Corpora of the Russian Language*

Victor Zakharov

Saint-Petersburg State University, Saint-Petersburg, Russia
vz1311@yandex.ru

Abstract. The paper describes corpora of the Russian language and the state of the art of Russian corpus linguistics. The main attention is paid to the Russian National Corpus and to specialized corpora.

Keywords: Russian corpus linguistics, corpora, the Russian language, specialized corpora.

1 Introduction

1.1 Prehistory of Russian Corpus Linguistics

In recent years creation of different text corpora became one of the cutting edge directions in the applied linguistics. In Western countries the corpus linguistics shaped itself as a separate linguistic universe in early 90's, even though the concept of the corpus and the first physical corpora had been known long before.

The earliest Russian corpus was built in 1980s at the University of Uppsala (Sweden). But another project influenced this direction in Russia so much that it should be mentioned. In 1960–70s the **Frequency Dictionary of Russian** was created by L.N. Zasorina (a printed version, 1977 [1]). Text database for the dictionary counted about 1 million tokens. During its compilation a huge number of notorious issues of corpus linguistics was discussed: representativeness, tokenization, normalization, lemmatization. So it was the earliest computerized corpus of Russian that doesn't exist nowadays.

In 1980s the **Computer Fund of the Russian Language** project started. The idea belonged to the academician Andrei Yershov. It was formulated in his paper "On methodology of constructing dialogue systems: the phenomenon of business prose" [2]. The idea was stated as follows: "Any progress in the field of constructing models and algorithms will remain a purely academic exercise, unless a most important problem of creating a Computer fund of the Russian language is solved. It is to be hoped that creation of such a Computer fund by linguists, qualified for the task, will precede construction of large systems for application purposes. This would minimize labor costs and simultaneously would protect the 'tissues' of the Russian language from arbitrary and incompetent intervention". The Fund was to include the following databases: 1) general lexicon of Russian, 2) databases of various dictionaries, 3) terminology database, 4) information system for the Russian grammar, 5) other subsystems (phonetics, dialectology, diachronic lexis etc), 6) and last but not least, collection of texts, i.e. corpus. Unfortunately, the bulk of the accumulated results was either abandoned or lost.

* Invited talk.

I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, pp. 1–13, 2013.

1.2 History

The most renowned Russian corpus for many years was the **Uppsala Corpus of Russian Texts**. By now its linguistic material is neither up to date in terms of the volume (one million word occurrences), nor complies with modern conceptions of a national corpus at all. The Uppsala corpus has 600 texts, its volume is 1 million tokens, equally divided between specialized texts and fiction. The aim of the corpus was to represent literary language, and thus the collection doesn't cover spoken language. Full specialized texts from 1985 till 1989 were selected for the corpus and fiction from 1960 through 1988. Texts were presented in Latin alphabet.

The Uppsala corpus belongs to so called **Tübingen Russian corpora** that were created during the project "Linguistische Datenstrukturen. Theoretische und empirische Grundlagen der Grammatikforschung" (SFB 441) of the Tübingen University in 1990-2000s with online access [3].

Russian newspaper corpus was built at the Department of Philology of the Moscow State University in 2000-2002 at the Laboratory of General and Computational Lexicology and Lexicography (1 million tokens in total, online version is limited to 200 thousand tokens). Texts and text items are automatically or semi-automatically marked by various tags: the source, text volume, genre, date of the publication etc. (for texts); grammatical, lexical, morphemic or other categories (for words) [3].

2 Modern Corpora of Russian

The National Corpus of the Russian Language, hereinafter referred to as the Russian National Corpus, is the most popular one among linguists for both being the most well known and the opportunities which it presents. However, being unable to go into a deeper analysis within the framework of this paper, we will zero on its general characteristics together with its most unique features. Also, to show the state of the art in modern Russian corpus linguistics we will touch in greater detail upon other corpora that are not so much known but are worth mentioning.

2.1 The Russian National Corpus

The Russian National Corpus (RNC)[1] includes primarily original prose representing standard Russian but also, albeit in smaller volumes, translated works (parallel with the original texts) and poetry, as well as texts, representing the non-standard forms of modern Russian [4,5]. It was started in 2003 and from April, 2004 is accessible via Internet. The corpus size in total is about 500 million tokens (March 2013).

