Spoken Russian in the Russian National Corpus (RNC)

Elena Grishina

Institute of Russian Language, RAS
121019, Volkhonka, 18/2, Moscow, Russia
rudi2007@yandex.ru

Abstract

The RNC now it is a 120 million-word collection of Russian text, thus, it is the most representative and authoritative corpus of the Russian language. It is available in the Internet at www.ruscorpora.ru. The RNC contains texts of all genres and types, which covers Russian from 19 up to 21 centuries.

The practice of national corpora constructing has revealed that it's indispensable to include in the RNC the sub-corpora of spoken language. Therefore, the constructors of the RNC have an intention to include in it about 10 million words of Spoken Russian.

Oral speech in the Corpus is represented in the standard Russian orthography. Although this decision made impossible any phonetic exploration of the Spoken Russian Corpus, but studying Spoken Russian from any other linguistic point of view is completely available.

In addition to traditional annotations (metatextual and morphological), in Spoken Sub-corpus there is sociological annotation.

Unlike the standard oral speech, which is spontaneous and isn't intended to be reproduced, Multimedia Spoken Russian (MSR) is otherwise in great deal premeditated and evidently meant to be reproduced. MSR is also to be included in the RNC: first of all we plan to make the very interesting and provocative part of the RNC from the textual ingredient of about 300 Russian films.

1. What is RNC?

Russian National Corpus (RNC) is a collection of Russian texts of 19–21 centuries, which are supplied with different types of annotation – morphological, semantic, metatextual. The project is carried out and managed by researchers from various institutes, scientific research sentres and universities of Russia (mainly in Moscow and Saint-Peterburg). The work of the RNC group is supported by the program "Philology and Informatics" of Russian Academy of Sciences.

The RNC has been functioning since April 2003 and is accessible at www.ruscorpora.ru. Now it contains circa 125 million words. In the next three years we plan to bring its capacity up to 200 million words[1]. Therefore, the RNC has become one of the most representative and authoritative corpora of Russian. Exactly for this reason our Corpus includes in its name the adjective "national".

The RNC contains all genres of Russian written texts – fiction (prose and drama) and non-fiction (periodicals, scientific texts, memoirs, letters, business documents, theological writings, and so on). A user of the RNC can form his subcorpus according to any parameter of metatextual, morphological and semantic annotation and according to a combination of all possible parameters.

2. The Necessity of Spoken Data in a Corpus

The practice of worldwide corpus construction proved that information of any language in any corpus can't be evaluated as full without presentation of Spoken Speech. Stenographs of Spoken English constitute about 10% of volume of the British National Corpus (about 10 million words)[2]. The Czech National Corpus (CNK) also contains the spoken sub-corpora, the so-called Prague spoken corpus, but it is not so large – about 800 000 words, 300 recorded conversations[3].

At first the RNC was considered to be a corpus of written texts but as the corpus enlarges the possibility to include in the collection spoken texts appears. The researchers had at their disposal an amount of dialect[4] and literary spoken texts. For extralinguistic reasons the sociolinguistic balance of the spoken sub-corpus of the RNC isn't possible at the moment, because the solution of this problem demands great resources (the founders of the CNC point out similar problems) However, we assume that the large capacity of the sub-corpus will average, or even compensate, the lack of sociolinguistic balance. At the moment the prospective capacity of the spoken sub-corpus is 10 million words – which is comparable with the spoken sub-corpus of the BNC. At present volunteers from many prominent Russian philological scientific centers work at replenishment of the spoken part of the RNC. Furthermore collections of texts were furnished by various scientific institutions, namely the Institute of Russian Language (Russian Academy of Sciences), philological faculties of the Saint-Petersburg State University and the Saratov State University, etc.

3. The Presentation of Spoken Data in RNC

It's obvious that spoken texts must be represented quite differently from written texts. The first problem, we faced with, was the choice of a form of representation of material in the corpus – as that's quite clear that spoken and written texts cannot be represented in a similar way. There are three methods of data representation – audio record, transcription and standard orthographic record. Audio record in combination with normal spelling seems to be an ideal variant but this decision presents two difficulties: 1) site designers have to develop software which brings any part of a record into correlation with the appropriate part of spelling, 2) the PC of RNC user has to meet some technical requirements which let user turn to the record as well as to the spelling.

