Pragmatic Markers and Parts of Speech: on the Problems of Annotation of the Speech Corpus

Natalia Bogdanova-Beglarian[0000-0002-7652-0358] and Kristina Zaides[0000-0001-7528-0420]

St. Petersburg State University, 7/9 Universitetskaya nab., 199034, St. Petersburg, Russia
nvbogdanova_2005@mail.ru, zaidi94@mail.ru

Abstract. The article considers the range of possibilities of pragmatic markers (PM) annotation: from the speaker's code to the speaker's commentaries for all difficult cases. The research is based on the material of two corpora of everyday Russian speech – "One Day of Speech" (ORD; dialogues / polylogues) and "Balanced Annotated Text Collection" (SAT; monologues). Two main annotation levels have become the objects of research in this paper: the part of speech of the original lexical unit, from which the basic version of the PM had derived (POS), and the model of formation of the PM which consist of more than one word (Model). The research shows the low feasibility of trying to fit PM into the system of traditional parts of speech, and, conversely, the importance and role of defining a model of formation of PM for their systematic description. In any case, the automatic annotation of corpus material turns out to be considerably difficult.

Keywords: Spoken Speech, Speech Corpus, Pragmatic Marker, Pragmaticalization, Part of Speech, Model of Formation.

Introduction

A speech corpus, by definition, should include not only a set of texts, but also their annotation [Zakharov 2005: 4; Plungian 2005: 6]. Two corpora of everyday Russian speech, which became the sources of observations for this research, are annotated: "One Day of Speech" (ORD; dialogues / polylogues) (see about that: [Russkij yazyk … 2016; Bogdanova-Beglarian et al. 2016 a, b) and "Balanced Annotated Text Collection" (SAT; monologues) (see: [Zvukovoj korpus … 2013]). The corpus "One Day of Speech" was formed using the method of long hours monitoring. This method is traditionally used in Japanese field linguistics studies (see: [Shibata 1983; Campbell 2004]), and, furthermore, was implemented during data collection for the spoken language part of the British National Corpus [Burnard 2020]. The advantage of this method lies in the receiving for the analysis such spoken material which is the closest to the natural everyday speech. During the process of the corpus development, this method has been first-time used as applied to the Russian language. The specificity of the corpus "Balanced Annotated Text Collection" is that it includes monologues received from the socially balanced groups of native speakers. The monologues follow 4 most common communicative scenarios: reading of a text, retelling of a text, de-

scription of a picture and a story. This balancing allows comparing monologues of one speaker, produced in different communicative situations, and monologues of different speakers, produced in similar communicative situations.

Apart from other types of annotation, in both corpora the pragmatic markers (PM) were annotated. PM constitute a significant part of structural units of any oral text: 2,77 % in the whole material, 2,83 and 2,57 % in dialogue and monologue, respectively.

1 Annotation of PM: Literature Background

In this paper, the term "pragmatic markers" describes particular discourse units (words, expressions, and phrases) with a weakened (sometimes even vanished) referential meaning, which have a variety of functions in the discourse: marking the start and the end of the speech, pause-filling, speech-reflection, etc. The term "pragmatic marker" was developed by B. Fraser who defined them as a class of words that signal some important for the speaker messages towards the speech [Fraser 1996]. The majority of the researchers more often use the term "discourse markers" (DM), referring to a group of discourse units which structure the text or label different kinds of relations between its parts. [Baranov et al. 1993; Shiffrin 1996; Lenk 1998; Kiseleva, Paillard 1998, 2003; Schourup 1999].

However, there are some important differences between discourse and pragmatic markers as the objects of the investigation [Bogdanova-Beglarian 2018]: DM are put in the text consciously by the speaker, and PM are inserted unconsciously as speech automatisms; DM in some cases have lexical and grammatical meanings, and PM usually lose both lexical meaning and grammatical properties completely; DM, in general, convey speaker's attitude toward the speech, and PM express speaker's relation to the speech process itself and verbalize difficulties in speech production; moreover, most of DM can be found in the language dictionaries, while PM are left out of the lexicographic description.

Thus, the meanings of the terms PM and DM do not fully coincide. However, the specificity of their annotation in the corpus material is similar in many respects. This paper presents several results of the first attempt of pragmatic markers annotation understood as specific elements of merely oral speech. The annotation of discourse markers, a wider range of units, as shown, was carried out in the different corpora.

D. Verdonik, M. Rojc, and M. Stabej annotated DMs in the corpus of Slovenian telephone conversations TURDIS and analyzed as DM, among the DM in the conventional understanding, hesitations and backchannel expressions. For most cases, as the researchers noticed, "it is not possible to say that a discourse marker performs only one of <…> pragmatic functions" [Verdonik, Rojc, Stabej 2007: 162]. The authors suggest manual annotation of markers since even if the development of an algorithm, trained on the manually annotated data, is possible, subsequent manual correction is needed because of the existing ambiguation of markers and content words.

