Multimodal Russian Corpus (MURCO): First Steps

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

Abstract

The paper introduces the Multimodal Russian Corpus (MURCO), which has been created in the framework of the Russian National Corpus (RNC). The MURCO provides the users with the great amount of phonetic, orthoepic, intonational information related to Russian. Moreover, the deeply annotated part of the MURCO contains the data concerning Russian gesticulation, speech act system, types of vocal gestures and interjections in Russian, and so on. The Corpus is on free access. The paper describes the main types of annotation and the interface structure of the MURCO. The MURCO consists of two parts, the second part being the subset of the first: 1) the whole Corpus, which is annotated from the lexical (lemmatization), morphological, semantic, accentological, metatextual, socioligical point of view (these types of annotation are standard for the RNC), and also from the point of view of phonetics (the orthoepic annotation and the mark-up of accentological word structure), 2) the deeply annotated MURCO, which is annotated in addition from the point of view of gesticulation and speech act structure.

1. Introduction

As the programs of LREC'2008 and LREC'2010 have shown, the construction and the creation of multimodal corpora are doubtless the mainstream of the contemporary corpus linguistics. The elaboration of the multimodal corpora follows 4 lines: 1) speech act classification and identification of the types of the dialogue moves, which are specific for various real situations (Strauß et al., 2008; Möller et al., 2008; Georgila et al., 2008; Kostoulas et al., 2008; Brutti et al., 2008, Marasek & Gubrynowicz, 2008; Nallasamy et al., 2008); 2) identification and specification of the human affects and emotions and their connections with speech and gesticulation (Forbes-Riley et al., 2008; Gnjatovíc & Rösner, 2008; Wilson, 2008; Devillers & Martin, 2008; Sainz et al., 2008; Fék et al., 2008; Cullen et al., 2008); 3) investigations in the area of thematic development of dialogue, including the problems of anaphora and the reference as a whole (van Son et al., 2008; Stoia et al., 2008; Gallo et al., 2008; Wilks et al., 2008); 4) creation of the specialized gesture corpora or the gesture components of the multimodal corpora, and elaboration of the gesture classifications and the set of the parameters of gesture description (van Son et al., 2008; Savino et al., 2008; Knight & Tennent, 2008; Blache et al., 2008).

The construction and creation of multimodal corpora come across some commercial and legal obstacles. Firstly, the multimodal corpora, which have been created as the parts of various business projects, very often become inaccessible for an ordinary user. Secondly, the multimodal corpora dealing with the real persons as the informants face the legal difficulties concerning copyright offence and privacy invasion.

It seems that the decisions and suggestions that have been chosen by the MURCO constructors (in spite of their shortcomings) let us to cross the mentioned obstacles and to create the resource which can be useful for the researchers in the diverse fields of linguistics.

2. Basic MURCO Principles

2.1 Spoken Component of RNC

So far, the RNC contains the Spoken Subcorpus (just now its volume is circa 8 million tokens), but this subcorpus does not include the oral speech proper – it includes only the transcripts of the spoken texts (Grishina, 2006). The structure of the Spoken Subcorpus of the RNC is as follows:

Types of texts	Million tokens	Percentage
Public spoken Russian	4.4	51%
Private spoken Russian	0.8	10%
Movie speech	3.4	39%

Table 1: Spoken Russian in RNC

It is absolutely natural that to supplement and to replenish the Spoken Subcorpus of the RNC, or, to be more precise, to transform it, we have to work out the generally accessible and relatively fair-sized multimodal corpus. To avoid the legal problems mentioned above, we have decided to use the cinematographic material in the MURCO.

Naturally, in the future we are also going to include in the MURCO the patterns of the public and private spoken Russian, but the cinematographic Russian is the most appropriate material to begin the project with. It should be mentioned inter alia that the usage of the cinematographic material to elaborate and test the annotation system of the pioneering corpus is far more promising than the usage of the "natural" (public or private) spoken Russian. The main reason for it is the fact that the cinema includes exceptionally manifold set of situations, and this situational variety results in the linguistic variety. Therefore, to annotate the movie Russian we need greater number of

definitions and more elaborated system of concepts than to annotate the "real-life" Russian. In other words, the exercised annotation of the movie Russian will be useful for the markup of the "natural" Russian, but the opposite is not right.

There are some features, which distinguish the natural spoken speech and the cinematographic one (first of all we mean the parameter of the text coherence), but the differences though remarkable are not crucial (see (Grishina, 2007a, 2007b) about the usage of the discourse markers in the movie transcripts; the strategy of their usage is virtually the same in the natural and cinematographic spoken Russian); that is to say, the higher coherence of the movie transcripts in comparison with the transcripts of the natural spoken texts does not turn the former into the written texts: they remain spoken ones (Forchini, 2009).