The corpus allows us to study the variability and volatility of linguistic phenomena frequencies, as well as to obtain reliable results in the following areas: 1) the study of morphological variants of words and their evolution; 2) the study of word-formation options and related issues; 3) the study of changes in syntactic relations; 4) the research of changes in the system of Russian accent; 5) a study of lexical variation, in particular, changes in synonym series and lexical groups, as well as semantic relations in them.

[1] http://ruscorpora.ru

Within the main corpus the RNC includes the following subcorpora:

1) The Main Corpus. It includes texts representing standard Russian and may be subdivided into 2 parts: modern written texts (from the 1950s to the present day) and early texts (from the middle of the 18th to the middle of the 20th centuries; pre-1918 texts are given in modern orthography). The main corpus counts in total 230 million tokens. The search is carried out in both groups. It is possible to choose one of them and add search parameters on the *Customize your corpus* page.

The part of modern texts is the largest one of the subcorpora. Texts are represented in proportion to their share in real-life usage. For example, the share of fiction does not exceed 40%.

Every text included in the main corpus is subject to metatagging and morphological tagging. Morphological tagging is carried out automatically. In a small part of the main corpus (around 6 mln tokens) grammatical homonyms are disambiguated by hand, and results of automated morphological analysis are corrected. This part is the model morphological corpus and serves as a testing ground for various search algorithms and programs of morphological analysis and automated processing. Disambiguated texts are automatically supplied with indicators of stress. Stress annotation may be turned off for printing or saving the search results.

2) The Corpus of Spoken Russian. It represents real-life Russian speech and includes the recordings of public and spontaneous spoken Russian and the transcripts of the Russian movies. To record the spoken specimens the standard spelling was used. The corpus contains the patterns of different genres/types and of different geographic origins. The corpus covers the time frame from 1930 to 2007.

3) Deeply Annotated Corpus (treebank). This corpus contains texts augmented with morphosyntactic annotation. Besides the morphological information, every sentence has its syntax structure (disambiguated). The corpus uses dependency trees as its annotation formalism (Fig. 1).

Fig. 1. Dependency tree of a Russian sentence

Nodes in such a tree are words of the sentence, while its edges are labeled with names of syntax relationships. The repertory of relationships for the corpus, as well as other specific linguistic decisions has been developed in the Laboratory for Computational Linguistics, Institute for Information Transmission, Russian Academy of Sciences that compiled the corpus.

4) Parallel Text Corpus. The site contains the parallel text corpora for the following languages: English, German, French, Spanish, Italian, Polish, Ukrainian, Belorussian, and multilingual.

5) Dialectal Corpus. The dialectal corpus contains recordings of dialectal speech (presented in loosely standardized orthography) from different regions of Russia. The corpus employs special tags for specifically dialectal morphological features (including those absent in standard language); moreover, purely dialectal lexemes are supplied with commentary.

6) Poetry Corpus. At the moment the poetry corpus covers the time frame between 1750 and 1890s, but also includes some poets of the 20th century. Apart from the usual morphological tagging, there is a number of tags adapted for poetry.

7) The poetic part of the **Accentological Subcorpus** includes Russian poetry of 18–21 centuries with the marked arses (potentially stressed syllables). It gives a user the possibility to figure out the real stress of a word-form according to the simple set of rules. For example, it is possible to search for texts written in various poetic meters.

8) Educational Corpus. The educational corpus is a small disambiguated corpus adapted for the Russian school educational program.

9) Newspaper Corpus. It covers articles from the media of the 2000s.

10) Multimodal/Multimedia Corpus (see below).

2.2 Semantic Annotation in the Russian National Corpus

The main corpus of the RNC contains semantic annotation, too [6]. Semantic annotation in the main corpus is a unique feature of RNC that makes it distinct from other national corpora.

Semantic and derivational parameters involved are *person*, *substance*, *space*, *movement*, *diminutive*, etc. There are three groups of tags assigned to words to reflect lexical and semantic information: class, lexical and semantic features, derivational features. The set of semantic and lexical parameters is different for different parts of speech. Moreover, nouns are divided into three subclasses (concrete nouns, abstract nouns, and proper names), each with its own hierarchy of tags.

Lexical and semantic tags are grouped as follows: taxonomy, mereology, topology, causation, auxiliary status, evaluation. A word in the semantic dictionary is assigned a set of characteristics along many other parameters.

