Pragmatic Markers in Russian Spoken Speech: an Experience of Systematization and Annotation for the Improvement of NLP Tasks

Natalia Bogdanova-Beglarian, Olga Blinova, Tatiana Sherstinova, Gregory Martynenko, Kristina Zaides
Saint Petersburg State University
Saint Petersburg, Russia
{n.bogdanova, o.blinova, t.sherstinova, g.martynenko}@spbu.ru, kristina.zaides@student.spbu.ru

Abstract—**Pragmatic markers are an integral part of spontaneous spoken speech, however, they still have no systematic scientific description. These speech elements perform mostly pragmatic functions and are characterized by almost complete absence (or significant weakening) of lexical and/or grammatical meaning. The frequency of pragmatic markers in speech exceeds that of almost all content words. Because of that, for the improvement of many current NLP tasks, it is very important to obtain proper systematization of pragmatic markers and to develop effective and reliable schemes for their annotation. In current research, we describe the preliminary set of pragmatic markers categories and present the results of two stages of their pilot annotation made independently by a group of experts.**

I. Introduction

Spoken speech is a sufficiently complex object for any automated systems of audio signal processing, due to its native (natural) uncertainty and diffuseness. Moreover, the detailed analysis of authentic spoken data made in real-life environment reveal the number of phenomena, which are extremely important for the process of oral communication, but which still remain out of view of linguists and specialists in speech technologies, because they rarely occur in usual "laboratory speech". The ignorance of these real-speech elements leads to noticeable shortcomings in work of NLP systems when they face spontaneous speech of everyday spoken discourse. Therefore, for many speech technologies systems it is important to have better understanding of the set of elements used in everyday discourse, as well as of the rules of their functioning.

Pragmatic markers should be noted first in the list of real-speech phenomena. They are speech elements that are devoid (either completely or partially) of lexical and sometimes grammatical meaning. Their main task is to perform different pragmatic functions. Pragmatic markers are an integral part of spoken speech but up to present they still have no systematic scientific description. The aim of the research presented in this paper was to propose a working scheme for systematization of pragmatic markers of Russian everyday speech, to develop annotation rules for these speech elements, and to test this rules by means of independent annotation of "live" speech recordings by several experts.

II. Studies of pragmatic markers: state-of-the-art

In this paper, by "pragmatic markers" we mean discourse units (words and multiword expressions) with a weakened referential meaning, which perform a variety of pragmatic or procedural speaker's tasks [1]. The semantics and the grammar of original forms, from which pragmatic markers derive, are replaced by their pragmatic function in spoken texts (e. g., Russian '*это самое*', '*как его* (*её/их*)', '*скажем так*', etc.).

The term "pragmatic marker" was first introduced by B. Fraser [2, 3]. According to Fraser, pragmatic markers signal important aspects of the speaker's message. They are related with "the ways in which the linguistically encoded information of sentence meaning provides an indication of the direct, literal messages intended by the speaker" [3]. Being embedded in sentences, pragmatic markers are "separate and distinct from the propositional content of the sentence" [Ibid.].

It should be mentioned that in contemporary linguistic studies several other terms have been used earlier to describe pragmatic units — "discourse particles", "discourse markers", "modal particles", etc. The most frequent term here is "discourse markers" (DMs) that are sufficiently well described in the scientific literature [4–12]. The common characteristic for discourse markers and pragmatic markers is their ability to help the speaker to structure the discourse.

Let us first consider the distinction between pragmatic markers (PMs) and discourse markers (DMs). DMs can express the speaker's conscious attitude toward the subject of speech (for example, introductory words), and PM verbalize speaker's attitude to the process of speech production, including all difficulties and hesitations in the course of this process. Further, while DMs are pronounced by the speaker consciously, being valuable units of any (oral and written) discourse and having have both lexical and grammatical meaning (*now*, *again*, *by the way*, *of course*, *probably*, *first*, etc.), PMs are pronounced unconsciously, as speech automatisms, they are absent of lexical meaning being almost completely "agrammatical". DMs are included in the dictionaries of the Russian language, and PMs are still outside of lexicographic description.

The presented research is a first approach to annotate and study PMs using methods of corpus linguistics. Though the meanings of terms PMr and DM do not fully coincide, the approaches of their corpus studies and annotation have much in common. Because of that, on the stage of elaboration of PMs annotation scheme and methods of analysis, we took into account the recent works on DMs annotation, too. Some of them are briefly described in this section below.

Among other types of annotation (morphological tagging, syntactic parsing etc.), the pragmatic one is considered to be one of the most important for speech corpora and supposes annotation of speech acts, turn-taking or pragmatic markers [13, 14]. In many cases and especially for small datasets, pragmatic annotation is performed manually since there are still no appropriate tools for its automatic analysis. Another problem of corpus pragmatics concerns the fact that most of pragmatically annotated corpora are "coded for selected aspects of pragmatic interests only" [15].

In [16], the annotation of discourse markers is implemented by the EXMARaLDA annotation tools, which allows to mark two or more functions for each marker in different contexts. It is done manually or semi-automatically based on the prescribed list of possible DM-functions.

The MDMA (Model for Discourse Marker Annotation) project is aimed at labelling discourse markers in spoken data through the manual selection of DMs by the coders and their further semantic, syntactic and pragmatic annotation; this methodology is also named "back-and-forth from theory to data" [17]. The results of this research showed that only markers in the initial position of the sentence are well identified by the algorithm based on statistical modeling (conditional trees and multifactorial analysis) [18].