The researchers reckoned this problems too complicated and effortful.

[1] The prehistory and the contemporary state of the RNC are possible to learn from (НКРЯ, 2005).

[2] See (BNC, 2000).

[3] See (Čermák, 2001).

[4] As for dialectal texts in RNC see (Летучий, 2005).

The transcription representation also poses some questions: 1) volunteers have to learn and use correctly appointed phonetic transcription, 2) user has to learn this transcription as well, 3) in Russia there are several rival transcription systems and that's quite an unsolvable task to reconcile and bring this trends together.

Taking into account all pros and contras authors consider the third method of data representation as the most appropriate and usable – so, all spoken text in the RNC are represented in simple orthographical spelling. This decision also has some drawbacks – obviously, it's impossible to use the corpus for study the phonetics of Russian. As a user can't turn to the sound or transcriptional origin, a number of ambiguities and vagueness remain. But this way of data presentation allows to work quite productive with all other language levels – morphology, word-formation, syntax and semantic as well as with rhetorical aspects of conversation. As for other ways of presentation of spoken text – we reckon that they're more suitable not for national corpus but for specialized phonetical corpora (see, f. e., Кибрик, Подлесская, 2003).

We would like to remark especially that this decision is similar to the one the authors of the BNC have made. In the BNC all initial material was collected in tape form, afterwards, these records were transcribed but nevertheless the authors of the BNC made no attempt to reflect phonetic or prosodic characteristic of the speech and all spoken texts are orthographically transcribed (with the exception of so-called vocal pauses, regionalisms and dialectisms) (Burnage, Dunlop, 1992).

One more problem is a punctuation. We decided not to give punctuation marks according to the punctuation rules but to replace all the marks with slash characters (/) and to keep only the following marks: ., ..., !, ?.

Thus, slashes have no meaning and serve only for making reading more comfortable.

Here is the example, how looks out an extract of spoken text in of the RNC.

Модератор: Ну хорошо. Мы об этом поговорим еще. Кто-то еще? Вы там просто были в Белоруссии. А у кого-то вот / какие еще / просто мнения есть? Что вы думаете о современной Белоруссии?

БОРИС: Ну / мне кажется / что это передовая республика была / ну / бывшая республика / нерядовая. Во-первых / у нас здесь в Воронеже минская вся продукция / хорошая / мы даже покупаем / покупаем / хорошие холодильники / товары / опыт хороший / то есть там поставлено дело / мне кажется / нормально.

4. Standard Spoken Russian and the Types of Its Annotation

Each record in the RNC has some associated descriptive information particular to it, which let the user sort texts in one way or another and form his own subcorpora[5], i.e. each document is provided with meta-information. Below we are about to describe elements of tagging in cases when spoken texts differs from written texts or have some shades.

1. Author. An author of spoken text is an author of a monologue or a participant of a dialogue, a conversation. It's clear that the author isn't named when he is unknown – that's quite an often case in spoken texts. If the author's name (or names) is (are) known, the following restriction comes into force – it's quite desirable that any spoken text has no more than two authors. Thus, a main participant is reckoned an author of text, for example, in case of interview an author is the interviewee, not the questioner.

2. Title. Ordinarily spoken text has no title and it is to be generated artificially from the following fields: 1) author of the text 2) type/genre of the text 3) topic of the text 4) the date and 5) the place of the recording. If any of the attributes is not defined, it is simply omitted. If the spoken text was published, all appropriate information, i.e. the name of the editor, title and date of the edition and all necessary bibliographic information is included in the field **title.** In cases when a microdialogue has situational determination, the locus of the conversation is to be included in the title (for example, a dialogue in a drugstore, a conversation in a supermarket, in a police station, etc.).

