes rather mental reaction or reaction of nervous system. External manifestation is mimicry or unregulated fine motility.	Врачи за нее тревожились. Vrachi trevozhilis za nee (Doctors worried about her)
Волноваться Volnovatsa		It is closer to trevozhitsa, but indicates the general anxiety of the subject. Manifestation is not only exaggerated motion activity but abnormal behavior	Я плохо спала — волновалась за Саню. Ja ploho spala, volnovalas’ za Sanyu. (I had a bad night, I worried about Sanya)

Table 5: The definitions of the verbs in the New Explanatory Dictionary of Russian Synonyms

Bespokoitsa is a dominant of the synonymic row. All members are stylistically neutral and differ only semantically. It should be mentioned that the vocabulary does not include in the row a verb *prezhivat’*, although such meaning of the verb is registered in explanatory dictionaries:

ПЕРЕЖИВАТЬ, несов. 1. См. пережить. 2. *за кого-что*. Волноваться, беспокоиться о ком-чем-н. (*разг*). П. за сына. П. за любимую команду. 3. Мучиться, страдать по какому-н. поводу (*разг*.). Поссорился с женой, теперь переживает.

Perezhivat’, imperf. 1. See perezhit’(= ‘survive’). 2. *za kogo-chto*. To worry about smb-smth, be distressed for smb-smth (*coll*.). To be distressed for the son; for a favourite command. 3. To feel sore about smth (*coll*.). He has quarrelled with his wife, now he feels sore about it (Ozhegov, 1999).

All verbs under study can have a dependent word. That is how it is described in the NERDS and in the vocabulary “Government in the Russian language” («Upravlenie v russkom jazyke» - URJ) by D.E. Rozental’ (Rozental’, 2005)

The most complete description is offered in the NERDS; at that the sequence of models corresponds to the degree of preference of a variant with relation to the standard.

Dictio-nary	Verb	Without objective comple-ment	**За** + Accusativ smb/smth	**о/об/обо** + Prepositional smb/smth	**Из-за, насчет, по поводу** + Genitive smb/smth	Subordi nate clause
NEDRS	беспокоиться	+	smb	smb	smth	+
	тревожиться	+	smb	smb	smth	+
	волноваться	+	smb	--	smth	+
URJ	беспокоиться	no data	smb/smth colloquial	smb/smth	no data	no data
	тревожиться	no data	smb/smth	smb/smth colloquial	no data	no data
	волноваться	no data	smb/smth	smb/smth	no data	no data
	переживать	no data	smb/smth substandard	smb/ssmth colloquial	smb/smth	no data

Table 6: The description of verb government in the NEDRS and URJ

As it is obvious from the table, recommendations of the two dictionaries differ. The NERDS restricts the semantic of a governed noun (e.g. **беспокоиться за** + Acc smb: object must be a living being; **беспокоиться из-за** + Gen smth: a cause of anxiety – circumstances, events). In the URJ comparing to the NEDRS stylistic restrictions are more strong, but semantic demands are more free. How these instructions correlate with real data received with the help of the RNC?

6.3. Data from the RNC

Absolute frequency of concerned words and joint frequency of verbs and prepositions in verb-noun collocations in the entire corpus are shown in the table below (according to http://corpus.leeds.ac.uk/ruscorpora.ru)

lemma preposition	**Беспокоиться Total=3167**	**Волноваться Total=4850**	**Тревожиться Total=523**	**Переживать Total=9302**
за	324	325	88	606
о/об/обо	707	126	93	252
из-за **насчет** **от**	9 60 -	- - 102	7 - 13	128 - 215

Table 7: Frequency distribution of the verbs in collocations

Correlation of absolute frequency of concerned verbs shows that a verb *bespokoitsa* is real generally used and can be considered as dominant. It should be taken into account that high frequency of verbs *volnovatsa* and *perezhivat* comes out from the fact that these polysemantic verbs are used in different meanings, not only in that one we are interested in (e.g., *perezhit' vojnu* – to survive a war, *etc*.).

From the other hand a high frequency of these verbs and prepositions suggests that there are no reasons to introduce stylistic restrictions on any of these collocations.

