Russian Pragmatic Markers Database: Developing Speech Technologies for Everyday Spoken Discourse

Natalia Bogdanova-Beglarian[1], Olga Blinova[1,2], Tatiana Sherstinova[2,1], Ekaterina Troshchenkova[1]

[1]Saint Petersburg State University,
[2]National Research University Higher School of Economics,

Saint Petersburg, Russia

{n.bogdanova, o.blinova, t.sherstinova, e.troshchenkova}@spbu.ru

Abstract — **The paper presents recent results obtained within the ongoing project dedicated to the study of Russian pragmatic markers. Pragmatic markers are obligatory elements of natural speech in any language; moreover, they are considered to be functionally important for speech production and overcoming inevitable speech difficulties. A correct understanding of use and functions of pragmatic markers is a prerequisite for solution of many applied tasks related to speech technologies. The research is carried out on the data of two speech corpora — ORD corpus of Russian Everyday Speech known as "One Day of Speech" corpus and SAT corpus "Balanced Annotated Collection of Texts", which consists primarily of monologues. The article describes the database of Russian pragmatic markers designed to support both linguistic and pragmatic studies of spoken Russian and the development of speech technologies for everyday discourse. Besides, it presents actual statistical data on pragmatic markers distribution in natural speech depending on different factors.**

I. Introduction

The articles is focused on pragmatic markers (PMs) [1, 2] as an object of research interest. PMs are a group of elements within the class of auxiliary units of oral discourse. They are frequently used in oral speech [3]. Most of the PMs are the result of pragmaticalization process, which is very active in natural spoken language [4, 5]. In this process certain grammatical forms or individual lexemes go to the communicative-pragmatic level of the language and become purely pragmatic units [6, 7]. This process may be accompanied by changes in their use (for example, unrealized valency, non-standard word order, etc.). As a result, it is the function of the unit which it performs in the structure of oral discourse and which can be called the pragmatic meaning of this unit that becomes the main aspect of the pragmaticalized unit. Pragmatic markers are characteristic of unprepared (spontaneous) oral speech, both monologic and dialogical, they are extremely frequent in our everyday communication, sometimes depend in their use on the individual speaker, social relations of the interlocutors or the communicative situation itself, and therefore require detailed consideration, a word-by-word description and fixing in a special dictionary of pragmatic units [7].

The inventory of pragmatic markers is quite variable, but more or less universal in terms of functions, which makes it possible to create a general PMs classification, annotate real speech material on the basis of this classification and obtain specific numerical data on the conditions and features of the the PM use. Thus, it becomes possible to describe the PM system for Russian everyday speech as a whole as an important and integral part of oral discourse.

A correct understanding of use and functions of pragmatic markers is a prerequisite for solving many applied tasks related to speech technologies.

The article describes the database of Russian pragmatic markers designed to support both linguistic and pragmatic studies of spoken Russian and the development of speech technologies for everyday discourse. Besides, it presents actual statistical data on pragmatic markers distribution in natural speech depending on different factors.

II. Research Data

The research is carried out on the data of two speech corpora — ORD corpus of Russian Everyday Speech known as "One Day of Speech" corpus [9, 10, 11] and SAT corpus "Balanced Annotated Collection of Texts", which consists primarily of monologues [12].

A. ORD corpus

The One Day of Speech Corpus (ORD) is, at the moment, the most representative corpus of sound recordings of everyday speech communication in Russian, made according to the method of continuous multi-hour monitoring of speech. The technique implies that all the speech activity of an informant volunteer during the day is recorded with a voice recorder, literally "hanging on the informant's neck" [9]. The sound recordings were made in 2007–2016 in St. Petersburg. The current statistical characteristics of the corpus are as follows: 1250 hours of sound obtained from 128 informants and more than 1000 of their interlocutors, representing different social groups (sociolects) of the modern Russian city, 2800 macro-episodes of speech communication, 1 million word usages in text transcripts, 125 thousand word usages in the annotated subcorpus [13].