L. Crible and S. Zufferey implemented the annotation of DM in French and English spoken speech and written texts using the structure of four domains — ideational,

rhetorical, sequential, and interpersonal [Crible, Zufferey 2015]. The researchers stated that the inter-annotator agreement of manual annotation was from 34% (for English texts) to 52% (for French speech). Several issues of such annotation arose: e.g., the distinction of similar DM functions, the discovery of new functions and their clustering, and the ambiguation of markers and other words.

L. Crible and M.-J. Cuenca [Crible, Cuenca 2017] reported that most annotation models of DMs were developed for annotation of written discourse: the Rhetorical Structure Theory (RST) [Mann, Thompson 1988], the Penn Discourse Treebank [Prasad et al. 2007], and the Cognitive approach to Cognitive Relations (CCR) [Sanders et al. 1992]. The researchers annotated discourse markers in the French-English corpus DisFrEn without applying the prescribed DM list [Ibid.]. The following problems appeared during the annotation: the presence of truncated structures in spoken speech, the ambiguity of some DM, and the multifunctionality of markers. Therefore, the authors concluded that the automatic annotation of DM is not possible.

Regarding the semi-automatic annotation of pragmatic units, the EXMARaLDA annotation tools, for instance, should be mentioned. It allows marking two or more functions for each discourse marker in different contexts manually or semi-automatically, using prescribed list of discourse functions, but does not allow the process of annotation being completely automatic [Crible 2018]. The automatic tool for the annotation of discourse markers is provided in the MDMA (Model for Discourse Marker Annotation) project, which uses the methodology named "back-and-forth from theory to data" [Université catholique… 2020]. Within this project, manual selection of DM in the spoken speech and their further semantic, syntactic and pragmatic annotation is made for the NLP-tasks. The results of the research showed that only the initial position of the marker in the sentence let the algorithm based on statistical modeling identify the marker and its particular function [Bolly et al. 2017].

PRAGMATEXT model of annotation includes the list of tagged pragmatic functions, e.g., labelling emotional language, discourse relations, discourse modality, speech act, etc. The researchers used this model for the first attempt of discourse markers annotation in the multilingual parallel corpus (Arabic-Spanish-English) [Samy, Gonzalez-Ledesma 2008]. At the first stage, the Spanish part of the corpus was annotated at the discourse markers level. At the second stage, the comparison with the DM in texts in another two languages was made using a bilingual dictionary. Non-ambiguous DM in Arabic and English texts were automatically tagged the same way as in the Spanish texts. The ambiguous markers were disambiguated manually considering their prosodic features and position within the sentence. The authors intend to develop the automatic disambiguation tool for the purposes of the DM annotation; however, it is prevented by such factors as DM categorical, syntactic, and discursive ambiguity, as well as the absence of the clear distinction between DMs and idiomatic expressions since the DM tend to form the lasts.

Thus, as it was shown above, pragmatic markers annotation of corpus spoken speech data can be performed manually and semi-automatically with necessary checking.

2 Annotation of PM and Types of Pragmatic Markers

The annotation was implemented at the several levels: the particular usage of PM (PM); its functions in this particular usage (Function PM); the commentaries for introducing the optional information and marking the difficult cases which show the troubles in the detection of PM and their functions (Comment PM); the basic version of PM (excluding its structural versions and/or inflectional paradigm) (Standard); the parts-of-speech tagging of the source lexical unit, from which the basic version of the PM derived (POS); the model of formation of the PM which consist of more than one word (Model); the speaker's code (Speaker PM); and phrase commentary (Phrase) [Bogdanova-Beglarian et al. 2018, 2019b].

The main functional types of PM in oral discourse turned to be the following: A – marker-approximator (*vrode 'like', kak by 'kinda'*), G – boundary marker (starting, final, and navigational), D – deictic marker (*vot (...) vot 'like ... this'*), Z – all types of replacement markers (for someone else's speech, whole set or its parts: *bla-bla-bla, i vse dela 'and all that', i vs'o takoe prochee 'and all that stuff'*), K – "xeno" marker (*tipa (togo chto) 'sort of', takoj 'like'*), M – meta-communicative marker (*da 'yeah', (ja) ne znaju '(I) don't know', znaesh' 'you know', smotri 'look'*), F – reflexive marker (*skazhem tak 'let's say', ili kak tam? 'or whatever'*), H – hesitation marker (*eto 'what', tam 'em', eto samoe 'whatchamacallit'*) [Ibid.].