3. Place of the recording. In written corpus, there is no such attribute – the location is defined only for publishing house – but it's quite obvious that for most of spoken colloquial texts and for all dialectal texts this characteristic is extremely important.

4. Date of the recording. This attribute corresponds with the date of publishing of the work in the written corpus.

5. Area of functioning. There are two main sub-units for the scope of functioning of the spoken language: 1) ***Spoken Public Speech***, which is presented wittingly to assumed listeners and which is a priori to be recorded, and 2) ***Spoken Private Speech,*** which isn't directed to outer listeners and doesn't suppose to be recorded. We would like to note that in the BNC spoken texts are divided into two similar parts: 1) **Context-Governed Part of the Spoken Corpus** — corresponds with the ***Spoken Public Speech,*** 2) **the Demographic Part of the Spoken Corpus** — corresponds with ***Private Speech*** and includes records of situation-governed texts (see BNC, 2000).

6. Genre of text. For ***Spoken Public Speech:*** monologue – lecture, speech, comment, narrative, presentation, homily, report, etc., dialogue (two or more participants) – talk, discussion, interview, conference, hearings, press conference, seminar, debate. ***For Spoken Private Speech:*** monologue – narration, story, dialogue – everyday talk, conversation, remembrance, wrangle, telephone conversation. For more details about genres of spoken texts in Russian linguistic tradition, see (Розанова, Китайгородская, 1999).

The ***Spoken Public Speech*** has a developed system of self-denomination (vide supra). It was our aim to surmount this variety and to consolidate this classification as much as possible. Thus, for example, talk show isn't considered as a separate genre, as all texts of this kind can be completely distributed among such basic genres as discussion and debate.

Genre scheme of ***Spoken Private Speech*** is organized quite in a different way: private speech has no genre self-denomination, so it is an entire united *element of conver-*

[5] As for general principles of metatextual annotation see (Савчук, 2005).

sation. The aim of the researchers was to designate different areas of this global conversation in a uniform, user-evident way, describe them and put in the corpus as separated units.

For genre description and designation, a set of so-called ***dominants*** was used. Some dominant is peculiar to some specific genre but in other kinds of texts it is slackened, so that every genre has it's own set of dominants. For example, in public monologues we should point out such dominants as *teaching* (for lections and homilies), *description of the current events* (comments, presentation), *narration about past events* (narrative, story), *an impact on hearer* (speech), *information* (report).

For ***spoken private speech,*** a dominant of standardization is extremely important, namely, a speaker has to keep the rules and observe the standard structure to be understood. Exactly this dominant, but not a monologue-dialogue opposition, has decisive importance in classification of spoken private speech, inasmuch as quite pure monologue in private speech doesn't exist.

Just according to this characteristic, micro-dialogues and telephone conversation could be separated from other types of private speech. For standardized dialogues, locus (i.e. where exactly, in what situation the (standardized) conversation takes place) is a determinant dominant.

As the standardization dominant weakens, other ones take action: if the speech is denial-oriented (argument) or it is an interlocution per se (idle talk, everyday conversation), if it is a telling about the past (narrative) or it's a description of some current events, etc.

Remarks designation. Record of speech contains some elements, which analogue one can only in plays find, i.e. ***metatextual remarks.*** It would be methodologically incorrectly not to distinct these metatextual remarks from the body of the text – it could, for example, appreciably change frequency characteristics of a text. At the same time, if to refuse from use of these remarks, that could lead to ambiguities and vagueness in the text.

There are two types of metatextual remarks. First, it's remarks as such (like "everybody is mute", "he laughs", etc.) These common remarks are marked doubly by two unique tokens, which aren't used anywhere in the text, for example: #everybody is mute#.

Further, by parsing the text, the program finds such cases and they are shifted automatically at another level of the text. Thus, although there are still visible during common reading of the text, they are not taken into account by statistic analysis.

Another type of metatextual notes is an indication of the author of a remarque. They contain a sociological annotation, i.e. sex, age, profession, etc. of the author are indicated. This annotation is also extremely important (in particular for sociolinguistic researches), as it permits to create a user-defined sub-corpus of utterances of, for example, programmers, or teenagers, or women of 55 and upward, etc.