2.2 Outputting Units in MURCO

The MURCO is the collection of the clixts. A ***clixt*** is the pair of a *clip* and the corresponding *text* (i.e. the corresponding part of a movie transcript). It is supposed that a user will have the opportunity to download not only the text component of a clixt (=marked up transcript), but also its sound/video component, so after downloading a user may employ any program to analyze it. The duration of a clip is within the interval of 5-20 sec.

As we have mentioned above, just now the total volume of cinematographic transcripts in the Spoken Subcorpus of the RNC is 3.4 million tokens. In the near future we will bring it up to 5 million tokens. Therefore, if we manage to transform this subcorpus into multimodal state, we will obtain one of the largest open multimodal corpora, so the task is ambitious enough.

3. Types of Annotation in MURCO

Since a clixt contains sound (=speech) and/or video tracks, it will be annotated from the point of view of text, sound and video. Therefore, the total structure of the MURCO annotation ought to be as follows:

Annotation zone	Types of annotation
Text	Standard RNC annotation Speech act annotation
Sound	Orthoepic annotation
Video	Gesture annotation

Table 2.

So, we see that some types of annotation in the MURCO are standard and quite usual for the RNC; the other ones are absolutely new and specific only for the MURCO.

The standard RNC annotation includes 5 types (RNC, 2006; RNC, 2009):

- metatextual annotation
- morphological annotation
- semantic annotation
- accentological annotation
- sociological annotation

All these types of annotation will be preserved in the MURCO.

Three types of annotation, which are specific for the MURCO, are as follows:

- the orthoepic annotation
- the speech act annotation
- the gesture annotation.

We'll describe these new annotation types below. It ought to be mentioned that the orthoepic annotation differs from the speech act and gesture annotation from the point of view of the obligation degree. Since the orthoepic annotation is planned to be automatic, it will be obligatory in all texts which will be included in the MURCO. On the contrary, the speech act and gesture annotations are manual; therefore, they will be used in the so called "deeply annotated" texts[1].

3.1 Standard RNC and MURCO Annotation

3.1.1. Metatextual Annotation

Every text in the RNC is supplied with the extralinguistic sociological information, which characterizes a text as a whole. This information forms the so called metatextual annotation. The main items of the metatextual information concern the author's characteristics (name, age) and the text-as-a-whole characteristics (title, date of creation, genre, and so on). In the MURCO the metatextual annotation of a movie transcript as a whole will be attributed to every clixt derivable of this movie.

3.1.2. Morphological Annotation

The morphological annotation in the RNC is provided with the automatic morphological parser "MyStem", which has been elaborated by the team of the Yandex, the biggest Russian search engine. Every token in the RNC is supplied with morphological information. The morphological string contains a lemma, a part of speech, the constant grammatical characteristics (e.g. gender for nouns, aspect and transitivity for verbs), the variable grammatical characteristics (e.g. case for nouns, gender-case-number for adjectives, person for verbs, and so on). The search is possible according to all these parameters.

3.1.3. Semantic Annotation

The texts in the RNC are semantically tagged with the program named *Semmarkup* (elaborated by A. Polyakov), which is based on the semantic dictionary of the RNC. The latter, in its turn, is founded on the database *Lexicograph*, elaborated under the leadership of E. Paducheva and E. Rakhilina (Russian Academy of Sciences).

Every word in the RNC is supplied with the semantic characteristics, which includes three types of tags:

- Class (a name, a reflexive pronoun, etc.)
- Lexical and semantic features (a lexeme's thematic

[1] Naturally, we do not plan to annotate every film which will be included in the MURCO, from the point of view of its speech act and gesture structure. We suppose the volume of the deeply annotated subcorpus of the MURCO ought to be about 1 million tokens.

class, indications of causality or assessment, etc.)
- Derivational features (a diminutive, an adjectival adverb, etc.)

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.

3.1.4. Accentological Annotation

The annotation of the *spoken* texts (including the movie transcripts) in the RNC contains in addition the accentological (Grishina, 2008; Grishina, 2009a; Savchuk, 2009; Grishina et al., 2010) and sociological (Grishina & Savchuk, 2009) information.

It is widely known that the stress in Russian is free and mobile, so the accentological information and the possibility of finding a word with this or that location of the stress mark is very important for a user. Moreover, the transcripts of the movies give us the possibility to reflect the real (as opposed to normative) Russian accentological system.

3.1.5. Sociological Annotation

The sociological annotation includes the data relating to a speaker (his age, sex, occupation, and if a speaker is an actor, then his name). Strictly speaking, the main features of the sociological annotation coincide with the main traits of the metatextual annotation. However, as long as spoken monologues are very seldom, in spoken subcorpus the sociological information ought to be attached not only to the text as a whole, but to every cue of a text.