B. SAT corpus

The SAT (Balanced Annotated Text Library) corpus contains monologues of various types, in the form of sound recordings and their transcripts. The sampling of informants is balanced in terms of their social characteristics. To collect

data, an experimental author's program was used, including, in particular, retelling of the two texts – a text with a plot and without one, as well as descriptions of the two images – also one with a plot and a non-plot one, plus a free story on a given topic. SAT Speech Collection contains several modules of professionally homogeneous groups: 1) speech of medical personnel (MED); 2) the speech of lawyers (JUR); 3) speech of teachers of Russian as a foreign language (RKI), 4) speech of students (STUD), 5) speech of "computer experts" (COMP). At the moment, the collection contains materials obtained from 96 informants, and has 500 monologue texts and 20 hours of sound [12].

III. Pragmatic Markers Annotation

Continuous PM annotation was carried out at the following levels [3]:

Level 1. PM – a pragmatic marker in the form as presented in the transcript (filled in ELAN [14] annotation environment).

Level 2. Function PM – functions of the PM that must be indicated simultaneously, at the same level, in alphabetical order (filled in ELAN).

Level 3. Speaker PM – a speaker code (filled in ELAN).

Level 4. Comment PM – level of comments. This layer is intended for including optional information, as well as for marking complicated cases in which the identification of PMs and their functions was difficult (filled in ELAN).

Level 5. Standard – the standard version of the PM (without taking into account structural variants and / or the inflectional paradigm) (filled out according to the results of downloading the annotation levels in MS Excel).

Level 6. POS – part of speech marking of the original lexical unit from which the standard version of the PM evolved. Implemented automatically using the MyStem program (Yandex-technology) and then verified manually.

Level 7. Model – derivation model for PMs that consist from more than one word (filled in MS Excel).

Level 8. Frase – each PM use is correlated with the phrase context (filled in ELAN).

On the whole, 136 communicative macro-episodes were annotated for the ORD corpus, and 170 monologues for the SAT corpus. The total number of speakers in the annotated sub-corpus is 257 for the ORD and 34 for the SAT.

The phonetic features of PMs use were annotated selectively – for those PMs that have similar use positions with their lexical counterparts (eg, "tak" ("so") and "koroche" ("in brief"). Thus, material was obtained for comparing the phonetic realization of the PM and that of its lexical analogue.

Phonetic annotation was performed for 70 macro-episodes of the ORD corpus. Annotation Levels:

Level 1. PM – a pragmatic marker in the form as present in the transcript (filled in ELAN annotation environment).

Level 2. Function PM – PM functions (done in the ELAN).

Level 3. Speaker PM – a speaker code (done in the ELAN).

Level 4. Frase – phrase context (done in the ELAN).

Level 5. NM – marking here whether the given word is a PM or a lexical unit (done in MS Excel).

Level 6. Duration – the PM duration in milliseconds (measured in Praat).

Level 7. Intonation – melody movement (smooth, going up or down, complex) (measured in Praat).

Level 8. PitchDif – frequency difference in the stressed vowel, Hz (measured in Praat).

Level 9. VolDuration – duration of the stressed vowel, ms (measured in Praat).

Level 10. PitchKrut – the abruptness of the basic voice pitch frequency (VPF) change in the stressed vowel (calculated in MS Excel as the difference in the VPF divided by the duration of the stressed vowel).

Level 11. Position – the PM position in the utterance: I (isolated), H (at the beginning of the phrase), K (at the end of the phrase), B (inside the phrase) (filled in MS Excel).

The difficulties in PM annotation were described in [15].

IV. Pragmatic Markers Database Structure

The database contains data related to two speech corpora – SAT (monologic speech) and ORD (everyday speech, dialogues, polylogous communication). As to the material of the ORD corpus, two basic tables related to the description of speakers and episodes were imported from the ORD with minimal adaptation, however, as to the SAT corpus, the tables and the full catalogue of the corpus were created for the first time as part of this project. For the ORD, for the first time its data are correlated with psychological information about the informants. The two tables (SAT-PilotSubset and ORD-PilotSubset) represent a sample on which pilot PM annotation was performed in the amount of 15,000 word usages for the SAT corpus and 75,000 word usages for the ORD corpus (the third and fourth stage of pilot annotating). The results of the first two stages of this pilot annotation, performed simultaneously by 4 experts, are in the tables PM-Annotation-v1 and PM-Annotation-v2, and the results of the third and fourth stages of the pilot annotation can be found in the tables SAT-PM-Annotation and ORD-PM-Annotation.