The meta-language of tags is based on English notation; it is, however, possible to make a search using traditional Russian category names in the search "semantic features" form. The following are some tags from an inventory of available tags with examples in parenthesis.

Nouns: categories: r:concr – concrete nouns, r:abstr – abstract nouns, r:propn – proper names. Some tags for concrete nouns:

Taxonomy: t:hum – person (*человек* (human), *учитель* (teacher)), t:hum:etn – ethnonyms (*эфиоп* (Ethiopian), *итальянка* (Italian)), t:hum:kin – kinship terms (*брат* (brother), *бабушка* (grandmother)), t:animal – animals (*корова* (cow), *сорока* (magpie)), etc.

Mereology: pt:part – parts (*верхушка* (top)), pt:part& pc:plant – parts of plants (*ветка* (limb), *корень* (root)), pt:part& pc:constr – parts of buildings and constructions (*комната* (room), *дверь* (door)), etc.

Topology: top:contain – containers (*комната* (room), *озеро* (lake)), top:horiz – horizontal surfaces (*пол* (floor), *площадка* (ground, area)), etc.

Evaluation: ev – evaluation (neither positive nor negative) (*озорник* (mischief-maker)), ev:posit – positive evaluation (*умница* (clever man or woman)), ev:neg – negative evaluation (*негодяй* (scoundrel)).

Some tags for verbs:

t:move – movement (*бежать* (run), *бросить* (throw))

t:put – placement (*положить* (put), *спрятать* (hide))

t:impact – physical impact (*бить* (beat), *колоть* (prick))

t:be:exist – existence (*жить* (live), *происходить* (happen))

t:be:appear – start of existence (*возникнуть* (arise), *создать* (create))

t:be:disapp – end of existence (*убить* (kill), *улетучиться* (disappear))

t:loc – location (*лежать* (lie), *стоять* (stand)).

And these are just a few examples of 200 tags available in the corpus that are structured in a hierarchical way.

2.3 Search in the Russian National Corpus

For search RNC uses the search engine Yandex-Server which has been specially adapted for corpus needs. One can search by an exact form, by a set phrase, by lexico-grammatical and semantic features, by additional features such as a specified position (before or after punctuation marks, in the beginning or in the end of a sentence, capitalization, etc).

Words we are searching for could be combined with logical operators «AND», «OR» and «NOT». For compound searches parenthesis are used. For example, the query `S & (nom|acc)` yields nouns in nominative or accusative. It can be used with both left or right truncation. Distance between words could be set from minimum to maximum. The distance between words next to each other is 1 word; the distance of 0 is interpreted as concurrence of word-forms. For lexico-grammatical search, we can input a sequence of lexemes and/or word-forms with certain grammatical and/or semantic features. We can combine them in any way.

A simpler way to search for certain grammatical features is to use a selection window. The selection window contains a list of appropriate features, subdivided by categories: i.e., part of speech, case, gender, voice, number for morphology, etc. To invert selection within a category, one uses the equivalent of the "NO" operator.

The **semantic features** field allows for listing the semantic and derivational features of the lexeme. As a rule, semantic features have a hierarchy. In semantic search we must remember that words tagged as belonging to a category may often not belong to a subcategory: for example, verbs belonging to the "physical impact" category include verbs not belonging to the subcategories "creation" and "destruction", such as verbs of processing like *вымыть* (wash). There exists a capability to uncheck the boxes next to all subcategories.

By default all the tagged meanings of a given word are searchable. For instance, the parameter Human qualities selected in the Semantic features field will yield both *умный* 'intelligent', *верный* 'faithful', *коварный* 'perfidious' (where the parameter is present in its basic meaning), as well as *мягкий* 'soft' or *холодный* 'cold' that apply to human beings only metaphorically.

To refine the scope of the search, we could select one or two parameters:

"sem" — only the first meaning given in dictionaries is searched (thus human qualities will yield words like 'intelligent', 'faithful' or 'perfidious', but not those like 'soft' or 'cold'); "sem2" — the meanings other than the first ones are searched (thus only words like 'soft' or 'cold' will be found).

The search can be limited to a subcorpus which is chosen as one of the above mentioned subcorpora or as a combination of metadata features.

