

Text Navigators in Spoken Russian

Abstract. The paper describes the set of the linguistic units, which serve as text navigators in spoken Russian. These units help a listener to orient himself in the whole of a spoken Russian text and to make up the impossibility to accomplish the time return in the oral speech. The investigation has been realized on the evidence of the Spoken Sub-corpus of the Russian National Corpus (RNC) – to be more precise, on the evidence of the Movie Sub-corpus of the RNC. We have used the method of word-to-word comparison between a movie transcripts and corresponding scenario or drama. This comparison let us form the set of typical discrepancies between spoken and written Russian. The main part of this set is the collection of the Text Navigators.

Keywords: Russian National Corpus, Spoken Russian, Text Navigators, Spoken Text Structure

1 What is RNC?

The Russian National Corpus (RNC) is a collection of Russian texts of the 18–21 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 centres and universities of Russia. The work of the RNC group is supported by the Russian Academy of Sciences and by the scientific Russian foundations. The RNC has been functioning since April 2003 and is accessible at www.ruscorpora.ru. Just now it contains circa 140 million tokens. In the next two years we plan to bring its capacity up to 200 million tokens¹.

The RNC mainly consists of the written Russian texts. The Spoken Sub-corpus is nothing but 5 million tokens, the Movie Sub-corpus – circa 1 million tokens². But small as they are, these corpora are sure utterly helpful for linguists. The existent collections of spoken Russian mainly are not representative, so, turning to them, we can never distinguish a rule and an exception to it. Therefore, our Corpus, which volume is estimated at millions, not thousands, in spite of its weak points³, gives us the possibility to get the reliable statistic data in different linguistic fields – lexicon, morphology, semantics, linguistic history, word-formation⁴.

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

² We hope that in next two years these two corpora grow to a considerable degree.

³ The Spoken Sub-corpus of the RNC lacks phonetic constituent.

⁴ The basic principles of organization and the whole structure of the Spoken sub-corpus of the RNC are described in [X, 2005] and [X, 2006].

2 Movie Sub-corpus and its Linguistic Significance

As for the Movie Sub-corpus (MSC) of the RNC, we have no information, whether its analogues are in existence anywhere in the world. When we began to build the MSC, we had no ideas – what kind of speech comes to be in the movies? Is it written-to-be-spoken oral speech? Is it real spoken Russian? What kind of regularities we are expected to find in movie transcripts (**MT**) – specific for written speech, or for spoken one?

Now we can answer some of the questions with certainty. The movie speech is the *imitation* of real spoken Russian, and this imitation comes to existence as a result of the specific process of *appropriation* of the alien text by an actor. The appropriation of the alien text leads to the emergence of the set of markers of spoken Russian in it, so an audience of a movie never has its doubt as to realism and naturalness of the speech in the movie. If an appropriation fails – a movie as a whole fails, and we begin to speak about falseness, “playing out of tune”, and so on.

But, if the actual state of affairs is as indicated above, then we can define the set of the markers of spoken Russian. To fulfil this task, we ought to compare a MT with the corresponding scenario or drama (source text, **ST**) to list the obtained discrepancies⁵, to try to classify them and to describe the most important types of discrepancies.

These two tasks are to be solved in this paper.

3 Two types of MT~ST Discrepancies

The analysis of the MSC data shows that the discrepancies between MT and corresponding ST can be divided into two groups – *Speaker-oriented spoken markers* and *Listener-oriented spoken markers*.

The first group is not large, because the interests of the Listener in spoken dialogue prevail over the interests of the Speaker. Therefore, the latter has the opportunity to watch over his own interests only when they don't hamper in the interests of the Listener.

3.1 Speaker-Oriented Spoken Markers

Speaker-oriented spoken markers are of two types. **The first type** gives a speaker the possibility to save his energies when he produces his speech. This type of the spoken markers mainly comes 1) to the use of apocopic forms of words, f.e. *чтоб* [chtob] instead of *чтобы* [chtoby] ‘in order to’, *хоть* [hot'] instead of *хотя* [hot'a] ‘although’, and some others, or 2) to the use of ellipsis of conjunctions or parenthesis, f.e. *если...*, *Ø...* instead of *если...*, *то...* ‘if... then...’, *может* instead of *может быть* ‘probably’.