The main database tables are as follows:

1) *Table SAT-Episodes* (Table of episodes / sound files of the SAT corpus) – a general catalogue of SAT sound recordings (was prepared during the implementation of this project)

SubCorpus – Subcorpus (Medical personnel, Lawyers, etc.)

SCode – Speaker Code

Gender – gender of the speaker

Age – age of the speaker

City – Place of Birth / Place of residence over a long period of time

Edu – Education

Profession – Profession of the speaker

URK – Speech Competency Level

Extraversion – Extraversion-Introversion characteristic of the speaker

Temperament – Temperament of the speaker

2) *Table SAT-PilotSubset* (Subcorpus from the SAT corpus prepared for pilot PM annotation). Contains the following fields:

SubCorpus – Subcorpus (Medical personnel, Lawyers, etc.)

SCode – speaker Code

Gender – gender of the speaker

Age – age of the speaker

City – Place of Birth / Place of residence over a long period of time

Edu – Education

Profession – Profession of the speaker

URK – Speech Competency Level

Tokens – Number of Word Uses

3) Table SAT-PM-Annotation (SAT – PM Annotation) – the results of the pilot annotation of monologue speech (50128 word usages)

SubCorpus – Subcorpus (Medical personnel, Lawyers, etc.)

SCode – Informant Code

SCode-2 – Code of the informant with certain speech competence level

Context – Context of use

PM – Pragmatic marker

Function – Function (s)

Comment – Comment

4–5) The PM-Annotation-v1 and PM-Annotation-v2 tables have a similar structure and reflect the results of the first two stages of pilot annotation. The description fields are as follows:

Time – Temporary Recording Report,

PM-Main – Pragmatic marker (structural version of the PM),

PM-Variant – Pragmatic marker (variant of the PM in use),

Context – Context of use,

Function – Function(s),

Comment – Comment

Episode – Episode / Sound File,

SCode – Informant code (according to the transcript),

Annotator – a person who performed the annotation.

6) *Table ORD-SpeakersSubcorpus* (Information about informants and their interlocutors from that part of the corpus that is subject to annotation). The structure of this table to a significant extent reproduces the structure of the original table of the ORD corpus, which is the data source:

SCode – informant code,

SName – the name of the informant on the questionnaire,

Gender – gender of the informant,

Age – the age of the informant at the time of recording,

PBirth – place of birth,

MLang – mother tongue,

Langs – other languages that the informant speaks,

Nat – parents' nationality,

SClass – social background,

Edu – education level,

Diploma – qualification (specialization) according to the degree or other education certificates,

PProf – past occupations or work experience,

Prof – current profession or occupation,

Regions – Places of residence over a long period of time,

Comments – comments of the annotators,

AgeGroup – age group,

PBirthN – normalized birthplace,

EduN – normalized level of education,

ProfGroup – dominant professional group (occupation),

Status – social status and some other description fields.

7) *Table ORD-PsychoTypes* (Information on the psychological characteristics of informants for the ORD corpus obtained as a result of psychological tests processing). Data on psychotypes are only available for informants recorded not earlier than 2014 (code S62 and more) – for a total of 69 informants. It contains the following description fields:

SCode – informant code,

Extraversion – Extraversion / Introversion,

Neurotism – The level of neurotism,

LieLevel – Level of veracity,

Temperament – Temperament (according to Eysenck test).

8) *The ORD-Episodes table* (macro-episodes of everyday speech communication from the ORD corpus) was imported from the ORD. It contains the following description parameters:

SCode – speaker code,

SFName – sound file name,

NComType – a normalized type of communicative episode,

NSRole – the social role of the informant in this episode (normalized code),

NPlace – locus (place) of communication (normalized code),

SFileOrig – name of the source (archive) file,

Start – the starting point of the episode relative to the beginning of the source file,

End – the end point of the episode relative to the beginning of the source file,

EPlace – place of communication (text description),

EAction – the main action accompanying the conversation, or a pragmatic goal,

EWho – the main interlocutors for the informant in this episode,

Duration – episode duration (min.),

FonQuality – phonetic quality in code representation (1 – maximum),

Priority – priority in annonation (rank markings),

SceneName – episode content and comments,

ELAN – the presence of a transcript of sound recording (logical field),

DivSpeak – breeding the decrypted file into speakers (logical field),

Comments – Commentary.