The types of annotation which are specific for the special corpora (MURCO, poetical, dialectological, etc.) define the peculiarities of the appropriate interface in comparison with the interface of the RNC proper.

Search results can be presented twofold: a horizontal text (a broader context) and a concordance (Fig 2). In both cases grammatical and semantic features of any word can be checked out (Fig 2 shows that for the word *женщина* (women)).

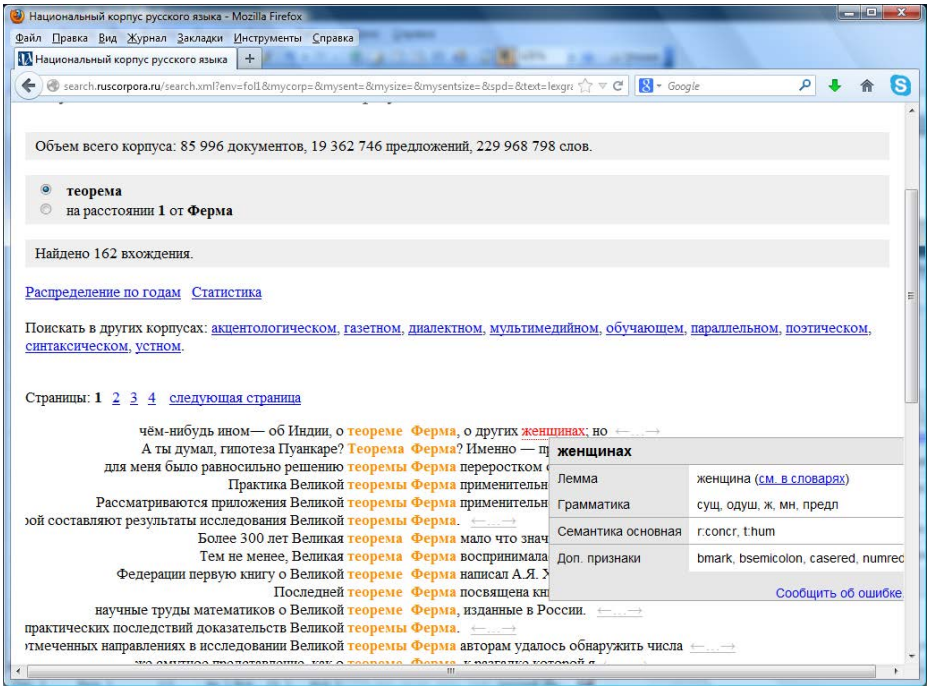

Fig. 2. Search results for the word combination *теорема Ферма* (Fermat's theorem)

From the search page one can get to *Графики* service (Charts) (*Распределение по годам* link (chronological distribution)).

2.4 Charts

In terms of functionality Charts service of RNC is similar to Google Books Ngram Viewer. It shows chronological distribution of lexical units (text forms, phrases), found

in the main corpus of RNC. You can get to this link from the search results page of RNC, Fig 2, as well as from the main menu, Fig 3.

You can set time limits too, e.g. from 1930 through 1960. Clicking the button *Построить* (Draw), we will get a chart, Fig 4, where each object we compare is shown in its own color with legend located in the top right corner (here *Черчилль* (Churchill), *Рузвельт* (Roosevelt), and *Франко* (Franco) are shown in comparison).