This task being fulfilled, the special program multiplies the sociological characteristics of a cue and assigns it to every token. Therefore, a user can formulate his morphological, semantic, lexical, accentological queries taking into account this or that sociological characteristics of a speaker (e.g. there is the possibility to form the subcorpora of masculine/feminine cues, of the speakers of certain age, of a certain actor, etc.; it is also possible to search this or that token/lexeme, morphological/semantic/accentological feature in combination with the sociological characteristics of a speaker).

3.2 Automatic, Semi-automatic and Manual Annotation in MURCO

3.2.1. Automatic Annotation

The annotation process in the MURCO may be automatic, semi-automatic and manual. Automatic annotation is provided with the corresponding parser. In the RNC and in the MURCO the morphological parser and semantic annotator are fully automatic.

The ***orthoepic*** annotation in the MURCO will also be fully automatic. We may automatically annotate the combinations of consonants and vowels within the word limits and at the word boundaries. The morphophonemic type of the Russian orthography gives us the possibility to pass on from the orthographical combination of the letters to the orthoepic combinations of the sounds. Therefore, we may analyze the history and contemporary situation as for the Russian pronunciation. For example, we may firstly obtain all word combinations, which include the letter combinations [d#l] and [t#l] (# means word boundary). Then, listening to the corresponding clips we may analyze the manner of the pronunciation of this letter combination. The obtained result seems to be very interesting: in the combination 'empty word + full word' the difference between voiced [d] and voiceless [t] persists, i.e. this word combination functions as one word; in the combination of two full words the voiced [d] sounds as the voiceless [t], i.e. the word boundary # functions here as a voiceless consonant; as for the word combination *vr'ad li* 'scarcely, hardly', its sounding ([dl] or [tl]) depends on the place and the date of the speaker's birth.

The Spoken Subcorpus of the RNC is partially accentuated (namely, in the movie transcripts, which form the considerable part of the Spoken Subcorpus, the stressed syllables are marked). Therefore, we may in automatic mode annotate the accentological structure of a word, e.g. we may mark first, second and so on pretonic vowels, first, second and so on post-tonic vowels, quantity of syllables, quality and number of a stressed vowel. It means that in the MURCO we may receive a set of the clips which fit our accentological query. For example, we may receive the clips which may illustrate different types of the vowel reduction in the second pretonic syllable in Russian.

3.2.2. Semi-automatic Annotation

The accentological and the sociological annotations in the RNC and in the MURCO are semi-automatic. To mark the stressed vowels, the spoken texts have been processed with the special program and after that they are corrected manually according to the real pronunciation. To mark the sociological characteristics of the spoken texts, they have been tagged manually and after that they are processed with the special program, so that the input markup of a cue is assigned to every token.

3.2.3. Manual Annotation

It is obvious enough, that we have no possibility to annotate the speech acts and the gestures in the movie clips automatically or in semi-automatic mode. One of the reasons for that (to say nothing of all technical difficulties) is the fact that to elaborate automatic or at least semi-automatic tagging of the speech acts and the gestures we need to have a test corpus to train a speech act or gesture tagger. So, we face the circularity: to obtain an automatic annotator we need a corpus, to obtain a corpus we need an automatic annotator.

Therefore, to annotate the speech acts and the gestures in the MURCO we may use the manual mode of annotation only. Maybe in future the MURCO will become one of the possible sources to create the sought-for speech act parser or gesture tagger.

It is well known that the main shortcoming of any manual annotation is the inability to provide the uniformity and commonality of the markup. In addition, the manual annotation includes a lot of chores which may be automat-

ed. These two circumstances cause the necessity to create the special workbenches for the annotators to make the process of the annotation the easiest one and the result of this process essentially normalized one. The workbenches "Marker" (the workbench to annotate speech acts) and "GesturesMarker" (the workbench to annotate gestures) offer an annotator the possibility to move from point to point answering the questions and selecting this or that variant among the displayed ones (the detailed description of both workbenches see in (Kudinov & Grishina, 2009)). In conclusion of the section we may summarize the stated above. The types of annotation in the MURCO may be characterized like this:

Method of annotation ***Assigned***	**Automatic (obligatory)**	**Semi-automatic (obligatory)**	**Manual (selected)**
to text	metatextual annotation	–	–
to word	morphological, semantic, orthoepic annotation, annotation of accentological word structure	sociological, accentological annotation	–
to clixt (text+clip)	metatextual annotation	–	speech act and gesture annotation

Table 3.