L. Crible and S. Zufferey conducted the annotation of DMs both in spoken speech and in written texts for two languages—French (annotators' native language) and English [19]. To identify the DM functions, the researchers used the structure of four domains — ideational, rhetorical, sequential, and interpersonal. It should be noted that the Penn Discourse Treebank (PDTB) [20] taxonomy is considered to be not suitable for DMs, as it was designed only for written texts. It turned out that the annotation of DMs in written texts is not an easy task, either. Thus, two experiments were designed: first, the potential candidate for DMs were found, which shows that coder who was expert in written text analysis pointed more DMs in written texts than in spoken ones, and vice versa with another annotator; second, the annotation of DM functions was conducted, which reflects the inter-annotator agreement from 34% (for English texts) to 52% (for spoken French). However, there were several issues during the annotation: the first problem was the distinction between ideational and rhetorical relations, the second issue concerned the distinctions between semantically overlapping functions, such as, e.g., conclusion and reformulation, and the third source of disagreement was authors' discovery of missing functions in the taxonomy, (e.g., meaning of a "goal"). As a result, a greater precision in the criteria used to disambiguate similar functions of DM and the inclusion of additional paragraphs in the instruction to ambiguous meanings of DM are adopted.

L. Crible and M.-J. Cuenca [21] also report that most annotation models of DMs were designed for written discourse: the Rhetorical Structure Theory (RST) [22], the Penn Discourse Treebank [20], and the Cognitive approach to Cognitive Relations (CCR) [23]. On the material DisFrEn, a French-English corpus, DMs were identified without using a closed list, if the words (conjunctions (*and*, *but*, *although*), adverbs (*actually*, *well*), prepositional phrases (*in fact*, *by the way*) or verb phrases (*you know*, *I mean*, *if you will*)) "met the criteria of procedurality, syntactic optionality, fixedness (i.e. grammaticalized), discourse-level scope and metadiscursive function (discourse relation, topic structure, turn exchange, speaker-hearer relationship)" [23]. The syntactic position, co-occurrence and sense disambiguation of DM were marked. The authors discuss the challenges of discourse markers annotation, such as truncated structures in spontaneous speech, the ambiguity of language in general and of some DMs in particular, the polysemy and the multifunctionality of some markers that can perform many different functions, in one and the same or in different contexts. All such cases make the automatic annotation of DMs not very possible. Hence, the researchers suggest not looking for two "necessary" for the discourse relations arguments and not trying to match the ideal formula "one form — one function".

D. Verdonik, M. Rojc, and M. Stabej [9] analyzed DMs in the corpus of Slovenian telephone conversations TURDIS, including into the list of DMs such phenomena as hesitations, traditionally understood DMs, and background signals, such as *I see*, *right*, *okay*, etc. They also try to deal with cases of markers co-occurrence, describing the most widespread chain of markers at the beginning of an utterance. The researchers reduce all the functions of DMs to the following: signaling connections to the propositional content, building relationship between the participants in a conversation, expressing the speaker's attitude to the content of the conversation, and organizing the course of a conversation. It is noted that for most cases "it is not possible to say that a discourse marker performs only one of these pragmatic functions" [9]. The authors point out that there is "no consistency on which expressions count as discourse markers, therefore we have to reconsider how to set a framework for annotation" [9] and suggest either to do manual annotation first, so that an algorithm can be trained on this database, or automatic one first, and after manually correct it. Their material allows performing automatic annotation for marking only hesitations, but many expressions that can function as DMs, as well as the elements of the propositional content of an utterance, need to be manually checked.

The first attempt to annotate discourse markers in a multilingual parallel corpus (Arabic-Spanish-English) is provided in [24]. The researchers use PRAGMATEXT model of annotation which includes the list of pragmatic tags, such as marking emotional language, discourse relations, discourse modality, evidentiality, metaphor, speech act, and deixis. After the POS-tagging, the annotation of markers in Spanish corpus was made using the monolingual Spanish lexicon of discourse markers. These texts were compared with texts in another two languages in order to define discourse markers with a help of bilingual lexicon. If the discourse marker is non-ambiguous, it is automatically tagged, as authors state. The ambiguous markers are disambiguated with the following context rules: prosodic features reflected through the punctuation (since corpus contains written texts) and the position of occurrence of DM within the sentence (inter-sentential segment). The statistical model for automatic disambiguation of DMs is planned to develop and to verify its results in future. As the factors which prevent the creation of automatic pragmatic annotation means, the authors mention following: the lack of consensus in the classification of DMs, their categorical, syntactic, and discursive ambiguity, and the distinction between DMs and idiomatic expressions, as well as the ability of DMs to form the idiomatic expressions.

Thus, it can be seen, that pragmatic annotation of corpus data may be performed in different ways and using a variety of tools, which can facilitate the process of manual annotation. However, the fully automatic annotation, even based on the approximately closed list of pragmatic markers, is not possible to date.

III. Preliminary set and functions of pragmatic markers in spoken Russian

Examination of everyday speech transcripts leads to the conclusion that in spoken Russian PMs perform many important pragmatic functions. First of all, we should mention marking speaker's word-search hesitation (the Russian examples may be ‘*как его*’(‘*её*’, ‘*их*’), ‘*это*’, ‘*это самое*’) and the reaction (or reflection) to the result of this search (‘*скажем так*’, ‘*или как его там*?’), as well as the metalanguage commenting on the process of speech production itself (‘*ну не знаю*’; ‘*что ещё*?’, ‘*знаешь’/ знаете*)’, ‘*прикинь/ прикиньте*’). Besides, there are hedging elements (e. g., ‘*типа того*’, ‘*или как лучше сказать*’, ‘*ну там*’), markers introducing someone else's words (‘*типа того что*’, ‘*вот мол*’, ‘*грит*’, ‘*такой*’) and other types of pragmatic elements.

The preliminary observations show that several dozens of such PMs, being quite different in respect to their structure and polyfunctionality, are capable to fulfill about 10 basic functions. Moreover, the frequency of PMs in speech exceeds that of almost all content words [25]. PMs are functionally important both for the process of speech generation and for overcoming inevitable speech difficulties. They are used in particular:

1) for marking the beginning or ending of utterances,
2) for attracting interlocutors' attention,
3) as a reaction to a mistake,
4) while searching for a word or other ways to proceed talking further,
5) for introducing someone else's speech into the narrative,
6) for introducing a new thought that just came to mind,
7) for correction of certain fragments of speech, and so on.

A preliminary version of pragmatic markers typology has been proposed by Natalia Bogdanova-Beglarian in recent academic papers [26], however it requires its validation on the basis of representative speech data, which is to be done in the nearest future.