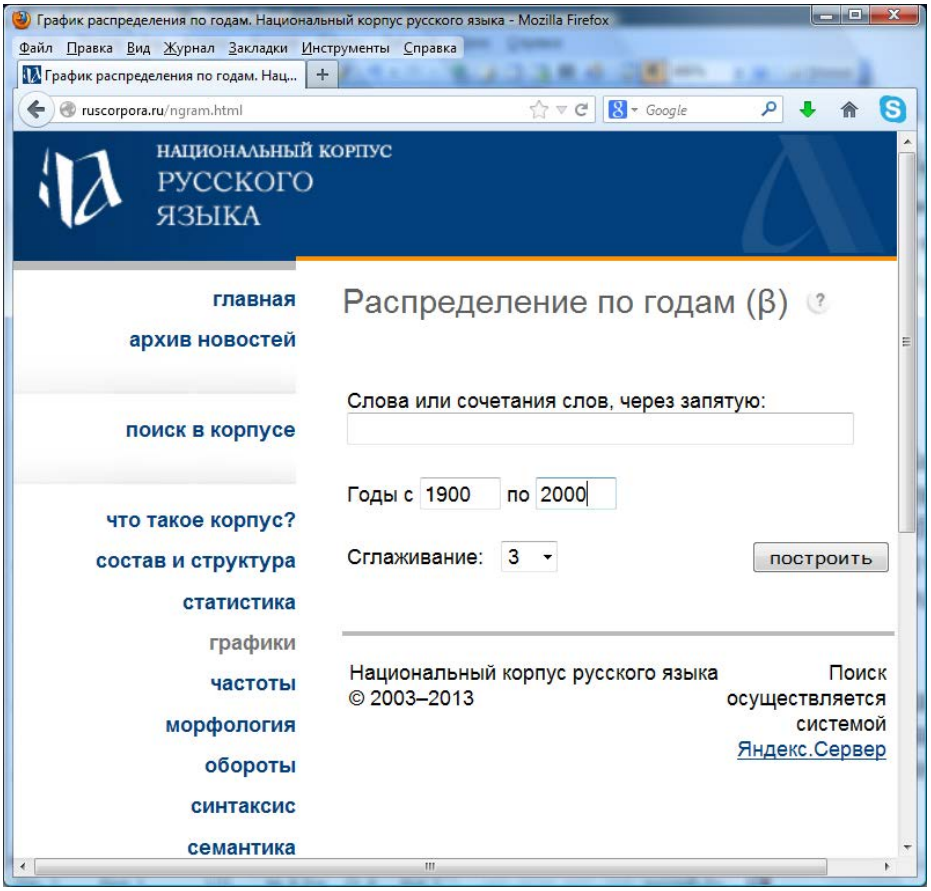

Fig. 3. Charts interface

The vertical axis is for relative usage frequency of a lexical unit. Mouse over any point on the curve, and you would see the relative usage frequency (in ipm), for a respective year. Smoothing the charts allows the general trend to be seen beyond random frequency volatility. Thus, with the 10 years smoothing the word frequency is averaged over 5 prior and 5 consecutive years, i.e. the average for 11 years is taken for any given year.