9) *Table ORD-PilotSubSet* (subsample of pilot annotation for the ORD corpus) – displays information which of the files from the ORD were used during each stage of pilot annotation. It contains the following description fields:

SFile – Episode / sound file,

Annotation – Pilot annotation stage,

NumAnnotators – Number of independent annotators,

Tokens – Volume of the text (word usages).

10) *Table ORD-PM-Annotation* (ORD – Annotation PM):

Episode – Episode / Sound File

SCode – Informant code (according to the annotation)

SCode-2 – Informant code (according to the database)

Context – Context of use

PM – Pragmatic marker

Function – Function (s)

Comment – Comment

Annotator – a person who performed the annotation)

The results of expert annotation of pragmatic markers were put into a database. Marked PM annotation levels were downloaded from the ELAN program into MS Excel, combined in MS Excel with automatic POS markings and expert annotation data (in the form of a table), and converted to MS Access tables. Thus, the database was expanded due to the introduction of new tables – ORD-PM-300000 (PM dialogic speech – 300 thousand tokens), SAT-PM-50,000 (PM monologic speech – 50 thousand tokens), PHON-PM- Praat (results of selective phonetic annotation), FL-POS (correlation of basic PM variants with “canonical” parts of speech), etc.

The database was formed for the entire annotated corpus material – 321504 tokens for dialogic speech and 50128 tokens for monologue speech. In the database, information from annotation files is correlated with information about the type of communicative scenario and other conditions of communication, as well as about the social and psychological characteristics of the speakers. As a result, new database tables were obtained – ORD-PM-CommScen, ORD-PM-Socio, etc., which allow studying the use of pragmatic markers for different sociolects, psychological types of a speaker, and for different communicative situations in oral speech.

V. Frequency Lists of Pragmatic Markers

Statistical processing of the PM expert annotation results was carried out, thus, quantitative information was obtained on the frequency of use for individual PMs and their types in oral speech. In particular, the following results were obtained:

A. PM Frequency lists for the whole sample

On the basis of all annotated material, 356 PM variants were found. The total share of PMs in speech is 27,753 ipm, or 2.77%. Out of the total number of PMs 82 variants were recorded both in dialogical and in monologic speech. The most frequent PM variants were the following (in brackets here and below, unless specified otherwise, the relative frequency is indicated in ipm – items per million):

vot [here] (6227), *tam* [there] (2917), *da* [yes] (1540), *kak by* [as if] (1256), *tak* [so] (1232), *znachit* [it means] (1041), *govorit* [(s)he says] (989), *nu vot* [well here] (741), *eto* [this]

(681), *znayesh'* [y'know] (664), *slushay* [listen] (648), *eto samoye* [this what I mean] (498), *koroche* [in brief] (397), *takoy* [such (masculine)] (385), *ponimayesh'* [d'you understand] (365), *tipa* [sort of] (361), *govoryu* [I say] (320), *ne znayu* [I don't know] (296), *etot* [this] (227), *voobshche* [in general] (223), *takiye* [such(plural)] (223), *vot tak vot* [so it's like this] (215), *vidish'* [d'you see] (207), *vso* [everything] (194), *v printsipe* [in principle] (182), *takaya* [such (feminine)] (178), *vot eto vot* [this here this] (130), *eti* [these] (122), *na samom dele* [actually] (117).