The second type of Speaker-oriented spoken markers gives a speaker the possibility to realize the surplus of speech intonation in his speech: the emphasis in an oral

⁵ This task was fulfilled in [X, 2007].

phrase often needs some extra words for its realization, which mean nothing, but implement the movement of intonation and the energy of emphasis. In this case are regularly used 1) iterations of different kind, f.e. (1) ^I*Ах ты наглец / а?* ^{II}*Ну ты наглец!* ^{III}*Ну ты наглец!* ^{IV}*Ну наглец!* (MT) instead of *Ах ты наглец!* (ST) ‘You are really an impudent fellow!’; 2) words-explicitors – *очень* ‘very’, *всегда* ‘always’, *никогда* ‘never’ and some others, f.e. (2) *Вам здесь будет **очень** удобно* (MT) instead of *Вам здесь будет удобно* (ST) ‘Here you’ll feel [**very**] comfortable’; 3) interjections, which have not been provided in ST, but become relevant for the real speaker, f.e. (3) *Господи **боже мой** / такая удача!* (MT) instead of *Такое везение!* (ST) ‘[**Oh, boy!**] Really it’s a good luck!’

3.2 Listener-Oriented Spoken Markers

Listener-oriented spoken markers are also of two types. **The first type** is used by a speaker to make a listener to perceive speaker’s words with due attention. For this purpose are regularly used 1) contact words/forms, f.e. the second person or the imperative of the verb *понимать* ‘understand’, *знать* ‘know’, *слушать* ‘listen’, *слышать* ‘hear’: (4) *Вы **знаете** / последнее время некогда читать-то было* (MT) instead of *Последнее время мне некогда было читать* (ST) ‘[**You know**] Lately I’ve had no time to read fiction’; 2) words of address, which name the listeners of the speech and so attract their attention, f.e. (5) *Ребятам / ну / ну не надо!* (MT) instead of *Не надо!* (ST) ‘[**Boys!**] Don’t do it!’ (see also the example (20)).

The second group of Listener-oriented spoken markers is very large and various. It includes different linguistic tools, by means of which a speaker let a listener 1) understand the composition of the speech, 2) distinguish more or less important parts of it, 3) learn the types of utterances, which are included in the speech, and so on. These very sorts of the markers we name Text Navigators.

4 Types of Text Navigators

4.1 Delimitation of Phrase Boundaries

The most frequent Text Navigators delimitate the phrase beginning and the phrase end. For these purposes the particles *ну₁* and *ну₂*, *а₁* and *а₂*, *да₁* and *да₂* are used.

1. The beginning of an assertion is indicated with the particle *ну₁*. It has a lot of meanings (see [Шмелев, 2005]), but its main destination is to let a listener learn that a speaker is going to ascertain something in the forthcoming phrase, f.e. (6) *Ну / дней / может быть / десять...* (MT) instead of *Дней, может, десять...* (ST) ‘[**Well,**] Probably about ten days...’ The particle *ну₁* is normally unstressed or receives the compulsory stress, being placed between two pauses.
2. The beginning of an imperatival phrase is indicated with the particle *ну₂*, f.e. (7) *Ну / в путь!* (MT) instead of *В путь!* (ST) ‘[**Well,**] Let’s go!’ In special cases the particle *ну₂* designates the beginning of a question and serves as a demand to answer it, f.e. (8) *Ну / молодой человек / вы служить собираетесь?* (MT) instead