4. MURCO Interface

4.1 Orthoepic Queries

The orthoepic annotation in the MURCO is founded on the morphophonemic principle of the Russian orthography, which means that there are quite transparent correspondence between the word spelling and the word pronunciation. Therefore, we receive the possibility to annotate the combinations of letters to obtain the pronunciation of the correspondent sounds.

The crucial types of sound combinations in Russian are as follows:

- C…C = combination of two or more consonants within the word limits
- V…V = combination of two or more vowels within the word limits
- C…C#C…C = combination of consonants at the word boundaries
- V…V#V…V = combination of vowels at the word boundaries
- C…C#V…V = combination of the consonants before the vowels at the word boundaries

Obviously, it is quite easy to annotate such combinations of letters in a text automatically. Consequently, to any tokens in the MURCO will be assigned the set of the letter combinations, which may present a difficulty. Naturally, all these combinations suppose to become searchable.

4.2 Queries on Accentological Word Structure

It is well known that the dynamic quality of Russian stress leads to the great degree of the reduction of the unstressed syllables in a word. Consequently, it is very important to give a user an opportunity to obtain information, concerning the position of the stressed syllable and the quality of the stressed vowel, the position/quality of the pre- and post-tonic vowels, and so on. Owing to the fact that the majority of the clixts in the MURCO are accentuated, it is possible to annotate the accentological structure of any token in automatic mode. The content of the possible requests is defined in line with the Table 4:

	quality of vowel	**number of syllable**
stressed vowel	A	1
pre-tonic vowel	B	2
post-tonic vowel	C	3
	quantity	
syllables	4	

Table 4.

In the table cells A–C a user may specify the letter designation of a vowel (in the stressed, pre- and post-tonic syllable), in the cells 1–3 – the number of the corresponding syllable, in the cell 4 – the quantity of the syllables in a word. All these parameters are independent, so a user can freely combine them if necessary. For example, a user may request all tokens containing 1) the second post-tonic syllable, 2) the stressed syllable *o*, 3) three syllables, 4) vowel *o* in the second pre-tonic syllable, while a token has 4 syllables and the stressed vowel *o*.

All these parameters are very important for the phoneticians, specialists in orthoepy, dialectologists, and investigators in the area of the history of Russian. In addition, the importance of orthoepic and accentological annotation can scarcely be overestimated, having in mind the professional interests of the teachers of Russian, uppermost as a foreign language.

4.3 Speech Act Queries

4.3.1 Sociolinguistic Characteristics of Clixt

1. ***Quantity of participants*** (1, 2, 3, many). We distinguish clixts with one, two, three and many participants. Since we describe a clixt from the point of view of speech, "a participant" here means "a speaker". Therefore, if one of the characters of a clixts is silent (even if this character is gesticulating), this character is not considered as a participant of this clixt. The physiological activities (see below) are not regarded as speech specimens, so if a character in a clixts only sighs, spits, groans, and so on, this character is not considered as a participant of this clixt.

2. ***Participants' sex*** (Mas, Fem, Mixed). So, there are three possibilities here: Male (all the participants of a clixt

are of male sex), Female (all the participants are women/girls), Mixed (there are men and women in a clixt).
3. ***Language*** (Russian, Russian with accent, Foreign (Ukrainian, English, and so on), Quasi, Secret… the list is open). Naturally, the main language used in the MURCO is Russian. But also there are a lot of inclusions of foreign languages, which ought to be marked. It should be also noted that an annotator has the possibility to mark up the occurrences of "Russian with accent" (for instance, south-Russian dialect, north-Russian dialect, uncertain Russian dialect). Also an annotator may mark up the usage of a Quasi-Language (the participants of a clixts speak non-existent language) and a Secret Language (the participants of a clixts speak a secret language, which is familiar to them, but is incomprehensible to the profane; this secret language may be generated from the natural Russian according to the definite set of the rules or may be a kind of argot or social/professional slang).
4. ***Social situation*** (Telephone call, Dinner speech, Talk with authorities… the list is open). The main social situation, which is marked up in the MURCO, is "non-specific situation". It means that the participants of a clixt are connected with the non-official or private relations. If the relationships between the clixt participants are official and public, the fact is specially marked. Among others, we tick off Telephone calls, Dinner speeches, Talks with authorities, Shop talks, Restaurant and Taxi orders, and so on. Bearing in mind, that the annotation of the kind may be combined also with the gesture annotation, it gives us the opportunity to analyze the special social and gesture formulas, which are specific for this or that social situation.