The automatic annotation of such material seems almost impossible: the very specificity of oral spontaneous speech, which is difficult to any systematization, causes too many problems. For instance, the syntagmatic division of spontaneous speech itself is problematic, which is relevant for the distinction of various boundary PM (G): starting, navigational, and final. It is also difficult to establish a distinction between the formally similar PM and meaningful units of discourse, that are pragmaticalized in the speech and sometimes are at the different stages of the pragmaticalization from the lexemes to the pragmatemes (*on prishol, a **tam** nikogo net 'he came but no one was there'* (adverb of place) – *on **tam** prishol **tam**, a nikogo net 'he em came em but no one was there'* (two PM used in hesitative and rhythm-forming functions)).

The most PM of Russian speech are polyfunctional, which leads to the necessity of identifying the main and additional functions of each marker in every particular case (*on **tam** prishol **tam** 'he em came em'* – HR). At last, spontaneous speech reveals such feature of PM as their "magnetism", attraction of one PM to another if they have one common (synonymous) function. Consequently, the need to distinguish different PM which consist of more than one word, on the one side, and a chain of markers, on the other side, appears: *eto kak jego 'what whatchamacallit'* (one marker) or *eto + kak jego 'er + whatchamacallit'* (a chain of markers) [Bogdanova-Beglarian et al. 2019a].

3 Pragmatic Markers and Parts of Speech

The set of the PM revealed in the material shows that PM have different "origins" in the field of parts of speech: particles (*vot 'here', von 'there'*), verbs (*znat' 'to know', govorit' 'to speak', smotret' 'to look', dumat' 'to think'*), including gerund (*govor'a*

'speaking'), adverbs (*tam 'there', tak 'that way', kuda 'where'*), pronouns (*etot 'this', samyj 'the most', on 'he', ona 'she'*), conjunctions/prepositions (*tipa 'kind of', vrode 'like'*).

The parts-of-speech tagging of corpus data was initially made automatically with the software "MyStem" (Yandex Technologies) and then checked manually. Only particular usages of PM were annotated. In the table 1, the results of this annotation in the ORD-corpus for the top of the frequency list of PM (first 20 ranks) are demonstrated. It can be already seen that certain difficulties during the automatic annotation with a help of "MyStem" software application arise, as well as insignificant divergence of two annotation types.

Thus, the program does not identify the integrity of the unit *kak by 'kinda'*, marking it as "adverbial pronoun + particle" (ADVPRO&PART), while, during the manual annotation, the experts choose the option which is closer to the nature of this unit – "particle / conjunction".

The software attributes to the marker *znachit 'well'* the tag "adverb and parenthesis" (ADV, parenth), whereas the manual annotation gives a variant "verb / parenthesis". The element *eto 'what'* in all PM (*eto 'what'* and *eto samoe 'whatchamacallit'*) is marked by the software as the "subject pronoun" (SPRO), although the traditional grammar, which became a base of manual annotation, treats this unit as the "adjective pronoun", that nominalized in the particular cases. The adjective nature of this unit is supported, for instance, by the ability of gender inflection (*eta 'what (fem.)', eto 'what (neutr.)', etot 'what (masc.)'*). The marker *tipa 'kind of'* in the automatic annotation is merely the "particle" (PART), although in the manual annotation it is "noun / preposition", which is required by the dictionaries in the first place.

However, even considering revealed inaccuracy of the automatic POS annotation of material, it is clear that the information about the POS of the original units, which have pragmaticalized and became pragmatic markers in oral speech, is rather a historical background which does not really describe new discourse units. For instance, the markers *tam* and *tak* as a PM lose all the adverbial properties [Turchanenko 2018], the word *da* as a meta-communicative marker falls into category of neither particle, nor conjunction [Shershneva 2015].

The verbal meta-communicative markers similarly lose the majority of their verbal characteristics in their new usage: verbs in indicative mood like *znaesh'/znaete 'you know', vidish'/vidite 'you see', ponimaesh'/ponimaete 'you know'* and verbs in imperative mood as *slushaj/slushajte 'listen', predstav'/predstav'te 'imagine'* leave merely formal number inflection [Bogdanova-Beglarian, Maslova 2019], the markers *(ja) ne znaju '(I) don't know'* и *znachit 'well'* completely lose any grammatical inflection and are used only in one fixed form [Bogdanova-Beglarian 2019], and the pragmatic "xeno" marker *govorit 'says'* is presented in the spoken speech solely in the present tense forms, more often phonetically reduced (*grit, gyt, gr'u, grim*, etc.) [Stojka 2019].