Numerous PM variations are an implementation of 59 standard (basic) PM forms. Of the total number of standard PM forms 25 can be seen both in dialogical and monological speech. The most frequent of them are as follows:

(...) *vot* [here] (7119), (...) *tam* [there] (2970), (...) *eto* [this] (…) (1827), (...) *da/da da da* [yes/yes, yes,yes] (1572), (...) *tak/tak tak tak* [so/so so so] (1357), (...) *kak by* [as if] (1353), *govorit/govoryu/govorim...* [I say/(s)he says/we say] (1337), *znachit* [it means] (...) (1062), *takoy* [such (masculine)] (1033), eto samoye [this what I mean] (879), (...) znayesh' [y'know, singular] (...)/(...) *znayete* [y'know, plural or respectful] (...) (839), vot (...) vot [here (…) here] (778), (...) *(po)slushay* [listen, singular] /(...) *(po)slushayte* [listen, plural or respectful] (750), (...) *ne znayu* [I don't know] (498), (...) *koroche govorya* [in brief] (462), (...) *tipa/tipa togo/tipa togo chto* [sort of/ sort of this/ sort of this like] (458), (...) *ponimayesh'* [do you understand, singular] / (...) *ponimayete* [do you understand, plural] (405), (...) *vso* [everything] (357), (...) *vidish'* (...) [d'you see, singular] / *vidite* [d'you see, plural or respectful] (255), *voobshche* [in general](231), (...) *dumayu* [I think] (...) (223), (...) *skazhem* [let's say] (...) (211), (...) *v printsipe* [in principle] (207), *vrode* [like] (...) (150), (...) *v obshchem* [in general] (130), *smotri* [look, singular]/*smotrite* [look, plural or respectful] (122), *na samom dele* [actually, in fact] (122), *(ty) predstavlyayesh'* [(d'y) imagine] (113), *shchas* [now, in a moment]/ *shchas shchas shchas* [in a moment, moment, moment] (93), (…) *tak daleye* [so on] (89).

Below one can see examples of some standard variant uses:

- uses of (…) **vot**: *vot / nu vot / da vot / i vot / tak vot / a vot / nu i vot / tak chto vot / vo / kak by vot /* etc.;
- uses of (…) **tam**: *tam / nu tam / vot tam / da tam*;
- uses of (...) **da / da da da**: da / da da da / nu da / vot da / da da da da / da da da da da;
- uses of (...) **eto** (…): eto / etot / eti / vot eto / etikh / v etom / vot etot / s etim / eta / etogo / vot eti / etu / v etot / s etimi / etoy / etom / vot etu / dlya etogo / na etom / etim / eto vot / na etikh / na eto / na etogo / nu eto / po etim / etomu / bez etogo / v eti / v etikh / v etikh vot / v eto / v etom to / vot eta / vot etikh / vot etogo / vot etoy / vso eto / za etim / iz etoy / ili etot / na etu / ne eto / nu tak eto / po vsemu etomu / po etoy / po etomu / pro etikh / s etoy / tak eto / eti vse / etimi / eto zhe / eto kak / eto eto eto / v etu / na etot / nad etoy, etc.

B. PM Frequency Lists for Everyday Dialogical Speech

In total, 315 variants of PM were registered in everyday dialogical speech. The total share of PM in dialogic speech was slightly larger than the average for the two samples – 28263 imp or 2.83%. The following options were the most common for dialogic speech:

vot (5779), *tam* (3151), *da* (1693), *tak* (1300), *kak by* (1295), *govorit* (1103), *znayesh'* (787), *slushay* (767), *eto* (758), *znachit* (758), *nu vot* (657), *eto samoye* (523), *koroche* (465), *ponimayesh'* (441), *takoy* (427), *tipa* (403), *govoryu* (360), *ne znayu* (341), *voobshche* (264), *takiye* (254), *etot* (254), *vidish'* (245), *vot tak vot* (235), *takaya* (201), *v printsipe* (163).