5. Multimedia Spoken Russian

The RNC is supposed to present the Russian language in all existent forms and conditions, from normative till extremely marginal. Most of them are already present in the corpus, tagged and accessible for users, some, for example poetry and songs, are to be included. Nevertheless, the RNC is lacking in one stratum of a language, the so-called multimedia texts, i.e. speech that is united with visual and acoustic perception. As we know, the sample corpora (the BNC and the ČNC first of all), which served as a pattern for creators of the RNC, contain this type of texts neither.

This lack is simple to explain: multimedia texts get into a gap between three basis forms of the language existence. That is written form, oral speech and e-speech (see Капанадзе, 2005) – all these forms are present in corpora, including the RNC. Multimedia texts are entirely spoken, so that they can't be referred to written or e-text. But it isn't either a spoken language per se, inasmuch as it hasn't its primary characteristics – spontaneity and non-reproductivity. Multimedia speech is always well prepared, written-to-be-spoken and, surely, not simply reproducible, but is intended for the constant reproduction.

Therefore, the lack of multimedia part in corpora is explicable but at the same time, it seems not to be founded in logic at all.

Here are classes of texts, which are considered to form the core of the multimedia sub-corpus. First of all, it's texts from feature and cartoon films, telecasts, broadcasts, television and radio commercials and libretti of operas, operettas and musicals. These texts are very important and influential in a culture (including the Russian one[6]), everyday spoken conversations abound with quotations from films. So, this lack might be compared with the lack in the RNC of «*Woe from Wit* » by A. Griboyedov or fables of I. Krylov, whereas the RNC contains texts with multiple quotations from the ones.

Here are in brief some features of annotation and internal design of multimedia texts. At this juncture matter concerns the cinematographic part of the corpus mainly, as the aim of the RNC creators within the next three years is to insert in the corpus textual constituent of at least 300 Russian films.

1. Author. Director and scriptwriter are considered as authors of films and TV-plays, but if a film has a literature basis, the author of the book must be indicated also. For example, among authors of a film "Ivan Vasilyevich changes profession" certainly M. Bulgakov must be named. All publicity text have a collective author, but some outstanding actors must be mentioned. Authors of TV-miniatures are an author of a text and a performer (for example, M. Zhvaneckij, R. Karcev). The situation with film's annotation is similar (see below). As for translated libretti not only an author of the text, but a translator must be named.

2. Field of functioning. Multimedia text differs in the field of functioning and there is no unity in this matter. For publicity and advertisement special subsection *multimedia publicity* in the section *publicity* is proposed to be introduced, other texts are to be attributed to a new field which can be preliminary designated as *multimedia fiction.*

3. Genre/kind of text. *Multimedia fiction* has it's own particular genre system. Some of them have been already

[6] For example, in a relatively small vocabulary (about 1000 words) (Шулежкова, 2003) 200 films and animated cartoons are mentioned. Quotations from these ones have penetrated in the Russian language and the Russian culture.

defined: it is comedy, thriller, action, science fiction, melodrama, drama. Most films and plays can be attributed to one of these kinds. However, there are still problems with determination of some genres, and we have to mention that similar problems have arisen while determination of kinds of fiction. In the issue a null, unmarked attribute has been introduced, a so-called *non-genre fiction.* Perhaps, a similar decision will be made for fiction multimedia but at the moment we don't have another material to be firmly convinced so it's too early to discuss the final decision. As for *multimedia publicity*, for the present there are two kinds of it: publicity (trailer) and advertisement.

Other attributes of the annotation of multimedia texts won't change in comparison with the general one.