- of *Вы, молодой человек, собираетесь служить?* (ST) ‘[Well,] young man, are you going to join the army?’ The particle $ну_2$ is always stressed.
3. The beginning of the phrase, which contains contradistinction to the previous text or to the previous action, is marked with the particle a_1 , f.e. (9) *А дядька правильно говорит...* (MT) instead of *Дядя правильно говорит...* (ST) ‘[But] our uncle is right, saying...’ We ought to mention that a question very often originates from a contradistinction, so the particle a_1 often marks the beginning of a question, f.e. (10) *А сколько же ей стукнуло?* (MT) instead of *Сколько же ей стукнуло?* (ST) ‘[And] How old she is?’ The particle a_1 is always unstressed.
 4. The beginning of an exclamation is indicated with the particle $да_1$, which is always unstressed, f.e. (11) *Да что вы такое говорите!* (MT) instead of *Вы что такое говорите!* (ST) ‘[But] you shouldn’t say so!’ We ought to point out that an imperatival phrase, which is primarily the exclamation from the point of view of the speaker, marks its beginning with $да_1$, not $ну_2$, f.e. (12) *Да молчи ты!* (MT) instead of *Молчи!* (ST) ‘[Come on,] shut up!’
 5. The end of a phrase is marked with the particles $ну_2$, a_2 , $да_2$, which are always stressed. The particle $ну_2$ at the end of a phrase is pronounced with the imperatival or interrogative intonation and means ‘do (this or that)!’ ‘say!’ ‘are you going to answer my question?’, the particle a_2 is pronounced with the interrogative intonation and means ‘and what is *your* opinion about this?’, the particle $да_2$ also is pronounced with interrogative intonation and means ‘do you agree with me?’. Examples: (13) *Я стал бы для тебя хуже? Ну?* (MT) instead of *Стал бы я для тебя хуже...?* (ST) ‘Would you think I’m worse than you thought? [Answer!]’, (14) *Вот народец / а?* (MT) instead of *Вот народец!* (ST) ‘This entire people are so bad [eh?]’, (15) *Вы ... / ходили в студию... / да?* (MT) instead of *Вы... ходили в студию...?* (ST) ‘You have been at the drama school? [Haven’t you?]’.

4.2 Explication of Role Structure

It is widely known that in Russian the meanings of the first and the second person are expressed with the verb endings and with the personal pronouns. In written Russian in many cases it’s enough to use only the 1-2 persons of the verb (as in Latin, for example). Our investigation has shown that in spoken Russian the verbs of 1-2 pp. are very often accompanied with the corresponding pronouns, f.e. (16) *Ты плюнь мне в рожу-то / я тебя прошу как родного!* (MT) instead of *Плюй мне в рожу, как родного прошу!* (ST) ‘[You] spit into my face, [I] beg [you], be good boy!’ (see also the ex. (4), (12)).

The explanation is as follows: a speaker doubles the meaning of the grammatical person to be sure that a listener grasps the role structure of the utterance correctly. The first and the second persons take on special significance in the speech interaction, because they refer to a speaker and to a listener, so the idea of the person must be expressed utterly distinctly.

4.3 Manifestation of Latent Performative

To facilitate the understanding of the text, a speaker very often manifests the latent performative, which makes the ground of his utterance. Written texts give a reader the opportunity to return to the beginning of a phrase and so to correct possible mistakes of his perception. In a spoken text the return of this kind is impossible. Therefore, in many cases a speaker let a listener know beforehand what type of utterance is going to take place just now, f.e. request: (17) *Теперь я вас прошу / пойдите туда вы...* (MT) instead of *Теперь пойдите туда вы...* (ST) ‘Now [I beg you] go there, I mean you personally...’; warning: (18) *Я в ужасном горе. Предупреждаю сразу* (MT) instead of *Я в ужасном горе* (ST) ‘[I warn you] I’m hard hit by grief’; question: (19) *Товарищ генерал... / я давно хотел спросить / как с йети быть?* (MT) instead of *Товарищ генерал, а как же с йети быть?* (ST) ‘Comrade general, [I always want to ask] what we can do with a snowman?’

4.4 Introduction of Metatext

A speaker often intends to give notice of the fact that in forthcoming utterance he means to say something very important, so a listener ought to be attentive. For this purpose mostly are used formulas *Так вот* or *Вот что*, which are the abbreviations of the phrase *<Так> вот <что>* [я вам сейчас скажу] ‘<So> just now I am going to say you the following’, f.e. (20) *Так вот / Надежда Георгиевна / я прошу вас позвонить моей жене* (MT) instead of *Я прошу вас позвонить моей жене* (ST) ‘[So] [Nadezhda Georgievna] [listen to me] I expect you to call my wife’. We name this kind of insertions the introduction of the metatext – the formulas *так вот* and *вот что* haven’t got their own meaning: their only destination is to mark the beginning of the important utterance.

Moreover, these formulas may transform and change into *вот так*, *вот* or *вот как*. These transformation, on the contrary, mark the end of the important utterance and are the abbreviation of the phrases *Вот так* [я думаю], *Вот* [что я думаю] and *Вот как* [я думаю!] ‘I’ve expressed my opinion just now’, ‘It’s my belief’, f.e. (21) *Заслуга не в том / чтобы не совершать ошибок / а в том / чтобы вовремя их исправлять. Вот так* (MT) instead of *Заслуга не в том, чтобы не делать ошибок, а в том, чтобы вовремя их исправлять* (ST) ‘Our merit is not to be faultless, but to correct our faults in time. [It’s my belief]’.

