Multimodal Parallel Russian Corpus (MultiPARC): Main Tasks and General Structure

Elena Grishina, Svetlana Savchuk, Dmitry Sichinava
Institute of Russian Language RAS
18/2 Volkhonka st., Moscow, Russia
rudi2007@yandex.ru, savsvetlana@mail.ru, mitrius@gmail.com

Abstract

The paper introduces a new project, the Multimodal Parallel Russian Corpus, which is planned to be created in the framework of the Russian National Corpus and to include different realizations of the same text: the screen versions and theatrical performances of the same drama, recitations of the same poetical text, and so on. The paper outlines some ways to use the MultiPARC data in linguistic studies.

1. Introduction

It is generally known that the main drawbacks and difficulties in the speech researches are connected with the fact that speech is not reproducible. It seems that we have no possibility to repeat the same utterance in the same context and in the same circumstances. These limitations lose their tension, when we deal with the etiquette formulas, and with other standard social reactions of a fixed linguistic structure. But unfortunately, the standard formulas of the kind are quite specific and may hardly represent a language as a whole. So, we may state that a spoken utterance is unique, in a sense that it takes place on one occasion only, here and now, and cannot be reproduced in combination with its initial consituation.

On the other hand, the question arises what part of this or that utterance is obligatory to all speakers in all possible circumstances, and what part of it may change along with the changes of speakers and circumstances. The only possible way to solve the problem is to let different speakers utter the phrase in the same circumstances. Naturally, the real life never gives us the possibility to put this into practice, laying aside the case of linguistic experiment. But the sphere of art lets us come near the solution.

To investigate the ways of the articulation of the same utterance by different speakers, but in the same circumstances, the RNC[1]-team decides to create a new module in the framework of the Multimodal Russian Corpus (MURCO[2]), which is supposed to be named Multimodal Parallel Russian Corpus (MultiPARC).

2. Three parts of MultiPARC

2.1 Recitation

We suppose that the Recitation zone of the MultiPARC will include the author's, the actor's, and the amateur performances of the same poetic or prosaic text. We plan to begin with the poetry of Anna Akhmatova, who is quite popular among professional actors and ordinary readership; besides, a lot of recordings of Akhmatova's recitations of her own poetry are easily available. There are no comparable corpora of the kind functioning at the present moment, as far as we know.

2.2 Production

MultiPARC will also include the different theatrical productions and screen versions of the same play. For example, we have at our disposal one radio play, three audio books, three screen versions, and seven theatrical performances of the Gogol's play "The Inspector General" ("Revizor"). As a result, the MultiPARC will give us the opportunity to align and compare 14 variants of the 1st phrase of the play: *I have called you together, gentlemen, to tell you an unpleasant piece of news. An Inspector-General is coming.* Naturally, every cue of the Gogol's play may be multiplied and compared in the same matter. And not only the Gogol's play, but also the plays of Chekhov, Vampilov, Rosov, Ostrovsky, Tolstoy, and so on. The only requirement to a play is as follows: it ought to be popular enough to have at least two different theatrical or screen versions.

The comparison of different realization of the same phrase, which is meant to be pronounced along with the same conditions and circumstances, but by the different actors, gives us the unique possibility to define, which features of this or that utterance are obligatory, which are optional, but frequent ones, and which are rare and specific only for one person.

Naturally, here we face the restrictions, which are connected with the artificiality of the theatrical and movie speech. Though, we definitely may come to some interesting and provoking conclusions concerning the basic features of spoken Russian, and probably of spoken communication as a whole.

2.3 Multilingual zone

The above section naturally brings us closer to the most debatable and open to question zone of the MultiPARC, namely the multilingual one. Here we suppose to dispose the theatrical productions and screen versions on the same play/novel, but in different languages (American and Russian screen versions of Tolstoy's "War and Peace",

[1] About the RNC see [RNC'2006, RNC'2009], [Grishina 2007], www.ruscorpora.ru; about the spoken subcorpora of the RNC see, among others, [Grishina 2006, 2007], [Grishina et al., 2010], [Savchuk 2009]).

[2] About the MURCO see, among others, [Grishina 2009a, 2009b, 2010].