Below is an exploratory typology of pragmatic markers, that we have used for annotating Russian spoken speech in our experiment [1]:

— *Discourse pragmatic markers*, which are used for structuring spoken texts. They include starting, guiding (navigational) and final markers (e. g., ‘*значит*’, ‘*вот*’, ‘*думаю что*’, ‘*знаешь вот*’, ‘*всё*’).

— *Search pragmatic markers* (‘*это самое*’, ‘*как его (её, их)*’, ‘*как это*’, ‘*что ещё*’), which provide the speaker with some extra time to find the proper word or expression.

— *Reflexives*, or pragmatic markers that reflect the speaker's reaction to his/her own words (‘*или как его*?’, ‘*или как там правильно сказать*?’, ‘*или как там?*’).

— *Approximators* are pragmatic markers used for the replacement of some enumeration or its part. For this purpose, various “substitutes” are used, which signs that enumeration is possible or can be continued. In the first case, the markers of complete replacement indicating the result of “the substitution strategy” action are used. In the second case, the markers of partial replacement, referring to the “combining strategy” of the speaker are employed (‘*все дела*’, ‘*всё такое прочее*’, ‘*(и) то (и) другое*’, ‘*то-сё*’, ‘*туда-сюда*’, ‘*пятое-десятое*’, ‘*бла-бла-бла*’).

— *Xenoindicators*, or markers that introduce someone else’s speech into the utterance (‘*грит*’/’*гыт*’, ‘*ах*’, ‘*типа того (что)*’, ‘*такой*’, ‘*вот*’, etc.).

— *Metacommunicative pragmatic markers* are metacomments to the speech, aimed at establishing a contact with interlocutor(s) or listener(s), as well as at speech comprehension by the speaker himself (‘*знаешь(-те)*’, ‘*понимаешь(-те)*’, ‘*да*’).

— *Deictic pragmatic markers*, whose function is related primarily to the discourse unit ‘*вот*’ (‘*вот здесь вот*’, ‘*вот такое вот*’, ‘*вот так вот*’).

— *Rhythm-forming pragmatic markers* are used to rhythmize spoken text (‘*вот*’, ‘*там*’, ‘*короче*’, ‘*так*’).

— *Pragmatic markers of self-correction* are used to correct an utterance (‘*это*’, ‘*это самое*’, ‘*не*’, ‘*не так*’).

— *Markers of speech “non-triviality”* (‘*так скажем*’, ‘*как это*’, ‘*так называемый*’).

— *Hesitation markers* (‘*там*’, ‘*это*’, ‘*м-м*’, ‘*э-э*’).

— *Interjectional pragmatic markers*, which differ from the interjections (which are often the etiquette ones) because they acquire either a new semantics, or a new pragmatics or prosodic design (‘*драсьте пожалста!*’, ‘*щас-щас-щас-щас*’, ‘*будет тебе*’).

It is expected that this provisional version is liable to undergo some changes or refinements in the process of its validation on representative spoken data [1].

IV. Research data

Data from two representative speech corpora are planned to be used to study functioning of pragmatic markers in spoken Russian [27]:

1) the corpus of Russian everyday speech “One Day of Speech”, known as the ORD corpus, which contains all types of everyday spoken discourse [28], and

2) the “Balanced Annotated Text Collection /Textotec” (SAT, containing monologic speech) [29].

An exploratory annotation of PM described in this paper made on the data of the first of them.

The corpus of Russian everyday speech “One Day of Speech” (ORD), which is currently the most representative resource for the analysis of spoken Russian, provided data for compiling a preliminary list of PMs in Russian spoken speech. Based on this list, a typology of Russian PMs was proposed, which is supposed to be suitable for automatic processing of large data sets [28], [30], [31].

The ORD corpus contains 1,250 hours of sound recordings obtained from 128 informants, which are native speakers of Russian, living in St. Petersburg, and more than 1,000 of their interlocutors, which represent various social groups. At present 2800 macroepisodes of everyday speech communication are described, speech transcripts exceed 1 million of words [31], [32]. The records were made by means of the 24 h-monitoring method and were further transcribed in the ELAN linguistic annotator [33].

For the pilot annotation of PMs, it was decided to use ORD subcorpus described in [34]. The subcorpus contains fragments of everyday communication of 12 respondents and their 10 interlocutors, lasting 1h 46 min in total. The episodes for this subcorpus have been selected from everyday (non-professional) conversations of respondents with their relatives, friends or colleagues. The amount of the speech material is comparable to a well-known spoken Russian corpus “Rasskazy o snovideniyax” (Night Dream Stories) [35].

The subcorpus contains episodes of speech communication of the balanced sample of respondents, representing two gender groups (6 men and 6 women), three age groups (four representatives for each – youth, middle-aged and seniors – group) and at least one representative of 4 social class groups: (a) high-level personnel, businessmen and self-employed individuals, (b) salaried employees, (c) students (including those who works), (d) unoccupied people, including non-working pensioners. Besides, the subcorpus contains speech of representatives of the following professions: 1) a worker, 2) a soldier, 3) an engineer, 4) an IT-specialist, 5) a teacher, 6) a physicist, 7) an art historian, 8) a marketer, 9) a lawyer, and 10) a musician [34].

In total, the research subcorpus contains 16060 tokens of speech transcripts, 10259 of which are words.

V. Development of annotation scheme

In order to elaborate an optimal approach to pragmatic markers annotation, two stages of trial annotation were carried out. In both cases, four experts took part in annotation of the same speech data. All experts have considerable experience in transcribing and analyzing sounding speech. It was considered important to provide independence of experts' estimations so that it would become possible to compare the results and to measure their consistency.