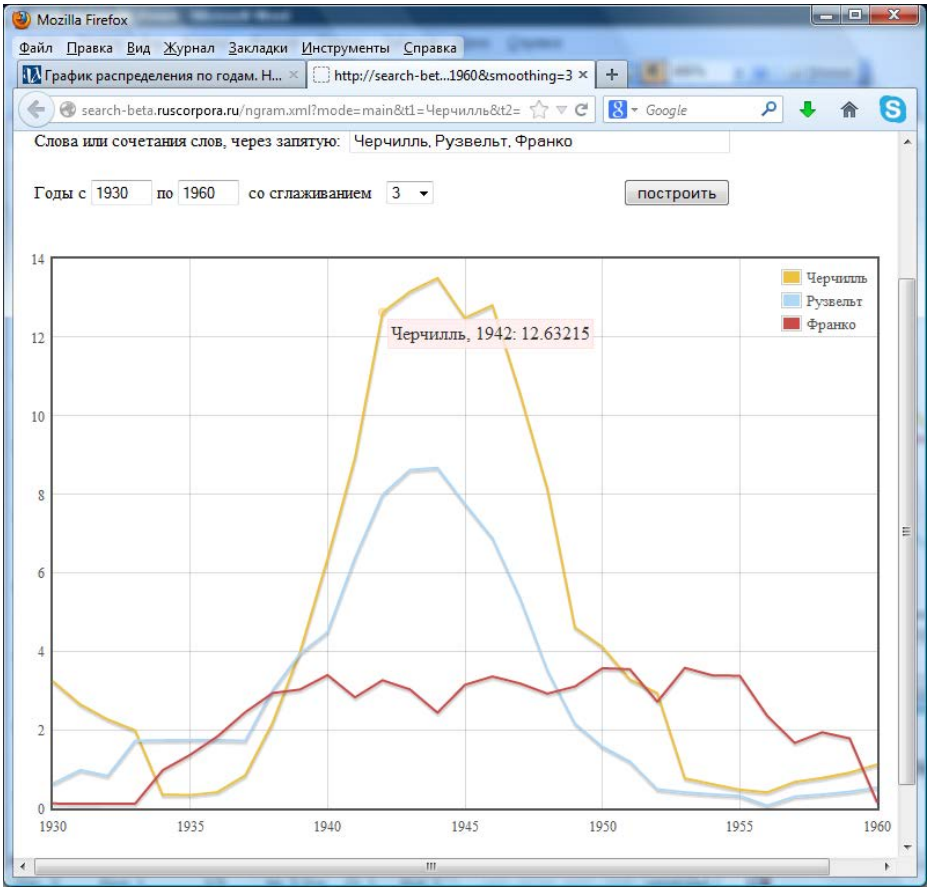

Fig. 4. Occurrence frequency of names Churchill, Roosevelt and Franco in the texts published from 1930 through 1960

There is a possibility to display tables with absolute frequency of occurrence for any year. Links from the tables allow to show examples from the corpus for each year.

From Charts one can easily jump to Google Ngram Viewer working on a Russian language collection of texts in Google Books. National Corpus and Google Ngram Viewer while having similar ideology, use different formulae to calculate relative frequencies.

2.5 Other Text Corpora of Russian

Helsinki Annotated Corpus (HANCO). The HANCO Corpus project has been running since 2001 in the Department of Slavonic and Baltic Languages and Literatures at the University of Helsinki. The corpus was envisaged to include morphological, syntactic and functional (semantic) information about approximately 100000 running words, extracted from a modern Russian magazines and representing the contemporary Russian language.

The main principles of creation are as follows [7]:

1. Targeting a wider audience. Potential users are not only a narrow circle of experts, but also students and teachers of Russian.
2. Focus on the accuracy of the grammatical description.
3. Multilevel grammatical information combined in the process of searching.
4. Possibility of alternative interpretations of linguistic facts.

For now the following types of linguistic information are to be included in the HANCO.

- Morphological information. In the HANCO, complete morphological description of every running word is given. The morphological analysis and the subsequent disambiguation procedure have been carried out automatically with further manual processing.
- Syntactic information. Syntactic information is given at three levels: word collocations, clauses, and sentences. The full description of units for every level is given according to the Academy Grammar of Russian.

Leeds University Corpora. In the 2000s a large number of corpora for different languages (including a Russian one) were created at the Leeds University by S. Sharoff.[2] Among them there is a version of the Russian National Corpus amounting to 116 mln words. The new frequency dictionary of Russian was created on its base. The interface to Russian corpora is supported by CQP IMS Corpus Workbench. It uses a special query language including Regular Expression language and allows for lexical and grammatical search. There are options to set the output interface. It is also possible to receive collocation lists calculated on the base of association measures such as MI, T-score, Log-likelihood. On this site[3] there is a collection of various software tools to process text corpus data.

Moshkov's Library Corpus. There is a big corpus of Russian texts, 680 mln words, on the site of NLP group created by A. Sokirko.[4] Using a powerful query language of the DDC corpus manager, it allows to search for lexical units taking into account parts of speech and morphological characteristics. At the same site there is a search service of bigrams (54 mln), calculated from a corpus by MI measure.

Sketch Engine Corpora. The English linguistic company Lexical Computing Ltd. (A. Kilgarriff) provides on a commercial basis an access to more than 140 corpora of various languages. Among them there is a few corpora of Russian, and primarily, the ruTenTen corpus created from Internet texts with the Wacky technology, totaling about 20 billion word usage. English researchers together with the Czech developers from the Masaryk University in Brno developed a special corpus manager, the Sketch Engine[5] [8]. The manager possesses many unique opportunities. Besides standard search with concordance output it issues lists of collocations based on individual syntactic models

[2] http://corpus.leeds.ac.uk/ruscorpora.html
[3] http://corpus.leeds.ac.uk/tools
[4] http://aot.ru/search1.html
[5] http://sketchengine.co.uk

(word sketches), forms a word frequency list, groups lexical units into lexico-semantic fields with internal clustering, and shows the strength of syntagmatic relations between lexemes.

2.6 Speech Corpora of Russian

Oral speech, and especially, the nonpublic oral improvised speech, according to many scientists, is the most important version of language, the closest to its "kernel", and showing the most characteristic models of language. Therefore it is important to dwell on Russian speech corpora.