4.3.1 Intensional Characteristics of Clixt

1. ***The types of the speech acts.*** The basic principle of the meaningful characteristics of a speech act in the MURCO is founded on 2 hypotheses: A) in the process of everyday communication a native speaker easily distinguishes one speech act from the other, otherwise the communication between the members of a speech community must fail; B) the main types of speech acts are embodied in the speech verbs of this or that language. These hypotheses, being adopted, let us build the faceted classification of the Russian speech acts, which basically addresses not the linguistic investigations concerning the different types of speech acts, but the natural linguistic intuition of an annotator and the experience of previous language usage, which has been engraved in the language itself. Naturally, this decision has a lot of drawbacks (and the most serious of them seems to be the unavoidable subjectivity of the annotation), but there seems to be no other choice. The striving to stick to the pure scientific and logical methods in the field of the speech act definitions leads us to the following risks: a) the impossibility to carry out any speech act annotation of the MURCO at all for lack of generally accepted scientific classification of the Russian (English, French, German, and so on) speech acts (let alone the fact that to create the classification of the kind we need the missing corpus with the manually annotated speech acts, so we face the circularity again); B) suppose we manage to elaborate the wanting speech act classification based on the pure logical and scientific grounds; may we be sure that this classification would be taken as equally logical by an annotator? We do not think so, because it is obvious enough that in the framework of the humanities the classification, which seems to be quite logical and objective to one person, is interpreted as absolutely subjective by the others. Therefore, it is far more preferable to rely upon and give credence to one's native language and one's everyday speech activities. In this paper we have no possibility to describe the speech act system of the MURCO in detail (see (Grishina, 2009b), where the interface of the MURCO is outlined), but we ought to mention that the list of the Russian speech acts includes about 150 items, grouped into 13 types (*Address or call, Agreement, Assertion, Citation, Complimentary, Critical utterance, Etiquette formula, Imperative, Joke, Modal utterance or performative, Negation, Question, Trade utterance*). The majority of these 150 speech acts corresponds to the Russian locutionary verbs, but there are the speech acts lacking the corresponding locutionary verbs, for instance, different types of questions (open, closed, indirect, critical, feedback), some types of negations (alienation), some etiquette formulas (Not at all!, etiquette modesty), and so on. This lack of correlations, however, does not change the main principles of the definition of the speech acts in the MURCO. To every clixt may be attached more than one type of speech act, and moreover, every speech act in a clixt may be characterized from different points of view (e.g., an assertion may be characterized at the same time as information, declaration, statement). Thus, the classification of speech act is not tree-like, it is faceted.
2. ***The completeness of an utterance.*** This markup zone gives an annotator a possibility to define the types of utterance breakings. On default an utterance is marked up as full one. The types of breakings are as follows: A) self-interruption – a speaker breaks his utterance under the influence of his own change of speech strategy; B) interruption – a speaker breaks his utterances under the influence of some external circumstance (for instance, a listener interrupts a speaker); C) unfinished utterance – a speaker has not intended to finish his utterance, for example, if its completion is absolutely predictable; D) gesture instead of word – the variant of the previous item: an utterances is finished with a gesture, not words; E) continued utterance – the variant of the item C: a speaker invites a listener to finish a speaker's utterance; F) question without answer – an unaccomplished question-answer complex; G) overlapping cues – the situation, when two or more cues are uttered simultaneously, so it is difficult to make them out.
3. ***The types of repetitions.*** It is widely known that the repetitions in the spoken speech are of great importance and go far beyond meaning transference. Within this annotation zone it is possible to mark up: A) the one-word /many-word/single/multiple repetitions; B) repetitions with intensifiers (*very, never, often, always, absolutely*

and so on); C) repetitions of the same text with different intonation; D) repetitions with the change of addressee (a speaker repeats the same text, addressing to different persons); E) repetitions during the overinterrogations – a) repetitions in answers: *I'm going to C h i t a. – Where? – C h i t a.* b) repetitions in questions: *I'm going t o C h i t a. – To C h i t a? – Yes.*; F) echo repetitions – a listener repeats a speaker's cue or its part with the same intonation; G) mimicking – a listener mimics a speaker's cue with the special mimicking intonation; H) envelope repetitions – the repetitions of a word at the beginning and at the end of a elementary discursive unit (EDU); I) relay repetitions – the repetitions of a word at the end of the previous EDU and at the beginning of the following EDU; J) simultaneous speaking – a cue or its part is uttered by two or more speakers at the same time; K) redirection of question – one person questions another, and this questioned person redirects the same question to the third person; L) imitation – a listener tries to imitate the speech behaviour of a speaker.

4. ***The manner of phonation.*** In this zone an annotator marks up different types of phonation and pronunciation of a cue. The types of phonation/pronunciation may be determined with a speaker's mental/physical state (crying, laughing, drunken, talking to oneself; articulation disorders, slip of the tongue, inarticulate cue, exercise stress, out of breath), a situation of speaking (declamation, reading, singing, dubbing-in, dictation); at this stage of annotation the special types of phonation are also marked (shout, whisper, ventriloquism, muffled shout, chanting, scanning, humming, parcelling out).