All annotations were made in ELAN, which provide immediate access to sounding of speech transcripts. The ORD transcribing conventions were described in [28]. For PM annotation, four additional tires were added to the ORD annotation files:

Tier 1. ***PM*** — On this tier PMs are given in the same form as they are presented in the speech transcripts. Here, it is obligatory to use standard spelling of words, despite their possible phonetic variation. For example, the reduced forms should be written in their full forms (e. g., not ‘*щас*’, but ‘*сейчас*’); the hyphens in repetitions and particles like ‘*-то*’, ‘*-ка*’ should not be put (e. g., ‘*сейчас сейчас сейчас сейчас*’, ‘*тут то*’, ‘*иди ка*’); all auxiliary symbols like the lengthening sign should be removed from this level (e. g., not ‘*да*(:)’, but ‘*да*’). However, a hyphen in indefinite pronouns with ‘*-то*’, ‘*-либо*’, ‘*-нибудь*’, ‘*кое-*’ (‘*кто-то*’, ‘*кому-либо*’, ‘*когда-нибудь*’, ‘*кое-что*’) should remain.

Tier 2. ***PM Function*** — This tier was introduced for annotating both the main and additional functions of pragmatic markers. All appropriate functions should be listed in a single tag, without spaces, and in alphabetical order. If it was impossible to distinguish the main function among others, the list of all relevant functions should be given (e. g., ‘*PX*’ = rhythmizing and hesitation (see below for the details).

Tier 3. ***PM Speaker*** — This is a tier, on which standard codes of speakers, participated in the conversation, should be given. This option allows to select automatically the PMs, pronounced by various speakers or groups of speakers.

Tier 4. ***PM Comment*** — The Comment-Tier is very useful for commenting PM actual usage, such as real phonetic pronouncing of the word, its prosody features, as well as various other characteristics of PM occurrence in speech.

A detailed step-by-step instruction for annotators was compiled. The experts performed the annotation of the same files independently of each other. The list of PM functions and the list of tags for them were slightly revised after each stage of annotation.

A. The first pilot annotation

For the first pilot annotation, an expanded list of pragmatic marker functions was proposed (see below).

Annotators were asked: 1) to mark all PMs which occur in recordings, spelling them in standard form on the PM-Tier and indicating the actual borders of PM in oscillogram, 2) to determine the main function of PM and to place the corresponding tag in the first place on the annotation on Function PM tier, while the remaining or additional functions should be listed further in the alphabetical order, 3) to give the code of correspondent speaker, and 4) to fill the Comment-tier when applicable.

Here is the list of 14 basic PM-functions used in the first stage of annotation:

1) APPR — approximator marker;

2) DEICT — deictic marker;

3) ZAMEST-PR — replacement marker for the whole set or its part;

4) ZAMEST-CHR — replacement marker for someone's speech, e. g., '*бла-бла-бла*';

5) XEN — quotational marker;

6) MET — meta-communicative marker;

7) NAVIG — navigational marker;

8) SEARCH — searching marker;

9) REFL — reflexive marker;

10) RHYTHM — rhythm-forming marker;

11) SELFCORR — marker of self-correction;

12) START — starting marker;

13) FIN — final marker;

14) HES — hesitation marker.

Besides, the annotators had the opportunity to assign to PMs some new functions when necessary and to express their uncertainty about marking a particular PM at the comment level. Besides, they were invited to describe phonetic features of PM with the help of special symbols.

Thus, on the *Comment PM* level, some additional features concerning the use of PMs could be noted. For example, phonetic reduction of the form (e. g., '*слышишь*' → '*слыш*', or '*говорят*' → '*грят*'), some special intonation form of PMs, and some pragmatic issues (e. g., the rhetorical function of the marker, expressed in the hypercorrection of the speaker, or speaker's desire to decorate his/her own speech, like in '*собственно говоря*').

On the Fig. 1 one may see the example of PM annotation in ELAN, and the Table 1 presents a fragment of annotation database, which was obtained by exporting these data from ELAN to MS Access.

The pilot annotation allowed to replenish the preliminary list of PMs by some new pragmatic units that had stayed out of the researchers' sight, such as: '*или ещё что-то такое*' (reflexive marker, marker-approximator), '*минуточку-минуточку*' (hesitation marker), '*значит так*' (navigational marker), '*то то*' (e. g, in the phrase like '*я говорю / за это вы мне поднимите* () *там вот это / то то*' – replacement marker), etc.

In the result of the first annotation state, it turned out that the best agreement between experts was achieved for pragmatic markers of the following types: *Meta-communicative pragmatic markers* ('*понимаешь*', '*знаешь*', '*слушай*', etc.), *Xenoindicators* ('*говорю*', '*типа*', '*такая*', etc.), and *Reflexives* ('*так сказать*'). Rather good results were showed when attributing *Approximators* ('*как бы*', '*типа*', etc.), too. However, there were less consent when annotating PMs, which may perform in speech several different functions (e. g., '*вот*', '*короче*', '*там*', etc.). Besides, the annotators faced difficulties when selecting one main function between two, which have much in common (e.g., between *Hesitation* and *Search* pragmatic markers). Moreover, the first stage of annotation highlighted the need for unification of PM spelling, indicating its major variant, as a number of variations turned out to be possible. For example, '(*я*) *не знаю*', '(*ну*) *не знаю*, '(*в*) *это самое*', '*понимаешь* (*ли*)', '(*я*) (*ж*) *говорю*', '(*ты*) *знаешь* (*что*)', etc.

Therefore, for the second pilot annotation it was decided to make some changes into PM annotation scheme. Thus, it was decided to use an invariant representation of each PM. For this purpose, the list of these variants has been compiled. The list of PM functions has also been slightly changed (e.g., it was agreed to combine *Hesitation* and *Search* pragmatic markers, as they are often occurred in the same position as marked by different coders). *Navigational*, *Starting*, and *Final* PMs have been also united into a new *Boundary* type.

B. The second pilot annotation

When preparing the instruction for the second trial annotation, it was decided, first, to use a shorter list of PM functions, and second, to remove the mandatory requirement of indicating the main function of PMs. The last change was introduced due to the fact that almost every PM in spoken speech turned out to be polyfunctional, and the hierarchy of its functions turned out to be a complex phenomenon, which is difficult to describe unambiguously and unanimously. In the new annotation scheme, annotators were asked to list all relevant functions in the alphabetical order. Notation has also been changed, as it seemed expedient to increase readability of multifunctional codes.