French and Russian screen versions of "Anna Karenina", British and Russian screen versions of "Sherlock Holmes", and so on).

This zone of the MultiPARC is intended for the investigation in two fields: 1) comparable types of pronunciation (pauses, intonation patterns, special phonetic features, like syllabification, chanting, and so on), which are often the same in different languages, 2) comparable researches in gesticulation, which has its specificity in different cultures. We think that this zone of the MultiPARC may become the subject of international cooperation.

3. MultiPARC interface

The MultiPARC in total is supposed to have the interface, which is adopted just now for the MURCO. The user's query will return to a user a set of clixts, i.e. a set of the pairs 'clip + corresponding text', the corresponding texts being richly annotated. But the MultiPARC seems to have some specific features. The investigation of movie and theatrical speech has shown that the actors regularly transform the original texts of a play (see [Grishina 2007]). We often meet the transformations of the following types:

1) additions
2) omissions
3) shifts and transpositions
4) synonymic equivalents
5) apocopes
6) restructuring, and some others.

(It should be noted parenthetically that these linguistic events take place also in poetry, though quite rarely.)

As a result, the real cue pronounced on the stage or on the screen may differ considerably from the corresponding cue in the prototypical text. Consequently, the MultiPARC interface ought to provide two types of queries: 1) query for the prototypical cue, 2) query for the real cue (see Pic. 1).

If a user makes a query, which refers to the prototypical cue, then he/she receives the clusters of the real cues (i.e. the complete set of the clixts, which correspond to this very prototypical cue). But if a user makes a query, which refers to the unit (word, construction, combination of letters, accent, and so on) included in a real cue, but missing in the prototypical one, then he/she receives in return only the real cues, which contain this unit.

4. Types of Annotation

Since the MultiPARC is the result of further development of MURCO, it is quite natural that it will be annotated under the MURCO standards. These are as follows:

- metatextual annotation
- morphological annotation
- semantic annotation
- accentological annotation
- sociological annotation
- orthoepic annotation
- annotation of the vocalic word structure

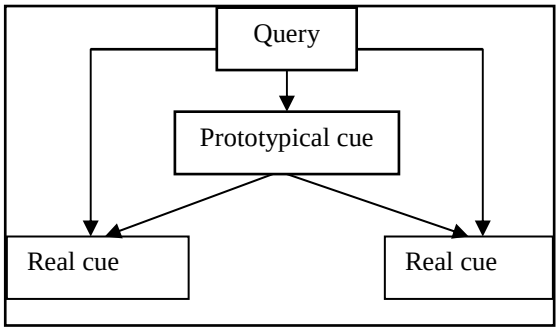
We have described all types of MURCO annotation earlier ([Grishina 2010]), so we need not to return to the question.

5. MultiPARC as Scientific Resource

MultiPARC is meant to be one of the resources for scientific researches, so its main task is the academic one. Being the academic resource, it lets us put and solve the scientific tasks, which concern following fields of investigation.

1. The regularities of the pause disposition in spoken Russian. The types of pauses from the point of view of their
1.1. obligatoriness
1.2. phonetic characteristics
1.3. duration
may be investigated systematically.
2. The regularities of the intonation patterns, which accompany the same lexical and syntactical structures.
3. The correspondence between punctuation marks and pause disposition.
4. The correspondence between the punctuation marks and intonation patterns.
5. The regularities of the change of the word order in spoken Russian in comparison with written Russian.
6. The set and ranking of clitics (proclitics and enclitics) in spoken Russian.
7. The correspondence between the communicative structure of a phrase (theme vs. rheme) and the most frequent manners of its pronunciation from the point of view of phonetics and intonation.

Below we mean to illustrate the above with some interesting observations.



Picture 1

6. Usage of MultiPARC

6.1 Syllabification in Spoken Russian

The trial version of the MultiPARC, which is being prepared just now, let us illustrate some types of its prospective usage in scientific studies. For example, we may investigate the role of some phonetic phenomena in Spoken Russian.

Let us analyze the beginning of the classic Gogol's play "The Inspector General" ("Revizor") from this point of view. The comparison of first 37 fragments gives us the possibility to analyze the main types of meaning of syllabification in Spoken Russian.