4. The internal text tagging There are at least two problems by internal tagging of multimedia corpus. First, it's a method of tagging of author's remarks. Here we'll use the same way as by tagging of author's words in dramas and Standard Spoken Speech. Therefore, these remarks are visible to user but belong to another structural level so that the constant repetition of the remarks wouldn't influence on the statistic. Besides, the eduction of author's words on another textual level indicates that it isn't the main textual body of a film but some annotation, informational addition of the RNC's creators which makes the corpus more user-friendly. Here is the example of such an annotated passage from the film "Gentlemen of Good Luck" by G. Danelia. A famous remark «Чуть что – сразу Косой!» (*On the least occasion – just Kosoj*) would be represented in such a way: «**Косой:** Чуть что – сразу Косой!»

But the incompleteness of such a representation is quite obvious. As many researchers of the film language remark, cites from films are inseparably linked with the intonation, some peculiarities of articulation of the actor (see Елистратов, 1999) – so we have to name him either. The full and right representation will look so: «**Косой / С. Крамаров:** Чуть что – сразу Косой!».

The second problem is concerned with the first one, it's how much supplementary auxiliary information the RNC must contain. It's obvious that one can't avoid including in the corpus remarks and explanations of hero's acts, the intrigue, some events. But these inclusions must be strictly limited, otherwise the size of annotation will exceed the text from the film. It's especially so for films with little text (for example, some films by A. Tarkovsky). Therefore, our main initial directions are to avoid the annotation as much as possible, not to explain the acts and events, but use some remarks only when it's extremely inevitable. In cases when the remark is required nevertheless, if, for example, some phrase is unclear or ambiguous without it, it is to be marked like remarks in dramatic texts and spoken texts, namely it is to be converted on another structural level and considered out of the textual body of the film.

The body of the text is to be divided into parts according to the principle of place- and event-unity and these parts are to be indicated like chapters in written texts.

That's the preliminary description of the multimedia part of the RNC. In conclusion we'll mention that at the moment we can't designate the capacity of this sub-corpora not only because there are no samples of representation of such material in other corpora, but because average textual extent of a film of 1,5-2 hours duration hasn't defined yet. Therefore, the capacity of the main part of this sub-corpus, cinema texts, isn't certain.

References

BNC (2000). The BNC Users Reference Guide, In *http://www.natcorp.ox.ac.uk/World/HTML.*

Burnage, G., Dunlop, D. (1992). Encoding the British National Corpus. In *Oxford University Computing Services Published in English Language Corpora: Design, Analysis and Exploitation, Papers from the 13th international conference on English Language research on computerized corpora*, Nijmegen. http://www.natcorp.ox.ac.uk/ using/papers/Burnage93a. htm.

Čermák, F. (2001). Language Corpora: The Czech Case In Text, Speech and Dialogue, In *TSD 2001*, eds. V. Matoušek, P. Mautner, R. Mouček, K. Taušer. Springer Berlin etc., pp. 21-30.

Елистратов, В.С. (1999). Русский кинемалогос (о целях и структуре словаря). In *Елистратов, В.С. Словарь крылатых слов (русский кинематограф).* М.: 1999, сс. 3-10.

Капанадзе, Л.А. (2005). На границе письменного и устного текста: структура и тенденции развития электронных жанров. In *Капанадзе, Л.А. Голоса и смыслы. Избранные работы по русскому языку.* М.: ИРЯ, сс. 305-320.

Кибрик, А.А., Подлесская, В.И. (2003). К созданию корпусов устной речи: принципы транскрибирования. In *Научно-техническая информация. Сер. 2. Информационные процессы и системы*, *№ 10,* М.: ВИНИТИ, сс. 5-12.

Летучий, А.Б. (2005). Корпус диалектных текстов: задачи и проблемы. In *НКРЯ, 2005*, сс. 215-232.

НКРЯ (2005). Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 344 с.

Розанова, Н.Н., Китайгородская, М.В. (1999). Речь москвичей. Коммуникативно-культурологический аспект. М.: «Русские словари», 396 с.

Савчук, С.О. (2005). Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции. In *НКРЯ, 2005*, сс. 62-88.

Шулежкова, С.Г. (2003). Словар крылатых выражений из области искусства, М.: «Русские словари», 430 с.

English translation – Sophie Piskunova (Russia)