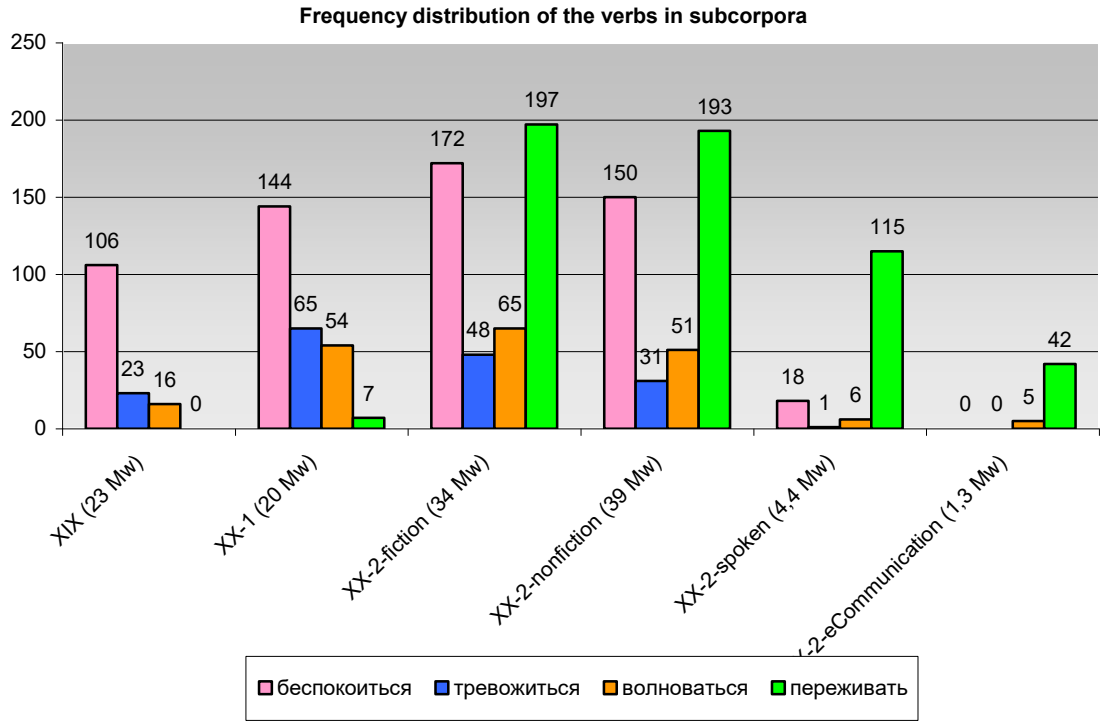
More detailed analysis of verb government is performed on the basis of manual data treatment in several subcorpora:

I. 1800-1899 XIX-fiction
II. 1800-1899 XIX-non-fiction
III. 1900-1949 XX-1-fiction
IV. 1900-1949 XX-1-non-fiction
V. 1950-2005 XX-2-fiction
VI. 1950-2005 XX-2- spoken
VII. 2000-2005 XX-2-eCommunication.

Verb	**Preposition**	**Objective comp.**	**Subcorpora**						
			19		**20-1**		**20-2**		
			I	**II**	**III**	**IV**	**V**	**VI**	**VII**
Беспокоиться	**о/об/обо**	smb	14	8	26	9	44	8	0
		smth	24	33	46	20	46	9	0
	за	smb	5	9	15	2	44	0	0
		smth	4	3	9	3	17	0	0
	из-за **насчет** **по поводу**	smb/ smth		3 1	2 6	5 1	2 4/15	1	
Всего			47	59	104	40	172	18	0
Тревожиться	**за**	smb	3	2	8	5	19	0	0
		smth	1	1	5	5	7	0	0
	о	smb	1		2	2	11	0	0
		smth	1		3	4	8	0	0
	из-за **насчет** **по поводу** **от**	smb/ smth		1	1 1	1 1	2/1	1	
Всего			7	16	20	45	48	1	0
Волнова-ться	**за**	smb	1	1	13	1	37	3	3
		smth	1		6	8	7		
	о	smb	1		2		1	1	
		smth	3	1	4	1	5	1	2
	из-за **насчет** **по поводу** **оттого** **от**	smb/ smth		3 1 1 3	2/3 6	4 1 3	2/12 1	1	
Всего			6	10	36	18	65	6	5
Пережи-вать, intrans.	**из-за**	smb	0	0	0	0	14	0	0
		smth		0	0	0	34	4	7
	о	smb	0	0	0	0	3	1	0
		smth	0	0	0	0	14	3	2
	за	smb	0	0	0	0	77	10	4
		smth	0	0	0	0	17	16	5
	по поводу	smth			0	0	14	5	1
	без доп.				8	1	24	76	23
Всего					8	1	197	115	42