5. ***The vocal gestures, interjections and physiological activities.*** In this zone an annotator marks up: A) the interjection, i.e. the non-verbal words, which have the standard written forms (for instance, *Oh* (meaning agitation, admiration, pity, mockery, distrust, and so on), *Ah* (meaning understanding, pain, fright, reply to address, scorn, and so on), *Uh huh* (meaning approval, agreement, backing-yes), and so on); B) the vocal gestures, i.e. non-verbal words, which lack the special written forms (for instance, iconic sounds, teasing sounds, feeling cold, intensity of feeling, and so on); C) physiological activities, i.e. a speaker's or a listener's physiological acts, for instance, sigh, cough, yawn, chuckle, whistle, spit, kiss, and so on. In fact, the deeply annotated part of the MURCO lets us investigate these important linguistic phenomena on a new level.

4.4 Gesture Queries

During last three decades the investigation of the role of the gesticulation in different languages has progressed to a large degree. Now it is the current opinion that it is time to elaborate the gesture corpora to base the investigation of the gesture systems on a hard ground (see the materials of LREC'2008 and their review and the main bibliography in (Grishina, 2009b)).

The MURCO seems to be the resource, which is generally accessible and quite considerable as for its volume, moreover, the MURCO is planned to include a lot of video tracks. So, it is absolutely necessary to provide a user with the annotation and interface concerning Russian gesticulation.

The basic principles and ideological grounds for our gesture classification we gave described earlier (Grishina, 2009b). So, in this paper we list the main items of the MURCO interface, concerning the gesticulation subject matter.

4.4.1 Sociolinguistic Characteristics of Gesture

1) ***The name of an actor*** (if it is known).

2) ***The sex of an actor*** (Male, Female).

3) ***The sex of a character*** (Male; Female; Unknown (for example, in the animated films); Men, playing female role (for example, John Travolta in *Hairspray*); Woman, playing male role (this is practically impossible); Men pretending to be a woman (for example, Dustin Hoffman in *Tootsie*); Woman pretending to be a man (for example, Julia Andrews in *Victor Victoria*)). It is obvious that the last 4 items are very important for the investigation of the gender specificity of the gesticulation.

4) ***The actor's age and the character's age*** (Child, Teenager, Adult, Aged, Unknown).

It should be mentioned that any specific social situation, in which the gesticulation takes place, ought to be marked up while annotating a clixt, so there is no necessity to mark it up once more the gestures being marked up.

4.4.2 Involved Objects

The gesticulation often enough supposes the object usage. This fact, naturally, ought to be mentioned while marking up this or that gesture. The objects in question may play three main roles.

1) ***The substitutes.*** These are the objects, which substitute any gesticulating human organ (for example, a pointer or a pencil instead of a speaker's forefinger in a deictic gesture *to show with a forefinger*).

2) ***The spoilers.*** These are the objects, which impede a gesticulating person and prevent him from pure gesticulating (for example, some clothes in the speaker's hand, which spoil *a greeting handshake*).

3) ***The accessories.*** These are the whole set of the objects, which are involved in the gesticulation (the substitute, the spoilers, and the adaptors). The latter are the objects, which act as the necessary components of this or that gesture and at the same time are not the part of the human body (for example, a watch is the adaptor for the gesture *to check time*, a surface is the adaptor for the gesture *to bang one's fist on smth*).

4.4.3 Repetition Factor

In the MURCO single and multiple gestures are distinguished (single gestures are labelled with perfectives, multiple ones with imperfectives).

4.4.4 Active Organ

The active organs of the gestures are distributed into 6 groups according to the main organs of the human body.

1. Main organ: ***head*** (brow, brows, chin, ear, eye, eyes, face, forehead, head, lips, lower lip, mouth, nose, tongue, upper lip, upper teeth)

2. Main organ: ***body*** (body, shoulder, shoulders, back)
3. Main organ: ***arm*** (arm, fingers, forefinger, forefinger+long finger, forefinger+long finger+fourth finger, forefinger+long finger+thumb, forefinger+thumb, fourth finger, hand, little finger, long finger, thumb)
4. Main organ: ***arms*** (arms, hands, forefingers, fingers)
5. Main organ: ***leg*** (foot, shin)
6. Main organ: ***legs*** (feet, legs)

4.4.5 Passive Organ

The set of the passive organs is specific for this or that active organ. The basic passive organs are as follows: No passive organ, arm, arms, back, body, breast/stomach, chin, eat, eye, face, fingers, hair, hand, head, hip, hips, lips, lower lip, mouth, neck, nose, shoulder, throat.