The Corpus of Spoken Russian is the collections of transcripts of the spoken texts of different types (private speech, public speech, and movie speech-tracks). Its volume just now is around 10,5 million tokens. These transcripts are annotated morphologically and semantically by the RNC annotation system. In addition, the corpus has its own annotation: the accentological and the sociological one. The accentological annotation presupposes that in every word-form the real rather than normative stress is marked. Therefore, a user can investigate the history of Russian accentological system and its normative requirements, which are specific for any particular period. The sociological annotation means that to every text the information on the sex, the age and the name of a speaker is assigned, so a user can form his own subcorpora according to all these parameters and their combinations.

The Spoken corpus of the RNC gives a user various possibilities, but all these tasks must not be connected or based on the real phonation. Therefore the **Multimodal Russian corpus (MURCO)** was formed as a part of the RNC [9]. Its material are fragments of movies of the 1930s through the 2000s. The main principle of the MURCO is the alignment of the text transcripts with the parallel sound and video tracks. Consequently, when a user makes his data query he may obtain not only a written text, annotated from different points of view, but also the corresponding sound and video material. This possibility let a researcher use the obtained information at his will. He utilize his own manner of phonetic transcription or speech and intonation analyzers of his own choice; he may pose and solve all types of research tasks connected with phonetics, etc. The total volume of the movie transcripts in the RNC is around 3,5 million tokens.

The types of annotation in the MURCO are as follows:

- orthoepic annotation: combinations of sounds are marked;
- annotation of accentological structure: the word structure in regards to the stress position is defined;
- speech act annotation: the types of speech acts and vocal gestures, used in a clip are described;
- gesture annotation: the type of gesticulation in a clip is described.

Another speech corpus worth dwelling upon is the **ORD corpus** developed at Institute of Philological Researches of the St. Petersburg State University [10].The abbreviation ORD stems from Russian *Odin Rechevoj Den'*, literally translated as *one day of speech*. The main aim of creating the ORD corpus is to collect recordings of actual speech which we use in our everyday communication. For the first series of recordings

a demographically balanced group of 30 persons representing various social and age strata in the population of St. Petersburg was selected. These individuals spent one day with recorders dangling around their necks and recording all their communications. So recorders were on while having breakfast at home with the family, then while preparing to go to work, on the way to work, speaking on the cell phone, then business and informal conversations at work with colleagues, lunch time, shopping, recreation, etc. In the result more than 240 hours of recording were obtained with 170 hours containing speech data quite suitable for further linguistic analysis, and more than 50 hours of recordings good enough for further phonetic analysis. The corpus was divided into 2202 communication episodes. 134 episodes are already transcribed in detail. At present, orthographic transcription of the corpus numbers more than 50000 word-forms. The corpus presents the unique linguistic material, allowing to perform fundamental research in many aspects including complex behaviour of people in real world. These utterly natural recordings are to be used for practical purposes: to verify scientific hypotheses, to make adjustments to improve systems of speech synthesis and speech recognition, etc.

2.7 Special Corpora

A special corpus is a balanced corpus, of smaller size, as a rule, meant specifically for certain research tasks helping to resolve corresponding problems of user choice.

SPbEFL LC, a Learner English corpus (English as a foreign language), started at Herzen University (Saint Petersburg, Russia) in 2009 is a multi-mother tongue (Russian, Chinese, Japanese, Korean, Thai, and Vietnamese) corpus that compiles written texts (essays and personal letters), monologues, and dialogues (in scripts). The contributors' pre-tested language proficiency is intermediate (26%) and advanced (74%). The language/text relevant criteria include medium, genre, topic, technicality and task setting. SPbEF LC is an attempt to compile a target-specific structure, a text collection in accord with essential corpus design criteria. Operated with reliable free tools, the corpus proves efficient enough in spotting and analyzing the learner language with reference to syntactically parsed texts, concordance, frequency and collocation lists.

The corpus is aimed at interlanguage studies based on the assumption that both the vocabulary and the sentence patterns presumably reflect the actual language fund that the learners subconsciously resort to in case of FL communication. SPbEFL corpus findings pinpoint "atypical" mistakes in learner interpretation and use of basic structures that address issues of both learner universals and new language learning and teaching materials [11].