C. PM Frequency Lists for Monologue Speech

In total, in everyday monologue speech, 134 variants of PM were registered. The smaller number of variants can be explained by a significantly smaller sample size (50 thousand word usages in SAT vs. 300 thousand word usages in ORD). The total share of PMs in monologic speech turned out to be slightly less than in dialogic speech – 25712 imp or 2.57%. The most common for monologic speech were the following options:

vot (8666), *znachit* (2584), *tam* (1645), *nu vot* (1201), *kak by* (1044), *tak* (861), *da* (705), *vso* (653), *skazhem tak* (522), *v obshchem-to* (444), *eto samoye* (365), *govorit* (365), *v printsipe* (365), *vot eti vot* (287), *eto* (261), *vot eto vot* (183), *eti samyye* (183), *v obshchem* (183), *tak vot* (157), *takoy* (157), *i tak daleye* (157), *tipa* (131), *vot eta vot* (104), *nu vso* (104), *govoryu* (104), *vot tak vot* (104).

D. Statistical distribution of the main PM functional types

For dialogic speech, the following functions are most common:

a) monofunctional: X – hesitational marker (8921), M – meta-communicative marker (3362), G – border delimitation marker (starting, final and navigational) (3108), P – rhythm-forming marker (pace-maker) (1799), K – xeno marker (1683), A – marker-approximator (1420), D – deictic marker (1252), Z – replacing marker (240), F – reflective marker (158).

b) multifunctional (combining several functions): GX (3242), AX (767), GM (365), GMX (254).

For monologue speech, the following functions are most common:

a) monofunctional: X – hesitational marker (10128), G – border delimitation marker (3524), P – rhythm-forming (pace-maker) marker (1592), M – metacommunicative marker (574), K – xeno marker (548), Z – replacing marker (313), F – reflective marker (131), A – marker-approximator (104), C – self-correction marker (52), D – deictic marker (26),

b) multifunctional: GX – border delimitation + hesitative (5195), AX – approximator + hesitative (1932), DX – deictic + hesitative (888), FX – reflective + hesitative (392), GF – border delimitation + reflexive (104), MX – metacommunicative + hesitative (104).

The following functions turned out to be more characteristic of dialogical speech (the difference in ipm is indicated in parentheses): M – metacommunicative marker (2788), A – marker-approximator (1315), D – deictic marker (1226), K – xeno- marker (1135), GM – border delimitation + metacommunicative (338), GMX – border delimitation + metacommunicative + hesitative (254), P – rhythm-forming marker (206), and for monologue: GX – border delimitation + hesitative (-1952), X – hesitative (-1207)), AX – approximator + hesitative (-1164), DX – deictic + hesitative (-782).

V. DATABASE OF ILLUSTRATIVE EXAMPLES

The study also led to creating a list of illustrative examples of the PM use. The base of illustrative examples is presented in two versions – a table of MS Excel format and a corpus of sound files in * .wav format. Examples of the PM use were taken from both speech corpora. The main selection criteria were the following factors: 1) the unambiguity of assigning the unit to the category of pragmatic markers, 2) if possible, the lack of homonymy of the functions (polyfunctionality) of the PM, 3) the quality of the corresponding speech fragment recorded in the field work (low noise level, lack of superimposed speech, sufficient level of speech intensity).

The table has the following description fields:

Level 1. PM – a pragmatic marker in its basic (standard) form;

Level 2. Frase – phrase context;

Level 3. Function PM – PM functions that must be indicated simultaneously, at the same level, in alphabetical order (done in the ELAN);

Level 4. Speaker PM – speaker code (filled in ELAN);

Level 5. Episode – link to the name of a communicative episode (ORD) or monologue code (SAT);

Level 6. SFile – name of the sound file (containing the PM use in the phrase).

The resulting list of examples will be used as illustrative material in the upcoming multimedia “Dictionary of Pragmatic Markers of Russian Everyday Speech”. Such a dictionary can be useful to specialists of various kinds: linguists, researchers of everyday Russian speech, creators of speech grammar (grammar of uses), which, no doubt, differs from the grammar of language, translators of spontaneous texts into other languages, for instance within the frame of a literary work, when translating characters' speech, to teachers of the Russian language to foreigners who have to learn how to perceive and correctly understand our spontaneous speech both verbally and in the written form, while reading texts in Russian. Among other dictionaries based on corpus material, the PM dictionary will help to create the most complete picture of the lexical specificity of everyday Russian speech.

The electronic version of such a dictionary assumes that the user can to listen to all the contexts that make up the illustrative fund of the dictionary entry. Each such entry is more likely to resemble a lexicographic essay, which is determined by the specifics of the material.