Table 8: Variants of prepositional government of the verbs in the subcorpora of the RNC

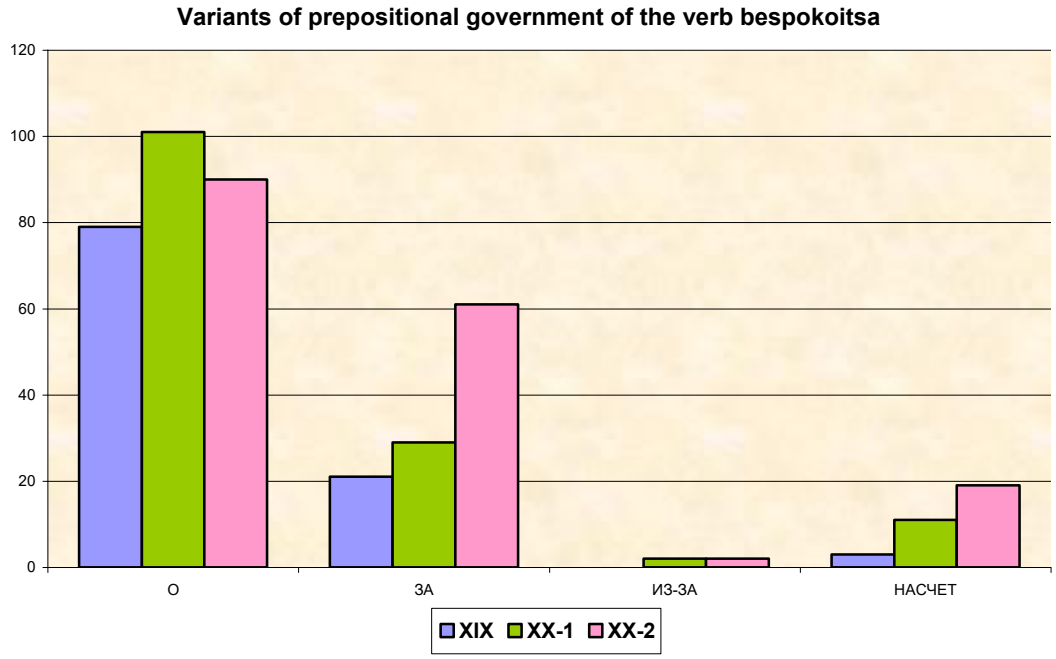
6.4. Data analyses and conclusions



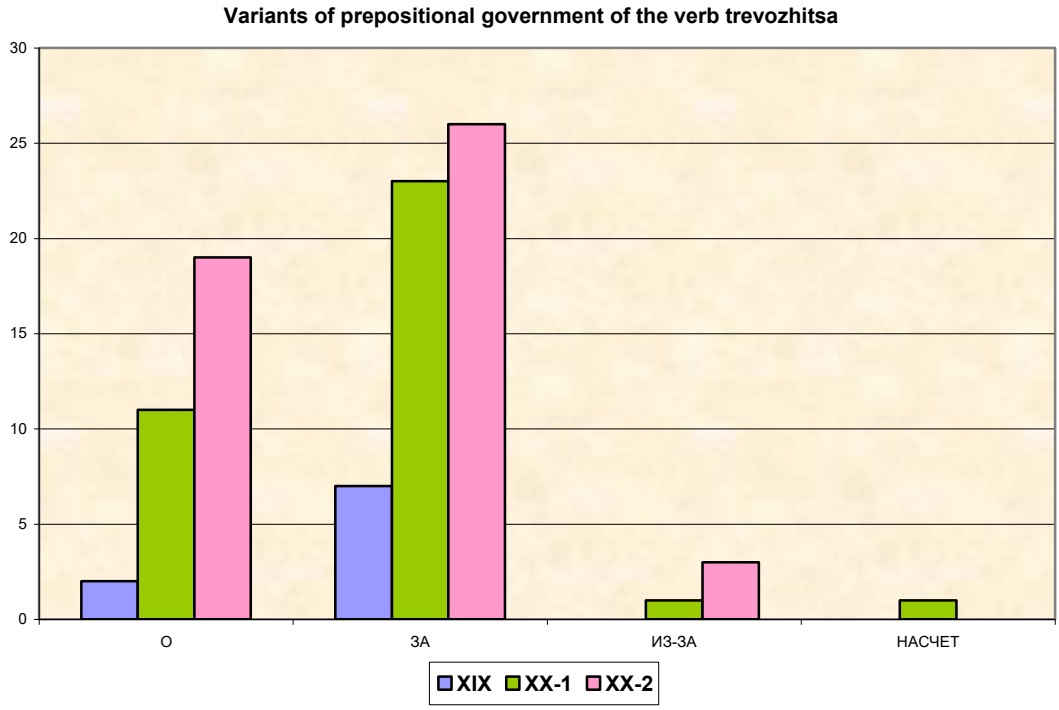
The correlation between verbs of the concerned synonymic row, analyzed by manual data treatment is represented on diagram. It shows, that the verb *bespokoitsa* (pink column) can be really considered as the dominant, as it keeps its high frequency for any period of time. The verb *trevozhitsa* (blue column) is on the contrary infrequent and its frequency has decreased since the XIX century. In the contemporary language it is regarded as bookish, literary. As you may see it is not fixed in the corpus of e-communication, and only once in the spoken subcorpus.