For the second pilot annotation, the following 10 main PM functions were approved:

1) A — marker-approximator ('*типа*', '*как бы*', etc.);

2) G — boundary marker, which includes former *starting*, *final*, and *navigational* markers ('*вот*', '*короче*', etc.);

3) D — deictic marker ('*вот этот вот*', '*вот сюда вот*', '*вот такой вот*');

4) Z — replacement marker referring to some whole set or its part ('*и так далее*', '*и всё такое*', '*то-сё*'), as well as for imitating someone else's speech ('*бла-бла-бла*');

5) K — "xeno" marker that introduces someone's speech ('*говорит*', '*типа*', etc.);

6) M — meta-communicative marker that refers to "communication about communication" ('*знаешь*', '*видишь*');

7) F — "reflexive" marker that expresses reflection on what is said ('*так сказать*', etc.);

8) R — rhythm-forming marker ('*вот*', '*там*', etc.);

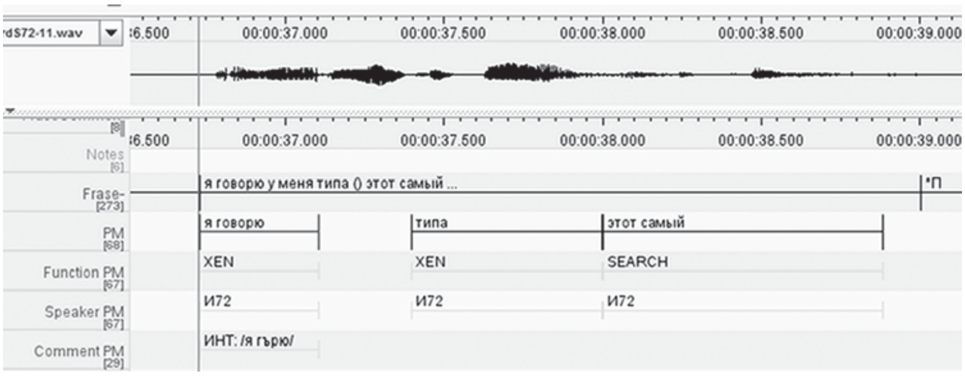
9) C — marker of self-correction ('*в смысле*, '*верней*', etc.);

10) H — hesitation marker, including searching one ('*это*', 'вот', '*там*', etc.).

The annotators agreed that it was easier to annotate according to the rules of the second instruction. The problem of determining both the main and additional functions of a particular marker was solved as well. And what is more important, the second annotation scheme allowed to achieve a better agreement of results obtained between annotators. The following section

TABLE I. THE EXAMPLE OF DATABASE OF ANNOTATION. FIRST STAGE (FRAGMENT)

No	File name	Start time	Duration	PM	Phrase	PM function	PM speaker	PM comment
1	ordS72-11_pm	5 420	740	*короче*	вчера короче / *П звонит Егорин брат //	HES NAVIG	И72	
2	ordS72-11_pm	13 150	180	*говорит*	мне говорит срочно / *П нужна твоя банковская карта //	XEN	И72	ИНТ: /гърит/
3	ordS72-11_pm	16 756	329	*я говорю*	я говорю / очень интересно //	XEN	И72	ИНТ: /гърю/
4	ordS72-11_pm	21 865	330	*говорит*	я говорит хочу купить два билета в театр //	XEN	И72	ИНТ: /гърит/
5	ordS72-11_pm	24 736	294	*типа*	типа купи мне два билета вот короче / *П сейчас немедленно / всё / горю //	XEN	И72	
6	ordS72-11_pm	25 960	240	*вот*	типа купи мне два билета вот короче / *П сейчас немедленно / всё / горю //	HES FIN	И72	
7	ordS72-11_pm	26 201	412	*короче*	типа купи мне два билета вот короче / *П сейчас немедленно / всё / горю //	HES FIN	И72	
9	ordS72-11_pm	29 246	475	*я говорю*	я говорю у меня(:) ...	XEN	И72	ИНТ: /гърю/
10	ordS72-11_pm	36 732	376	*я говорю*	я говорю у меня типа () этот самый ...	XEN	И72	ИНТ: /я гърю/
11	ordS72-11_pm	37 400	600	*типа*	я говорю у меня типа () этот самый ...	XEN	И72	
12	ordS72-11_pm	38 004	886	*этот самый*	я говорю у меня типа () этот самый ...	SEARCH	И72	
13	ordS72-11_pm	41 364	325	*я говорю*	Маестро$ / и () я говорю вряд ли чего-то можно на неё купить / *П потому что это Маестро$ // @ угу //	XEN	И72	ИНТ: /я грю/
14	ordS72-11_pm	47 673	302	*говорит*	хорошо говорит / я перезвоню //	XEN	И72	ИНТ: /грит/
16	ordS72-11_pm	56 830	327	*короче*	и он короче звонит / *П я уже в метро спускаюсь //	HES	И72	
17	ordS72-11_pm	72 220	335	*говорит*	и говорит / *П я тебе завтра занесу //	XEN	И72	ИНТ: /гъвърит/
18	ordS72-11_pm	74 051	299	*вот*	вот / *П типа бля буду / *П занесу //	HES START	И72	
19	ordS72-11_pm	74 500	310	*типа*	вот / *П типа бля буду / *П занесу //	XEN	И72	
20	ordS72-11_pm	89 980	740	*короче*	он / *П короче ...	HES SEARCH	И72	
21	ordS72-11_pm	95 959	486	*понимаешь*	наши новогодние подарки / *П светящийся ошейник для собаки / *П понимаешь это все где-то лежит / оно где-то есть / *П он был замечен вообще за такой () ерундой за всякой / *П что лучше даже не думать //	MET	И72	
22	ordS72-11_pm	108 140	285	*этот*	ну ты поаккуратнее ! *П так дала ему свою () тут () карту //	HES SEARCH	Ж1	
23	ordS72-11_pm	119 840	394	*такая*	а я такая говорю / Екатерины_Сергеевны% нет / *П она вышла //	XEN	Ж1	
24	ordS72-11_pm	120 234	134	*говорю*	а я такая говорю / Екатерины_Сергеевны% нет / *П она вышла //	XEN	Ж1	ИНТ: /гърю/
25	ordS72-11_pm	125 128	400	*я говорю*	Сергеевна% ? я даже не знал ! *П я говорю ну да // @ гм !	XEN	Ж1	ИНТ: /гърю/
26	ordS72-11_pm	129 766	439	*я говорю*	вон(?) / *П я говорю ну / *П по делам очевидно !	XEN	Ж1	ИНТ: /я грю/