6.1.1. The highest degree of quality

Hereinafter the first figure in the brackets refers to the number of the utterances with the syllabification, the second figure refers to the total number of the utterances, and the percentage means the comparative quantity of the syllabicated utterances (it will be recalled that we have compared 14 realizations – the

theatrical performances, movies, audio books – of the same play).

The syllabification is used to mark up the words and word-combinations, which include the component 'the highest degree of quality' in their meaning (hereinafter these words and words-combinations are bold-faced).

The corresponding illustrations are as follows.

It would be better, too, if there ***weren't so many of them***. (5-11-45%)

I have called you together, gentlemen, to tell you an ***unpleasant*** (6-14-43%) *piece of news.*

Upon my word, I never saw the likes of them — black and ***supernaturally*** (6-14-43%) *big.*

The attendants have turned the entrance hall where the petitioners usually wait into a poultry yard, and the geese and goslings ***go poking their beaks*** (5-12-42%) *between people's legs.*

Besides, the doctor would have ***a hard time*** (4-11-36%) *making the patients understand him.*

An ***extraordinary*** (4-13-31%) *situation, most extraordinary!*

He ***doesn't know a word*** (3-11-27%) *of Russian.*

Last night I kept dreaming of two rats — ***regular monsters***! (4-14-26%)

And I don't like your invalids to be smoking ***such strong tobacco***. (3-10-30%)

You ***especially*** (2-12-17%), *Artemy Filippovich.*

Why, you might gallop ***three years away from here*** (1-14-7%) *and reach nowhere.*

6.1.2. Important information, maxims and hints

The syllabification is used to mark the information of heightened importance. This group includes the suggestions and hints:

Yes, an Inspector from ***St. Petersburg***, (2-14-14%) ***incognito***. (9-14-64%) *And with* ***secret instructions***, (5-14-36%) *too.*

I had a sort of ***presentiment*** (5-14-36%) *of it.*

It means this, that ***Russia — yes — that Russia intends to go to war***, (9-13-69%) *and the* ***Government*** (4-13-31%) *has secretly commissioned an official to find out* ***if there is any treasonable activity anywhere***. (7-14-50%)

On the look-out, or not on the look-out, anyhow, gentlemen, I ***have given you warning.*** (3-14-22%)

In addition, this group includes the maxims. The maxims are the utterances stating something to be absolutely true, without any reference to time, place, and persons involved. Therefore the maxims are accompanied with the syllabification quite often to underline the importance and significance of the conveying ideas:

Treason in this little country town! (= 'It is impossible to have treason in this little country town') (4-14-29%)

The ***Government is shrewd.*** (2-14-14%) *It makes no difference that our town is so remote. The Government is* ***on the look-out all the same.*** (3-14-21%)

Our rule is: ***the nearer to nature the better***. (7-12-58%) *We use no expensive medicines.*

A ***man is a simple affair***. (3-13-23%) *If he dies, he'd die anyway. If he gets well, he'd get well anyway.*

6.1.3. Introduction of the other's speech

Third group of syllabification is quite specific. It includes the utterances, which introduce the other's speech or autoquotations. Generally, the introduction precedes the other's speech, but sometimes it summarizes the citation. This group also includes the introductions of one's thoughts and opinions:

"My dear friend, godfather and benefactor — [He mumbles, glancing rapidly down the page.] *— and* ***to let you know*** (4-14-26%)*"— Ah, that's it* [he begins to read the letter aloud]

Listen to what he ***writes*** (3-14-22%)

It means this, (4-13-31%) *that Russia — yes — that Russia intends to go to war*

My opinion is (2-13-15%), *Anton Antonovich, that the cause is a deep one and rather political in character*

I have made some arrangements for myself, and ***I advise you*** (2-12-17%) *to do the same.*

So, the tentative studying of the MultiPARC data has shown that it may give us the possibility to study the semantics and functions of different phonetic phenomena in Russian systematically.