As it is shown on diagram, this synonymic row consists of 3 verbs in the XIX century and of 4 verbs in the XX century. The added verb - *perezhivat* - was consolidated in the meaning 'to be distressed for *smb*, to be emotional over *smb-smth'* in the second half of the XX century. It had changed its grammatical characteristics and began to be used as intransitive. Sporadic cases of such usage of the verb are registered in the RNC in the first half of the XX century (in the novel "Golden Calf" by Ilf and Petrov, 1927). For the texts of the 1960-1980s it is a frequent usage noticed and commented on by some linguists (e.g. Dmitry Shmelev). In the contemporary spoken language (oral texts and electronic communication) the frequency of the verb *perezhivat* exceeds extremely the frequency of the other verbs of the row. In the whole, the verb *perezhivat* retains colloquial shade of meaning, but spheres of its usage are expanding.

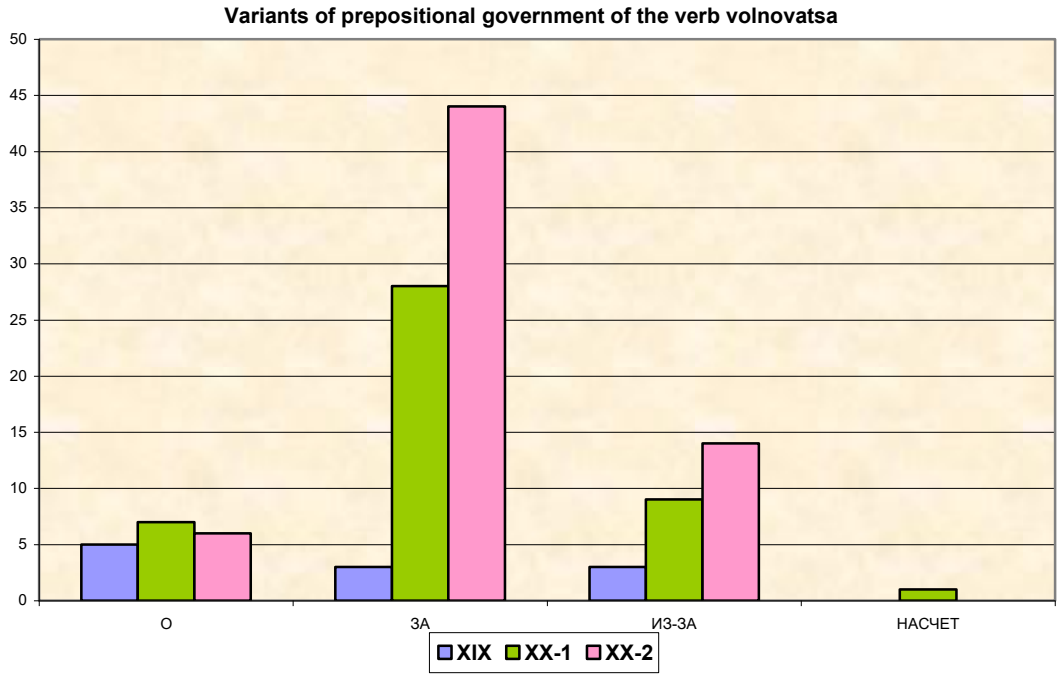
The analysis of different variants of verbal government shows that all the verbs under study tend to equalize the models of government. It is displayed on the diagrams: blue column is number of variants in the texts of the XIX century, green column – in the texts of the first half of XX century, pink one – in the texts of the second half of the XX century. The first three columns displays the usage of variants with preposition **O,** the second triplet is the usage of variants with **ЗА,** the rest are given for comparison.