Frase-: я говорю у меня типа () этот самый ... | *П

PM: я говорю | типа | этот самый

Function PM: XEN | XEN | SEARCH

Speaker PM: И72 | И72 | И72

Comment PM: ИНТ: /я гърю/

Fig. 1. The example of PM annotation in ELAN

concerns the description of inner-annotation agreement analysis, which has been performed in order to evaluate the effectiveness of current PM-annotation methodology and to receive a more detailed analysis of its approbation.

VI. Inter-annotator agreement

The inter-annotator agreement is used, in particular, for assessing the effectiveness (simplicity/comprehensibility, completeness, etc.) of guidelines that are implemented in manual annotation of corpus data. As a result of this assessment, researchers can find out how clearly the linguistic categories involved in the annotation are delineated.

There are three basic statistical measures for inter-rater reliability: Cohen's Kappa [36], Fleiss' kappa [37], and Krippendorff's Alpha [38].

In current research, inter-rater reliability was calculated using Kohen's Kappa. It is based on the observed proportions of agreement and disagreement between annotators in comparison with the expected proportions. This coefficient is recommended to be used in situations like ours, when a large number of items is to be annotated by a small group of raters [39].

Before data processing, it was necessary to consider that during the annotation process the variant implementations of the same PMs were allowed. For example, analyzing the phrase **П ну значит там (...) нахожу / диктую ей // *П* [ordS19-03], three annotators suggested three variants of the PM: '*ну значит*', '*ну значит там*', '*значит*', and the fourth annotator did not mark any PM at all in this utterance.

In the summary table, where the responses of annotators are given, not only the basic version of the PM '*значит*' is given, but also all its real implementations '*ну значит там*' and '*ну значит*'. Accordingly, we can evaluate not only the selection/non-selection of the PM in a particular utterance, but also its concrete form. In the described case, the table contains three lines corresponding to the three variants of PM marked in a particular utterance, and the data on the presence/absence of responses are presented in the form of codes "0" (if there is no answer) and "1" (if the corresponding variant of the PM has been annotated, and there is an answer). The fragment of the table, which presents the data processed for inter-annotator agreement is given in Table II.

TABLE II. The accepted way of coding annotators' answers

PM	Coder 1	Coder 2	Coder 3	Coder 4
я говорю	1	0	0	0
говорю	0	1	1	0
ну вот	1	1	0	1
вот	0	0	1	0
типа	0	1	0	0
такой	1	0	0	0
вот	1	1	1	1
он говорит	1	0	0	0
говорит	0	1	0	1
ну ладно	1	0	0	0

In the total, the table contains 1192 rows, which means that the total number of the PM annotated in the utterances of the pilot subcorpus by the four annotators is 1192 units. Several PM could be found and tagged in one utterance.

The table containing the answers of particular annotators was processed in R [40], [41]. As a result, the inter-annotator agreement index (kappa value) of 0.19 was obtained. The kappa is measured in the range from -1 to 1, and for us it was desirable to get a result close to 0.85, 0.9 or even to 1, since, according to accepted scales of interpretation, such indicators are considered high [25].

A low inter-annotator agreement index may indicate that either all annotators act differently (that is, the guideline should be improved), or there are those among them whose annotation strategies differ fundamentally from the strategies of the others. To test this assumption, we consistently excluded from the set of processed data the responses of the Annotator1, the Annotator2, the Annotator3, and the Annotator4. As a result, we obtained the best inter-annotator agreement coefficient after excluding the Annotator1 (kappa = 0.47).

In the pairwise comparison of the answers of the four annotators, the best inter-annotator agreement indices were obtained for the Annotators 2 and 3, while comparing all the annotators with the Annotator1 only negative results are obtained, see Table III below.

TABLE III. The pairwise comparison of annotators' answers

Coders	Kappa
Coder 1, Coder 2	-0,15
Coder 1, Coder 3	-0,10
Coder 1, Coder 4	-0,03
Coder 2, Coder 3	0,51
Coder 2, Coder 4	0,46
Coder 3, Coder 4	0,45

However, after bringing together all PM variants to the basic ones, we were able to identify a number of PMs, which the annotators extract most consistently. Among them PM '*видишь*', '*знаешь*', '*значит*', '*как бы*', '*понимаешь*', '*слушай*', etc. (the data are presented in the Table IV). In the rows for each PM, the number of uses marked by the particular annotator is given. For this PM list, the Krippendorff's Alpha coefficient is 0.89, that is, according to the scales of assessment, it can be estimated as high.

TABLE IV. The most unanimously tagged PM

PM	Coder 1	Coder 2	Coder 3	Coder 4
видишь	6	7	5	6
вся х…ня	1	1	1	1
знаешь	16	13	15	15
как бы	21	24	21	23
понимаешь	8	8	6	5
слушай	5	5	5	5

On the other hand, there is a group of relatively frequent PMs, which have been tagged less consistently — '*вообще*', '*вроде*', '*всё*', '*короче*', '*говорит*', '*думаю*' and other items (see Table V). For a given list, the obtained Krippendorff's alpha value is 0.14.