A new multimodal corpus of learners' spontaneous dialogues made according to a short outline is under construction in Irkutsk State Linguistic university (Russia). It is called **UMCO** *(Uchebnyj Multimodalnyj Corpus (Learner multimodal corpus))* and belongs to a category of learner's corpora. The dialogues are made up by the students of Chinese, Russian and German. Now UMCO consists of 25 video clips lasting from 1,5 to 3 minutes. ELAN is used as a corpus manager, for its multiple advantages, including the possibility of typing Chinese characters.

The corpus contains a number of parallel subcorpora where small thematic blocks of native speakers dialogues are aligned with those produced by the learners of the same language.

Among other specialized corpora worth mentioning are **Regensburg diachronic corpus of Russian** (texts in Old Russian), **Corpus of Old Russia manuscripts** (birch-bark letters), a parallel Corpus for translations of **The Tale of Igor's Campaign**, a Corpus of electronic Russian Heritage **Manuscript**, a historic **St. Petersburg Corpus of Hagiographic Texts** of XV–XVII centuries (SKAT), etc.

The demand for corpora of specialized texts can be comparable with that for national ones. Any specialized branch corpus gives a specialist the most important material: professional terms in their typical context thus providing means to monitor terminology evolution including the birth of new terms.

3 Conclusion: Corpus Oriented Researches

At the moment all corpora of the Russian language and mostly the RNC are used by both Russian and foreign researchers. The RNC has English interface and the help system in English. Its subcorpora with their special annotation provide various possibilities for linguistic studies. The RNC site has a special division called *Studiorum*. It includes some data of researches in Russian language.

The studies based upon the semantic annotation are of special interest. There are a few works which address word sense disambiguation and lexical constructions – the chains of lexical units, one of which is usually a lexical constant and others are variables [12]. The basic results obtained in the experiments have to do with revealing and classifying of different types of context markers to specify different meanings of target words. The type and degree of specification of the RNC semantic annotation could provide the rules for associating context tags of special semantic classes with different meanings.

References

1. Zasorina, L.N. (ed.): Chastotnyi slovar' russkogo yazyka. Moskva (1977)
2. Yershov, A.P.: K metodologii postroeniya dialogovykh sistem. Fenomen delovoi prozy. Novosibirsk (1979)
3. Reznikova, T.I.: Slavyanskaya korpusnaya lingvistika. In: Plungyan, V.A. (ed.) Natsionalnyi Korpus Russkogo Yazyka: 2006–2008, Saint-Petersburg, pp. 404–465 (2009)
4. Natsionalnyi korpus russkogo yazyka: 2003–2005, Moskva (2005)
5. Natsionalnyi korpus russkogo yazyka: 2006–2008, Saint-Petersburg (2009)
6. Lashevskaja, O.N., Shemanaeva, O.J.: Semantic Annotation Layer in Russian National Corpus: Lexical Classes of Nouns and Adjectives. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, pp. 3355–3358 (2008)
7. Kopotev, M., Mustajoki, A.: Printsipy sozdaniya Helsingskogo annotirovannogo korpusa russkikh tekstov (HANCO) v seti Internet (Principles of the Creation of the Helsinki Annotated Corpus HANCO). Nauchno-tekhnicheskaya Informatsiya 2(6), 33–37 (2003)

8. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. In: Proceedings of the XIth Euralex International Congress, pp. 105–116. Universite de Bretagne-Sud., Lorient (2004)
9. Grishina, E.: Multimodal Russian Corpus (MURCO): General Structure and User Interface. In: Levická, J., Garabík, R. (eds.) Slovko 2009. NLP, Corpus Linguistics, Corpus Based Grammar Research, Bratislava, Slovakia, pp. 119–131 (2009)
10. Sherstinova, T.: The structure of the ORD speech corpus of Russian everyday communication. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 258–265. Springer, Heidelberg (2009)
11. Kamshilova, O.: Learner Language analysis in SPbEFL Learner Corpus. In: Learner Language, Learner Corpora. LLLC 2012 Conference, October 5-6. The University of Oulu (2012), http://www.oulu.fi/hutk/sutvi/oppijankieli/LLLC/LLLC2012_abstracts.pdf
12. Lashevskaja, O., Mitrofanova, O.: Disambiguation of Taxonomy Markers in Context: Russian Nouns. In: Jokinen, K., Bick, E. (eds.) 17th Nordic Conference on Computational Linguistics (NODALIDA 2009), Odense, Denmark. NEALT Proceedings Series 2009, vol. 4, pp. 111–117 (2009)