As for the verb ***bespokoitsa***, there is an increase of the variant with preposition **ЗА+Acc**. For the texts of the XX century this variant is practically equivalent to the variant **О+Prep**. We also need to mention the increasing of frequency of the construction **НАСЧЕТ+Gen**.

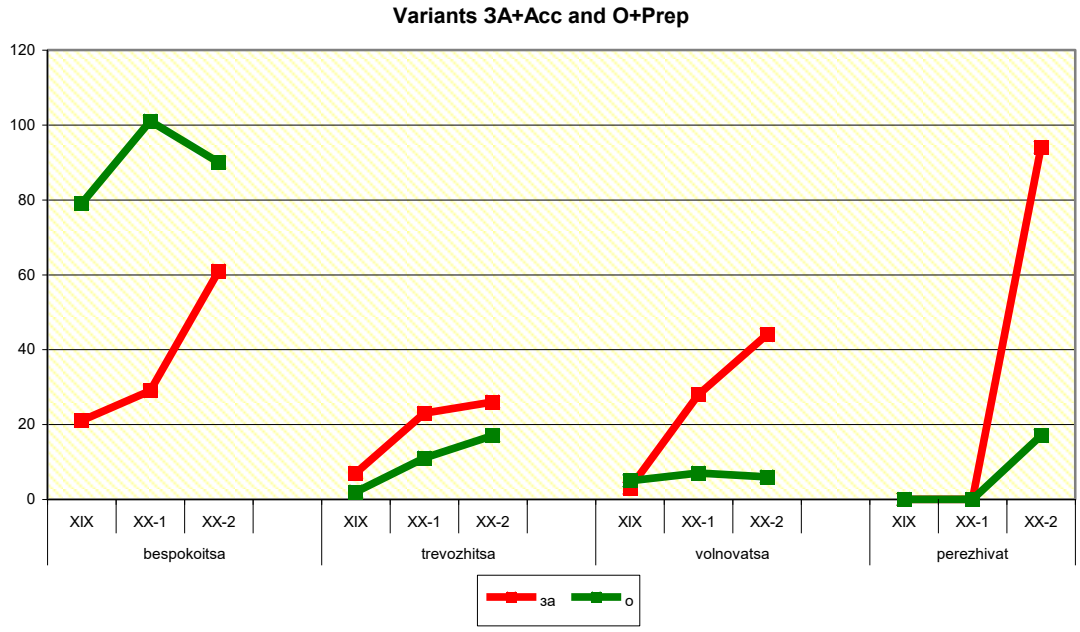


As far as the verb ***trevozhitsa*** is concerned, there is a tendency to reduce a gap between variants *trevozhitsa za kogo-to/trevozhitsa o kom-to*.



For the verb ***volnovatsa*** a variant *volnovatsa za kogo-to* is more preferable than *volnovatsa o kom-to* and it became so even more in the XX century. Thus we can't totally ignore this model (as the Dictionary of Synonyms does), as it is stably represented in the corpus during all the concerned period, including contemporary speech and electronic communication. It should be mentioned that the construction *volnovatsa iz-za kogo-to* or even more usable *iz-za chego-to* (a cause) is also rather frequent.

As for ***perezhivat*** the variant *za kogo-to* is preferable, *iz-za kogo-to* and *o kom-to* are infrequent.



On the whole we can conclude that the variant of government **V+3A+Acc** has turned out to be more productive than the other: during the period of the XIX-XX centuries the variants **V+3A+Acc** (red graph) and **V+O+Prep** (green graph) of the verbs having a main government model **V+O+Prep** (*bespokoitsa, trevozhitsa*) have become closer in a quantitative sense whereas variants of government of the verbs with a dominant model **V+3A+Acc** (eg., *volnovatsa*) diverge.