TABLE V. THE LEAST UNANIMOUSLY TAGGED PM

PM	Coder 1	Coder 2	Coder 3	Coder 4
вообще	0	0	0	20
вот	116	91	45	177
вроде	2	0	6	11
всё	4	0	0	23
говорит	53	40	42	20
да	21	2	7	104
думаю	4	4	2	11
короче	32	28	16	32
на самом деле	1	7	0	8
не знаю	15	16	2	16
так	22	13	12	39
такой	21	6	5	54
там	55	30	45	89
типа	14	9	11	15

VII. CONCLUSION

The conducted double pilot annotation of the subcorpus of everyday Russian speech, created on the basis of the ORD corpus, allowed to improve the proposed typology of pragmatic markers of spoken Russian [26] — to correct the PM list, to expand the scope of their functions, and to approve approaches to their annotation.

The comparison of the results of two stages of expert annotation allowed to clarify the functional resources of PMs in Russian oral spontaneous speech, and to develop an effective annotation scheme designed for processing of a large speech corpus. The obtained annotation results can be further processed automatically, despite the involvement of a relatively large number of PM-annotators.

The results of inter-annotator agreement measurement allowed to identify the group of PMs that are consistently recognized by all annotators. In regard to these PMs, we may say that their functions in speech are rather clear and unambiguous. However, we have revealed a group of other PMs, which are not less frequent in Russian then the first ones, but either they are not consistently perceived by annotators as being *pragmatic* markers, or there is no uniformity in attribution of their functions. In concern of these frequent elements of spoken Russian, additional investigations should be conducted.

Basing on the results of double pilot annotation presented in this paper, and taking into account the discrepancies that were revealed in expert decisions, PM annotation is to be carried out on representative subsets of two Russian speech corpora (ORD and SAT, the so-called Balanced Annotated Text Collection of monologues).

The methodology used in this research to identify a complete set of pragmatic markers can be regarded as a technology for verifying some generally accepted linguistic facts, akin to how an investigator in field linguistics evaluates the facts of a language, which he does not know. Finally, the proposed taxonomy of PMs in Russian spoken speech will be useful to developers of various NLP systems.

ACKNOWLEDGMENT

The presented research was supported by the Russian Science Foundation, project # 18-18-00242 "Pragmatic Markers in Russian Everyday Speech".

REFERENCES

[1] N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, G. Martynenko, E. Baeva, "Towards a description of pragmatic markers in Russian everyday speech", LNAI, Vol. 11096: Speech and Computer. *20th International Conference*, *SPECOM 2018*, Leipzig, Germany, September 18-22, 2018, Proceedings. Springer Publishing Company, 2018, pp. 42-48.

[2] B. Fraser, "An approach to discourse markers", *Journal of Pragmatics*, vol. 14, 1990, pp. 383-395.

[3] B. Fraser, "Pragmatic markers", *Pragmatics*, vol. 6, is. 2, 1996, pp. 167-190.

[4] J. Beliao and A. Lacheret, "Disfluency and discursive markers: when prosody and syntax plan discourse", *DiSS 2013: The 6th Workshop on Disfluency in Spontaneous Speech*, Stockholm, Sweden, № 54 (1), 2013, pp. 5-9.

[5] U. Lenk, *Marking Discourse Coherence: Functions of Discourse Markers in Spoken English*. Tuebingen: Narr, 1998, pp. 37-52.

[6] A. Popescu-Belis and S. Zufferey, "Automatic identification of discourse markers in dialogues: an in-depth study of like and well", *Computer Speech and Language*, Elsevier, The Netherlands, vol. 25, is. 3, 2011, pp. 499-518.

[7] D. Shiffrin, *Discourse Markers*. Cambridge: Cambridge University Press, 1996.

[8] L. Shourup, "Discourse markers", *Lingua*, Elsevier, The UK, № 107, 1999, pp. 227-265.

[9] D. Verdonik, M. Rojc, M. Stabej, "Annotating discourse markers in spontaneous speech corpora on an example for the Slovenian language", *Language Resources and Evaluation*, The Netherlands, № 41 (2), 2007, pp. 147-180.

[10] A. N. Baranov, V. A. Plungian, E. V. Rakhilina, *Putevoditel', po diskursivnym slovam russkogo yazyka* [*Guide to discursive words of Russian*]. Moscow: Pomovskii i Partnery, 1993 (in Russ.).

[11] *Diskursivnye slova russkogo jazyka: Opyt kontekstno-semanticheskogo opisanija* [*Discursive words of Russian: experience of contextual and semantic description*], K. Kiseleva and D. Paillard (eds). M.: Metatext, 1998 (in Russ.).

[12] *Diskursivnye slova russkogo jazyka: kontekstnoe var'irovanie i semanticheskoe edinstvo* [*Discourse words of Russian: contextual variation and semantic units*], K. Kiseleva and D. Paillard (eds). M.: Azbukovnik, 2003 (in Russ.).

[13] G. Leech, "Adding linguistic annotation", *Developing Linguistic Corpora: a Guide to Good Practice* / Ed. By M. Wynne. Oxford: Oxbow Books: 2005. Pp. 17-29.

[14] T. J. M. Sanders, W. P. M. S. Spooren, L. G. M. Noordman, "Toward a taxonomy of coherence relations", *Discourse Processes*, vol. 15, is. 1, 1992, pp. 1-35.

[15] C. Rühlemann, "What can a corpus tell us about pragmatics?", *The Routledge Handbook of Corpus Linguistics* / Ed. by A. O'Keeffe, M. McCarthy. London: Routledge, 2010. Pp. 288-301.

[16] L. Crible, *Discourse Markers and (Dis)fluency. Forms and Functions across Languages and Registers*. Amsterdam: John Benjamins, 2018.

[17] Université catholique de Louvain official website, MDMA – Model for Discourse Marker Annotation, Web: https://uclouvain.be/fr/instituts-recherche/ilc/valibel/mdma-model-for-discourse-marker-annotation.html.

[18] C. T. Bolly, L.Crible, L. Degand, D. Uygur-Distexhe, "Towards a model for discourse marker annotation: From potential to feature-based discourse markers", *Pragmatic Markers, Discourse Markers and Modal Particles: New perspectives* / Ed. by Ch. Fedriani and A. Sansó. Amsterdam: John Benjamins, 2017. Pp. 71-98.