VI. CORRELATIONS OF THE PM USE WITH DIFFERENT FACTORS

The combination of pragmatic annotation with the meta-information of the sound recordings of the ORD and SAT corpora allows us to study the PM use in different sociolects, by speakers of various psychological types, and for different communicative situations of everyday oral speech (see some data of this type in: [3]).

For instance, it is worth noting the use of metacommunicatives and xenomarkers, which are found in women’s speech much more often than in men's speech.

As to the use of PM by speakers with different psychotypes, the most frequent PMs in the speech of both groups were hesitative and border delimitation markers, but their distribution is not at all as uniform as it might seem at first glance. With introverts the use of the most frequent markers (X, G, M) stands out, the proportion of other functional types in their speech is small. Extroverts, on the other hand, use, firstly, all 10 types identified in the classification in their speech, and secondly, the use of the rest is equally frequent, with some exceptions. Introverts in the sample completely lack markers of self-correction – there were no such contexts for the analysis. Most likely, this means that introverts sustain internal dialogue more and think carefully about what they will say and with what words, which means they make fewer mistakes when articulating what they wanted to say (for more details see: [6]).

Corpus material makes it possible to obtain other data of this kind, which is the prospect of the proposed study.

VII. CONCLUSION

The database described in the article is an accumulating resource where data from a wide variety of categories are gathered and combined. It also provides basic information about the two speech corpora used in the study — ORD (dialogs) and SAT (monologues). This allows one to correlate various kinds of information (linguistic, pragmatic, extralinguistic, etc.), in particular, the results of expert annotation and the communication conditions of the corresponding speech episode.

An important result of the presented research is statistics on pragmatic markers use in Russian spoken language. A list of the main functional types of pragmatic markers has been prepared, empirical data have been obtained on the frequencies of basic (standard) PMs versions, as well as that of specific usages of these basic types, as well as for statistics on PMs use in speech of various types. These quantitative data are to be used for theoretical studies of everyday Russian and spoken discourse in general as well as for improvement of spoken NLP systems and speech technologies.

VIII. FURTHER RESEARCH

The study of PMs functioning in different types of spoken discourse (monologues and dialogues) will be continued. In particular, the study of communicative alignment in the use of PMs (interactive adjustment of the speech behavior of communication participants) will be carried out on the basis of dialogical speech.

In addition, we are going to find out about multifunctional pragmatic markers, capable of performing various pragmatic and procedural functions, in monologue speech and dialogues; to analyze differences in the use of PMs that have the same lexical and grammatical composition but perform different functions; to build "synonymous" series of pragmatic markers that can perform homogeneous functions in speech; to study sequences (chains) of pragmatic markers; to look into correlations between the function of the multifunctional pragmatic marker and its position in the utterance; to study structural variability of pragmatic markers (including phonetic variability, among others their ability to undergo phonetic reduction, and lexico-grammatical variability).

ACKNOWLEDGMENT

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

REFERENCES

[1] B. Fraser, "Pragmatic markers", Pragmatics, 6 (2), 1996, pp. 167–190.

[2] K. Aijmer, "Pragmatic Markers in Spoken Interlanguage", Nordic Journal of English Studies, 3, 2004, pp. 173–190.

[3] E. N. Torgersen, C. Gabrielatos, S. Hoffmann, S. Fox, "A corpus-based study of pragmatic markers in London English", Corpus linguistics and linguistic theory, 7(1), 2011, pp. 93–118.

[4] Degand, L., Evers-Vermeul, J. Grammaticalization or Pragmaticalization of Discourse Markers? More than a Terminological Issue // Journal of Historical Pragmatics. 16:1, 2015, pp. 59-85.

[5] G. Diewald, "Pragmaticalization (Defined) as Grammaticalization of Discourse Functions", Linguistics. 49(2), 2011. pp. 365–390.

[6] L.J. Brinton, *Pragmatic markers in English: Grammaticalization and discourse functions*. Mouton de Gruyter, Berlin, 1996.