4.4.6 Adaptor

Adaptor is the object, which is the necessary component of this or that gesture, but is not one of the organs of human body. The main types of adaptors are as follows: No adaptor, cloth, earth, external object, glasses, gloves, handset, headwear, heavy object, interlocutor, piece of furniture, pocket, sky, surface, tableware, tie, vessels, watch, wristlet.

4.4.7 Dimensional Characteristics of Gesture

1. ***Palm orientation:*** up, down, one opposite the other, to speaker's body, outside, perpendicularly to speaker's body
2. ***Direction of movement:*** backward, differently directed, does not matter, downwards, forward, forward-backward, from right to left, from the outside to the center, from within outside, horizontal circle, on its axis, outside, to oneself, to the center, upwards, vertical circle.

4.4.8 Gesture Meanings and Gesture Types

Till the moment we have marked out about 250 ***gesture meanings***, which are grouped into 14 ***gesture types***. The gesture types are as follows:

- *Adopted, Conventional, Corporate, Critical, Decorative, Deictic, Etiquette, Gestures – speech acts, Gestures of inner state, Iconic, Physiological, Regulating, Rhetorical, Searching*

Every type includes some gesture meaning. For example, some of the etiquette gestures are as follows:

- gratitude (*to applaud, to move one's head forward, twice-repeated kiss, to close one's eyes, to nod, to touch smb, to bow, to touch smb's hand, to kiss smb, to kiss smb's hand, press one's hands to one's breast*, and so on)
- apology (*to beat one's breast, to nod, to move one's chin outside, to press smb's hand to one's breast, to press one's hand to one's breast*)
- invitation (*to nod, to show smth with one's hand, to bow*), and so on.

So, the meaning of a gesture is described as a combination of 3 parameters: 1) its contextual meaning in this or that consituation, represented in a clip/clixt, 2) the type of task which is fulfilled with the gesture (=the gesture type), and 3) the traditional Russian name of the gesture (=the gesture name). The latter may be lacking, and in this case we ought to invent the missed name.

5. Conclusion

Thus we can see that the MURCO considerably extends searching possibilities up about the characteristics of spoken Russian. We may illustrate the fact with the queries, applying to the Russian greeting formulas (GF) (see Table 5).

Corpus / **Types of queries**	**Spoken Subcorpus of RNC**	**MURCO**
1. **Lexical queries**: the retrieve of the specific lexemes, used in GF (e.g. *zdravstvujte* 'how do you do?', *privet* 'hi!', and so on)	+	+
2. **Morphological queries**: the retrieve of the specific morphological characteristics of the GF lexemes (e.g. *zdravstvujte* (Pl or courtesy) vs *zdravstvuj* (Sg), *privet* (Noun) vs *privetstvuju* (Verb), and so on)	+	+
3. **Sociological queries:** the forming of the gender and chronological subcorpora to investigate the peculiarities of the GP usage	+	+
4. **Semantic & speech act queries:** the retrieve of all Russian GP simultaneously	–	+
5. **Orthoepic/accentological queries:** the retrieve of the types of the vowel contractions and the shortening of the consonant groups in the GF; the reduction of the pre- and post-tonic vowels in GF	–	+
6. **Speech act queries:** the retrieve of the types of repetitions, used in GF; the types of vocal gestures and interjections, accompanying the different types of GF; GF, used in the man/woman dialogues; see also the item 4	–	+
7. **Gesture queries:** the retrieve of the gestures, accompanying Russian GF	–	+

Table 5.

6. Acknowledgements

The work of the MURCO group is supported by the program "Genesis and Interaction of Social, Cultural and Language Communities" of the Russian Academy of Sciences. The author's investigation is supported by the RFBR[2] (RFFI) under the grant 08-06-00371a and the grant "Elaboration of Multimodal Russian Corpus (MURCO) within the framework of Russian National Corpus (www.ruscorpora.ru)".

[2] The Russian Fund of Basic Researches.

7. References

Blache, Ph., et al. (2008). Creating and exploiting multimodal annotated corpora In *LREC'2008*.

Brutti, A., et al. (2008). WOZ Acoustic Data Collection for Interactive TV. In *LREC'2008*.

Cullen, Ch., et al. (2008). Emotional Speech Corpus Construction, Annotation and Distribution. In *LREC'2008*.

Devillers, L., Martin, J.-C. (2008). Coding Emotional Events in Audiovisual Corpora. In *LREC'2008*.

Fék, M., et al. (2008). Multimodal Spontaneous Expressive Speech Corpus for Hungarian. In *LREC'2008*.

Forbes-Riley, K., et al. (2008). Uncertainty Corpus: Resource to Study User Affect in Complex Spoken Dialogue Systems. In *LREC'2008*.