Comparing our results with recommendations of vocabularies (for example, the well-known dictionary "Government in Russian" by Rozental, we can come to a conclusion that these rules look too strict, and don't meet the real usage. For example, a construction *perezhivat za kogo/chto* is attributed with a mark "substandard", i.e. it can't be used in the literary language. But that is not so, as the Corpus shows: the model is widely used in fiction (in the works of B. Vasilyev, V. Tendryakov, V. Grossman, V. Chivilikhin, I. Grekova, V. Shukshin, A. Rybakov, *etc*.), both in speech of personages and in narrator's text. Besides, what is more surprising is that this model is widely spread in non-fiction (memoirs, journalism). As for the variant ***trevozhitsa*** *O kom/chem*, marked in the dictionary as *colloquial*, the Corpus gives only one example for the verb *trevozhitsa* in the spoken texts one none in e-Communication.

Thus, the corpus approach to studying variants, as it was exemplified by the research of verbal government, enables to carry out the qualitative and quantitative analysis of units and constructions, to reveal trends in correspondence of competing variants, to trace the development of new phenomena, to amend lexicological descriptions and normative recommendations.

References

Apresjan Ju. (ed) (1999) Novyj objasnitel'nyj slovar' sinonimov russkogo jazyka (=The New Explanatory Dictionary of Russian Synonyms). Moscow: Jazyki russkoj kul'tury. - NEDRS

Butler, C.S. (2004). Corpus studies and functional linguistic theories. *Functions of language 11:2*, 147-186.

EAGLES (1996) (author J.M. Sinclair). Preliminary recommendations on text typology. Available on-line from www.ilc.cnr.it/EAGLES96/pub/eagles/ corpora/corpustyp.ps.gz

Galjashina, E.I. (2002) Problema differenciacii spontannoj i podgotovlennoj rechi, in Trudy mezhdunarodnogo seminara Dialog'2002 po komp'juternoj lingvistike i ee prilozhenijam. Available on-line from http://www.dialog-21.ru/materials/archive.asp?id=7287&y=2002&vol=6077

Graudina, L. K. (1980) Voprosy normalizacii russkogo jazyka: grammatika i varianty. Moscow: Nauka.

Nacional'nyj korpus russkogo jazyka: 2003-2005. Rezul'taty i perspektivy. (2005) Moscow: Indrik.

Ozhegov, S.I., N.Ju. Shvedova (1999) Tolkovyj slovar' russkogo jazyka. Moscow: Azbukovnik.

Reference Guide for the British National Corpus (XML Edition) (2007) Lou Burnard (ed.) Available on-line from http://www.natcorp.ox.ac.uk/XMLedition/URG/

Rozental', D.E. (2005) Upravlenie v russkom jazyke, in D.E. Rozental'. Russkij jazyk: Spravocnik-praktikum. Moscow: Oniks. - URJ

Sinclair, J.M. (2003) Corpora for lexicography, in P. van Sterkenberg (ed.). A Practical Guide to Lexicography, pp. 167–178. Amsterdam: Benjamins

Sharoff, S. (2004) Towards basic categories for describing properties of texts in a corpus, in Proceedings of Language Resources and Evaluation Conference (LREC04*)*. May, 2004, Lisbon, Portugal. Available on-line from http://www.comp.leeds.ac.uk/ ssharoff/texts/lrec-04.pdf

Sharov, S.A., S.O. Savchuk. (2004) Tipologija tekstov dlja predstavitel'nogo korpusa, in Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika-2004», pp. 352-362. SPb: St. Petersburg University Press.

Shmelev, D.N. (2002) On some problems of development and normalization of the modern Russian language, 1962 in D.N. Shmelev. Opera selecta, Moscow: Yazyki slavyanskoj kultury.

Vinogradov, V. V. (1978) Osnovnye etapy istorii russkogo jazyka, in V.V. Vinogradov. Izbrannye trudy. Istorija russkogo literaturnogo jazyka, pp. 10-64. Moscow: Nauka.

Опубликовано: Corpus-based Investigation of Language Change: the Case of RNC // Proceedings of the Corpus Linguistics Conference CL2007 University of Birmingham, UK 27-30 July 2007 / Edited by Matthew Davies, Paul Rayson, Susan Hunston, Pernilla Danielsson
http://ucrel.lancs.ac.uk/publications/CL2007/final/181/181_Paper.pdf