4.5 Lexicalization and Intensification of Emphasis

In spoken Russian a phrase emphasis is expressed with the help of stress, of special intonation and of word order. These three ways to emphasize a word or a word combination are so to speak “non-material”, because they lack special material carrier. Therefore, from the speaker’s point of view these types of emphasizing are not enough reliable. In consequence, in spoken Russian an emphasis often is supported with special words and constructions.

The lexicalization of emphasis is mostly accomplished with the particle *-то*, which follows the emphasized word, f.e. (22) *А ты-то счастлива? – Я / нет* (MT) instead of *А ты счастлива? – Я? Нет* (ST) ‘[**And you,**] Are you happy? – Me? No’.

It’s well known, that a phrase emphasis very often accompanies a rhema. Therefore, in spoken Russian we have the special construction, which frames a rhema and so a listener is guaranteed against the loss of it. The construction is *что [rhema] что ли?*⁶, f.e. (23) *Вы что / не знаете / что ли?* (MT) instead of *Неужели вы не знаете?* (ST) ‘You [**really**] know nothing about it, [**don’t you**]?’

The free word order in Russian allows a speaker to use one more way to intensify a rhema – he has the possibility to move it closely to the beginning of a phrase and to give a listener the chance to grasp it ahead of the unemphasized part of the utterance, f.e. (24) *Лида **взбалмошная**⁽²⁾ **немножко**⁽¹⁾* (MT) instead of *Лида **немного**⁽¹⁾ **взбалмошная**⁽²⁾* (ST) ‘Lida is slightly⁽¹⁾ **crotchety**⁽²⁾’.

5 Conclusions

So, in the issues 3–4 we have listed the most frequent and the most important types of the discrepancies between movie transcripts and corresponding source texts (dramas and scenarios). In the Table 1 we sum up the previous account.

Table 1. Types of Spoken Markers in Russian

Markers	Speaker-oriented	Listener-oriented
Apocopic forms of words	+	
Ellipsis of conjunctions and parenthesis	+	
Iterations	+	
Words-explicators	+	
Interjections	+	
Contact words/forms		+
Words of address		+
Text navigators		+

The central position among the spoken markers belongs to the text navigators, which help a listener to grasp the meaning of an utterance adequately. The text navigators divide into three groups – the text navigators, 1) quantifying an utterance, 2) qualifying an utterance, 3) structuring an utterance, see Table 2.

⁶ Both parts of the construction are not obligatory, so it may be *что [rhema]?* or *[rhema], что ли?*

Table 2. Types of Text Navigators in Spoken Russian

Text Navigators	Quantifying	Qualifying	Structuring
Delimitating phrase boundaries	+		
Explicating role structure			+
Manifesting latent performative		+	
Introducing metatext		+	
Materialization of emphasis			+

It seems, that our investigation of the spoken markers in Russian lays the foundation of the further analysis of spoken Russian. 1) First of all, now we have the formal grounds for distinguishing spoken and written Russian – in spoken Russian the density of some words (for example, *ну, а, да, вот, что, так, -то*, and some others) and of some grammatical forms, f.e. 1–2 pp. pronouns, is noticeably higher, than in written Russian. 2) Besides, we have received the formal reference points to analyze the composition of spoken Russian texts. 3) The last, but not the least, now we may put the question of the typology of spoken markers and their transformation from one language to another.

References

1. [НКРЯ, 2005] – Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 344 с.
2. [X, 2005] – X. Устная речь в Национальном корпусе русского языка //Национальный корпус русского языка: 2003-2005. Результаты и перспективы, М., 2005, с.94-110
3. [X, 2006] – X. Spoken Russian in the Russian National Corpus (RNC) // LREC2006: 5th International Conference on Language Resources and Evaluation. ELRA, 2006. P. 121-124
4. [X, 2007] – X. О маркерах разговорной речи (предварительное исследование подкорпуса кино в Национальном корпусе русского языка) // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007» (Бекасово, 30 мая – 3 июня 2007 г.), с. 147-156
5. [Шмелев, 2005] – Шмелев А.Д., «Заполнители пауз» как коммуникативные маркеры. // Язык. Личность. Текст. М.: 2005