Forchini, P. (2009). Spontaneity reloaded: American face-to-face and movie conversation compared. In *Corpus Linguistics 2009. Abstracts The 5th Corpus Linguistics Conference, 20-23 July 2009, Liverpool, p. 118*.

Gallo, C.G., et al. (2008). Production In A Multimodal Corpus: How Speakers Communicate Complex Actions. In *LREC'2008*.

Georgila, K., et al. (2008). A Fully Annotated Corpus for Studying the Effect of Cognitive Ageing on Users' Interactions with Spoken Dialogue Systems. In *LREC'2008*.

Gnjatovíc, M., Rösner, D. (2008). On the Role of the NIMITEK Corpus in Developing an Emotion Adaptive Spoken Dialogue System. In *LREC'2008*.

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.
At: http://docs.google.com/View?id=df52fjjj_3wd9mcrdg

Grishina, E. (2007a). O markerah razgovornoj rechi (predvaritel'noje issledovanije podkorpusa kino v Nacional'nom korpuse russkogo jazyka). In *Kompjuternaja lingvistika i intellektual'nyje tehnologii. Trudy mezhdunarodnoj konferencii "Dialog 2007"*. Moscow, RSGU, pp. 147-156.
At:
http://www.dialog-21.ru/dialog2007/materials/html/22.htm.

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.
At:
http://docs.google.com/Doc?docid=df52fjjj_11fmxsdzdh&hl=en.

Grishina, E. (2008). National'nyj corpus russkogo jazyka kak istochnik svedenij ob ustnoj rechi. *Rechevyje tekhnologii*, 3, pp. 50-62.
At: http://docs.google.com/View?id=df52fjjj_34g9d9w2dg

Grishina, E. (2009a) Korpus "Istorija russkogo udarenija". In *RNC'2009*, pp. 150-174.
At: http://docs.google.com/View?id=df52fjjj_37ghmg36cb

Grishina, E. (2009b). Multimedijnyj corpus russkogo jazyka (MURCO): problemy annotacii. In *RNC'2009*, pp. 175-214.
At: http://docs.google.com/View?id=df52fjjj_363wxt76dk

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., Savchuk, S. (2009). Ustnyj korpus v Nacional'nom korpuse russkogo jazyka: sostav i struktura. In *RNC'2009*, pp. 129-149.
At: http://docs.google.com/View?id=df52fjjj_39gh8wsffv

Knight, D., Tennent, P. (2008). Introducing DRS (The Digital Replay System): A tool for the future of Corpus Linguistic research and analysis. In *LREC'2008*.

Kostoulas, T., et al. (2008). A Real-World Emotional Speech Corpus for Modern Greek. In *LREC'2008*.

Kudinov, M., Grishina, E. (2009) Insrumenty poluavtomaticheskoj razmetki dl'a Mul'timedojnogo russkogo korpusa (MURCO). In: *Kop'juternaja lingvistika i intellektual'nyje tekhnologii (Mezhdunarodnaja konferencija "Dialog 20009", 8(15)). Computational Linguistics and Intellectual Technologies (Annual International Conference "Dialogue 2009", 8(15))*. Moscow: RSGU, pp. 249-261.
At: http://www.dialog-21.ru/dialog2009/materials/html/40.htm

LREC'2008. (2008). 6th International Conference on Language Resources and Evaluation. Marrakesh: ELRA.
At: http://www.lrec-conf.org/proceedings/lrec2008/

Marasek, K., Gubrynowicz, R. (2008). Design and Data Collection for Spoken Polish Dialogs Database. In *LREC'2008*.

Möller, S., et al. (2008). Corpus Analysis of Spoken Smart-Home Interactions with Older Users. *In LREC'2008*.

Nallasamy, U., et al. (2008). NineOneOne: Recognizing and Classifying Speech for Handling Minority Language Emergency Calls. In *LREC'2008*.

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.

Sainz, I., et al. (2008). Subjective evaluation of an emotional speech database for Basque. In *LREC'2008*.

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.

Savino, M., et al. (2008). Integrating Audio and Visual Information for Modelling Communicative Behaviours Perceived as Different. In *LREC'2008*.

Stoia, L., et al. (2008). SCARE: A Situated Corpus with Annotated Referring Expressions. *In LREC'2008*.

Strauß, P.-M., et al. (2008). The PIT Corpus of German Multi-Party Dialogues. In *LREC'2008*.

van Son, R.J.J.H., et al. (2008). The IFADV corpus: A free dialog video corpus. *In LREC'2008*.

Wilks, Y., et al. (2008). Dialogue, Speech and Images: The Companions Project Data Set. In *LREC'2008*.

Wilson, Th. (2008). Annotating Subjective Content in Meetings. In *LREC'2008*.