6.2 Types of pauses

The MultiPARC presents the data to investigate the types and the usage of the pauses in Spoken Russian. The preliminary analysis has shown that there are 4 types of pauses as for their frequency:

1) obligatory pauses; frequency 80-100%

I have called you together, gentlemen, to tell you an unpleasant piece of news. ‖ (14-14-100%) *An Inspector-General is coming.*

2) frequent pauses; frequency 50-79%

I advise you to take precautions, ‖ (11-14-79%) *as he may arrive any hour,* ‖ (8-14-57%) *if he hasn't already, and is not staying somewhere* ‖ (8-14-57%) *incognito.*

3) sporadic pauses; frequency 20-49%

Oh, that's a small ‖ (2-11-14%) *matter.*

4) unique pauses; frequency 8-19%.

Oh, as to ‖ (1-13-8%) *treatment, Christian Ivanovich and I have worked out* ‖ (1-13-8%) *our own system.*

Having distinguished the different types of pauses, we may analyze the correlation between

1) the frequency of pauses and the punctuation marks;

2) the duration of pauses and their frequency;

3) the types of pauses and the types of the syntactic boundaries;

4) we may also systematically investigate the expressive features of the unique pauses.

As for the last point, we may notice that breaking up the combination of an attribute and a determinatum (AD) into two parts with a pause is a quite seldom event. In 37 surveyed fragments of the Gogol's play we may see 21 combinations AD without any pauses between A and D, and only 7 combinations with the unique pauses: A‖D. As a result, the pause in the constructions like AD has a great expressivity and underlines the importance of the attribute.

7. Conclusion

We may see that the Multimodal Parallel Russian Corpus (MultiPARC) present the new type of the multimodal corpora. This corpus gives a researcher the possibility to analyze the spoken events from the point of view of their frequency,

singularity, expressiveness, semantic and syntactic specificity, and so on.

Moreover, the MultiPARC presents the data for the gestural investigations. For example, the eye behavior (namely, blinking), which is specific for the professional actors while declaiming poetry, is quite different from this of non-professional performers. Since the MultiPARC is planned to include video, we may obtain the gestural data from different screen versions and theatrical performances. So, the contrastive analysis of the data is available.

8. Acknowledgements

The work of the MURCO group and the authors' research are supported by the program "Corpus Linguistics" of the Russian Academy of Sciences and by the RFBR (The Russian Fund of Basic Researches) (RFFI) under the grants 10-06-00151 and 11-06-00030.

9. References

Grishina, E. (2006). Spoken Russian in the Russian National Corpus (RNC). In *LREC'2006: 5th International Conference on Language Resources and Evaluation. ELRA*, pp. 121-124.

Grishina, E. (2007b). Text Navigators in Spoken Russian. In *Proceedings of the workshop "Representation of Semantic Structure of Spoken Speech"* (*CAEPIA'2007, Spain, 2007, 12-16.11.07, Salamanca*). Salamanca, pp. 39-50.

Grishina, E. (2009a). Multimodal Russian Corpus (MURCO): types of annotation and annotator's workbenches. In Corpus Linguistics Conference CL2009, Universuty of Liverpool, UK, 20-23 July 2009,

Grishina, E. (2009b). Multimodal Russian Corpus (MURCO): general structure and user interface. In NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference, Smolenice, Slovakia, 25-27 November 2009. Proceedings. Tribun, 119-131, http://ruslang.academia.edu/ElenaGrishina/Papers/153531/Multimodal_Russian_Corpus_MURCO_general_structure_and_user_interface

Grishina, E., et al. (2010). Design and data collection for the Accentological corpus of Russian. In In *LREC'2010: 7th International Conference on Language Resources and Evaluation. ELRA* (forthcoming).

Grishina E. (2010) Multimodal Russian Corpus (MURCO): First Steps // 7th Conference on Language Resources and Evaluation LREC'2010, Valetta, Malta. l

RNC'2006. (2006). *Nacional'nyj korpus russkogo jazyka: 2003–2005. Rezul'taty i perspektivy.* Moscow: Indrik.

RNC'2009. (2009). *Nacional'nyj korpus russkogo jazyka: 2006–2008. Novyje rezul'taty i perspektivy.* Sankt-Peterburg: Nestor-Istorija.

Savchuk, S. (2009). Spoken Texts Representation in the Russian National Corpus: Spoken and Accentologic Sub-Corpora. In *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference, Smolenice, Slovakia, 25-27 November 2009. Proceedings.* Brno, Tribun, pp. 310-320.