[19] L. Crible, S. Zufferey, "Using a unified taxonomy to annotate discourse markers in speech and writing", *Proceedings of the 11th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation*, London, UK, 2015, pp. 14-22.

[20] R. Prasad, E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, The Penn Discourse Treebank 2.0 annotation manual. Technical report, Institute for Research in Cognitive Science, 2007, Web: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1203&context=ircs_reports.

[21] L. Crible, M.-J. Cuenca, "Discourse markers in speech: characteristics and challenges for corpus annotation", *Dialogue and Discourse*, vol. 8, no. 2, 2017, pp. 149-166.

[22] W. C. Mann, S. A. Thompson, "Rhetorical structure theory: Toward a

functional theory of text organization", *Text*, vol. 8,no. 3, 1988, pp. 243-281.

[23] T. J. M. Sanders, W. P. M. S. Spooren, L. G. M. Noordman, "Toward a taxonomy of coherence relations", *Discourse Processes*, vol. 15, is. 1, 1992, pp. 1-35.

[24] D. Samy, A. Gonzalez-Ledesma, "Pragmatic annotation of discourse markers in a multilingual parallel corpus (Arabic-Spanish-English)", *Proceedings of the International Conference on Language Resources and Evaluation*, *LREC 2008*, Marrakech, Morocco. Web: http://www.analedesma.es/wp-content/uploads/2010/04/doaingles.pdf.

[25] P. Artstein, M. Poesio, "Inter-coder agreement for computational linguistics", *Computational Linguistics*, 2008, vol. 34 (4), pp. 555-596.

[26] N. V. Bogdanova-Beglarian, "Pragmatemy v ustnoj povsednevnoj rechi: opredelenie pon'atia i obshchaja tipologia" ["Pragmatems in spoken everyday speech: definition and general typology"], *Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia* [*Perm University Herald. Russian and Foreign Philology*], iss. 3 (27), 2014, pp. 7-20 (in Russ.).

[27] N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, G. Martynenko, "Linguistic features and sociolinguistic variability in everyday spoken Russian", *SPECOM 2017*, LNAI, vol. 10458, 2017, pp. 503-511.

[28] A. Asinovsky, N. Bogdanova, M. Rusakova, A. Ryko, S. Stepanova, T. Sherstinova, "The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: creation principles and annotation". V. Matoušek, P. Mautner (eds.), *TSD 2009*, LNAI, vol. 57292009, Berlin-Heidelberg, Springer publ., 2009, pp. 250-257.

[29] N. V. Bogdanova-Beglarian, T. Ju. Sherstinova, K. D. Zajdes, "Korpus "Sbalansirovannaja annotirovannaja tekstoteka": metodika mnogourovnevogo analiza russkoj monologicheskoj rechi" [The corpus "Balanced Annotated Text Collection": Method of Multilevel Analysis of Russian Monological Speech] // *Analiz razgovornoj rechi (AR3-2017): trudy sed'mogo mezhdisciplinarnogo seminara* [*Analysis of Spoken Russian Speech (AR3-2017): Proceedings of the 7th Interdisciplinary Seminar*] / D. A. Kocharov, P. A. Skrelin (eds). St. Petersburg: Polytekhnica-print Publ., 2017, pp. 8-13 (in Russ.).

[30] N. Bogdanova-Beglaryan, A. Asinovskiy, O. Blinova, Ye. Markasova, A. Ryko, T. Sherstinova, "Zvukovoj korpus russkogo yazyka: novaja metodologia analiza ustnoj rechi [Sound Corpus of the Russian Language: a new methodology for analyzing the oral speech]. D. Shumska, K. Osga, (eds.), *Jazyk i metod: Russkij jazyk v lingvisticheskikh issledovaniakh XXI veka* [*Language and Method: The Russian Language in the Linguistic Studies of the 21st Century*], vol. 2, Kraków, 2015, pp. 357-372 (in Russ.).

[31] *Russkij jazyk povsednevnogo obshhenija: osobennosti funkcionirovanija v raznyh social'nyh gruppah* [*Everyday Russian Language in Different Social Groups*]. Collective monograph / Ed. by N. V. Bogdanova-Beglaryan. SPb.: LAJKA, 2016 (in Russ.).

[32] N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, G. Martynenko, "Linguistic features and sociolinguistic variability in everyday spoken Russian", *SPECOM 2017*, LNAI, vol. 10458, 2017, pp. 503-511.

[33] ELAN (Version 5.2) [Computer software]. (2018, April 04). Nijmegen: Max Planck Institute for Psycholinguistics, Web: https://tla.mpi.nl/tools/tla-tools/elan/.

[34] N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, G. Martynenko, "An exploratory study on sociolinguistic variation of Russian everyday speech", *SPECOM 2016,* LNAI, vol. 9811, Springer, Switzerland, 2016, pp. 100-107.

[35] A. Kibrik and V. Podlesskaya, (eds.), *Rasskazy o snovidenijakh. Korpusnoe issledovanie ustnogo russkogo diskursa* [*Night Dream Stories: A Corpus Study of Spoken Russian Discourse*]. Moscow: Languages of Slavic Cultures [Yazyki slavyanskikh kul'tur], 2009 (in Russ.).

[36] J. Cohen, "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, 1960, vol. 20 (1), pp. 37-46.

[37] J. L. Fleiss, "Measuring nominal scale agreement among many raters", *Psychological Bulletin*, 1971, vol. 76 (5), pp. 378-382.

[38] K. Krippendorff, *Content Analysis: An Introduction to its Methodology*. Beverly Hills, CA: Sage, 1980.

[39] S. T. Gross, "The kappa coefficient of agreement for multiple observers when the number of subjects is small", *Biometrics*, 1986, vol. 42 (4), pp. 883-93.

[40] B. Falissard "psy: Various procedures used in psychometry". R package version 1.1. 2012. Web: https://CRAN.R-project.org/package=psy.

[41] R. Lo Martire, "rel: Reliability coefficients". R package version 1.3.1. 2017. Web: https://CRAN.R-project.org/package=rel.