[7] S. Günther, K. Mutz, "Grammaticalization vs. Pragmaticalization? The Development of Pragmatic Markers in German and Italian", in W. Bisang, N. P. Himmelmann, B. Wiemer (eds.). What Makes Grammaticalization? A Look from its Fringes and its Components. Berlin: Language Arts & Disciplines, pp. 77-107, 2004.

[8] E. Graf, *Interjektionen im Russischen als Interaktive Einheiten*. Frankfurt am Main, 2011.

[9] A.N. Baranov, V.A. Plungyan, E.V. Rakhilina, *Putevoditel' po diskursivnym slovam russkogo yazyka* [Russian Discourse Words Guide]. Moscow, 1993. (in Russ.)

[10] 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", in: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNAI, vol. 5729. Springer, Berlin-Heidelberg, 2009. pp. 250–257.

[11] N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, O. Ermolova, E. Baeva, G. Martynenko, A. Ryko "Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech", in Ronzhin, A. et al. (eds.) SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, 2016, pp. 659–666.

[12] N. Bogdanova-Beglarian, O. Blinova, T. Sherstinova, G. Martynenko, "Corpus of Russian Everyday Speech "One Day of Speech": present state and prospects" [Korpus russkogo yazyka povsednevnogo obshcheniya «Odin rechevoj den'»: tekushchee sostoyanie i perspektivy], in Proceedings of the V.V. Vinogradov Institute of Russian language. Vol. 21. Russian National Corpus: research and development [Trudy Instituta russkogo yazyka im. V. V. Vinogradova. Vyp. 21. Nacional'nyj korpus russkogo yazyka: issledovaniya i razrabotki] / A. M. Moldovan, V. A. Plungyan (eds.). Moscow, 2019, pp. 101-110.

[13] N. Bogdanova-Beglarian, O. Blinova, K. Zaides, T. Sherstinova, " 'Balanced Annotated Collection of Texts' (SAT; monologues): studying the specifics of Russian monological speech ['Sbalansirovannaya annotirovannaya tekstoteka' (SAT): izuchenie specifiki russkoj monologicheskoj rechi], Proceedings of the V. V. Vinogradov Russian language Institute. Vol. 21. National corpus of the Russian language: research and development [Trudy Instituta russkogo yazyka im. V. V. Vinogradova. Vyp. 21. Nacional'nyj korpus russkogo yazyka: issledovaniya i razrabotki] / A. M. Moldovan, V. A. Plungyan (eds.). Moscow, 2019, pp. 111-126.

[14] N. Bogdanova-Beglarian, O. Blinova, G. Martynenko, T. Sherstinova, "Some Invariant Features of Russian Everyday Speech: Phonology, Morphology, Syntax", Komp'juternaja Lingvistika i Intellektual'nye Tehnologii [Computer Linguistics and Internet Technologies], 2(16), 2017, pp. 82-95.

[15] ELAN Web: https://tla.mpi.nl/tools/tla-tools/elan/.

[16] N.V. Bogdanova-Beglarian, O.V. Blinova, T.Ju. Sherstinova, E.V. Troshchenkova, D.A. Gorbunov, K.D. Zaides, "Pragmatic Markers of Russian Everyday Speech: the Revised Typology and Corpus-Based Study", Proceedings of the 25th Conference of Open Innovations Association FRUCT / S. Balandin, V. Niemi, T. Tuytina (eds.). Helsinki, Finland, 2019, pp. 57-63.

[17] N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, G. Martynenko, "Pragmatic Markers Distribution in Russian Everyday Speech: Frequency Lists and Other Statistics for Discourse Modeling", SPECOM 2019. Lecture Notes in Artificial Intelligence, LNAI, vol. 11658. Springer, Switzerland, 2019, pp. 433-443.

[18] D.A. Gorbunova, "Extravert vs Introvert: the Frequency of the Pragmatic Markers in Speech of Different Informants", Cognitive Studies of Language. Vol. XXXVII. Integrative Processes in Cognitive Linguistics: Papers of International Congress on Cognitive Linguistics. May 16-18, 2019. Moscow – Tambov – Nizhny Novgorod: DECOM, 2019, pp. 1040-1044. (In Russ.).