Table 1. The top of the frequency list of PM cases in the ORD-corpus: frequencies and POS tagging (for 300,000 tokens)

Rank	PM	Frequency	IPM	POS (aut.)	POS (man.)
1.	*vot 'er'*	1205	4017	PART	particle
2.	*tam 'em'*	657	2190	ADVPRO	adverbial pronoun
3.	*da 'yeah'*	353	1177	PART	particle / conjunction
4.	*tak 'this way'*	271	903	ADVPRO	adverbial pronoun
5.	*kak by 'kinda'*	270	900	ADVPRO&PART	particle / conjunction
6.	*govorit 'says'*	230	767	V	verb
7.	*znaesh' 'you know'*	164	547	V	verb
8.	*slushaj 'listen'*	160	533	V	verb
9.	*znachit 'well'*	158	527	ADV, parenth	verb / parenthesis
10.	*eto 'what'*	158	527	SPRO	adjective pronoun
11.	*nu vot 'well er'*	137	457	PART&PART	particle + particle
12.	*eto samoe 'whatchamacallit'*	109	363	SPRO&APRO	adjective pronoun + adjective pronoun
13.	*koroche 'in short'*	97	323	ADV, parenth	adverb / parenthesis
14.	*ponimaesh' 'you know'*	90	300	V	verb
15.	*takoj 'like'*	89	297	APRO	adjective pronoun
16.	*tipa 'sort of'*	84	280	PART	noun / preposition
17.	*govor'u 'I say'*	75	250	V	verb
18.	*ne znaju '(I) don't know'*	71	237	PART&V	particle + verb
19.	*voobshche 'generally'*	55	183	ADV, parenth	adverb / parenthesis
20.	*takie 'like'*	53	177	APRO	adjective pronoun

It could hardly be correct to refer all these pragmaticalized forms to the certain traditional POS categories.

4 Basic Versions of PM and Parts of Speech

The top of the frequency list of basic (standard) versions of PM seems slightly different than the one of particular usages of PM in the table (the data from the two corpora altogether):

(...) vot '(…) er' (IPM here and hereinafter – 7119),
(...) tam '(…) em' (2970),
(...) eto, eta, eti… (…) '(…) what… (…)' (1827),
(...) da/da da da '(…) yeah/ yeah yeah yeah' (1572),
(...) tak/tak tak tak '(…) well/well well well (1357),
(...) kak by '(…) kinda' (1353),
govorit/govor'u/govorim... 'says/say…' (1337),
znachit (...) 'well (…)' (1062),
takoj/takaja, takie 'like' (1033),
eto samoe/eti samye, etot samyj… 'whatchamacallit…' (879),

(...) znaesh' (...)/(...) znaete (...) '(...) you know (...)' (839),
vot (...) vot 'like (...) this' (778),
(...) (po)slushaj /(...) (po)slushajte '(...) listen' (750),
(...) ne znaju '(...) don't know' (498),
(...) koroche govor'a '(...) long story short' (462),
(...) tipa/tipa togo/tipa togo chto '(...) sort of' (458),
(...) ponimaesh' / (...) ponimaete '(...) you know (...)' (405),
(...) vs'o 'that's all' (357),
(...) vidish' (...) / vidite '(...) you see (...)' (255),
voobshche 'generally' (231),
(...) dumaju (...) '(...) think (...)' (223),
(...) skazhem (...) '(...) let's say (...)' (211),
(...) v principe '(...) basically' (207),
vrode (...) 'like (...)' (150),
(...) v obshchem '(...) anyway' (130),
smotri/smotrite 'look' (122),
na samom dele 'actually' (122),
(ty) predstavlyaesh' 'you know' (113),
shchas/shchas shchas shchas 'one moment' (93),
(...) tak dalee '(...) so on' (89).

One could see that the majority of markers has only generalized structure of basic version, with potential extension or restricted grammatical flexibility, cf.: VOT – *i vot 'and er', da vot 'well er'*; ZNAESH – *ty znaesh' 'you know', vot znaes'h 'er you know', nu znajete 'well you know'*, etc.; GOVORIT – *govor'u 'I say', govorish' 'you say', govorim 'we say'*, etc.; VRODE – *nu vrode 'well like', vrode kak 'like as', vrode by 'like well'*. Rare PM from the whole list of PM, annotated in the material, do not show such structural variability: *von 'err', prikin' 'guess', i tak dalee 'and so on', po idee 'normally'* and a few others. However, the deictic marker VOT (…) VOT *'like ... this'* exists merely as a structural model, which is filled by a new unit each time: *vot* ***tak*** *vot 'like this', vot* ***takoj*** *vot 'like this', vot* ***ots'uda*** *vot 'like this'*, etc. In fact, this marker does not have some single basic (standard) form. The automatic parts-of-speech tagging of such material not only seems difficult, but also has rather inaccurate results since it cannot consider the specificity of possible extensions.

5 Models of Formation of PM

The annotation of corpus material at the level of models of formation of the PM, which consist of more than one word (Model), is supposed to be the most informative and scientifically valuable. At least 12 such models have been identified:

1. PM, which initially consist of more than one word, that are basic versions (but not the source "lexicographic" version): *eto samoe 'whatchamacallit', kak jego (jejo, ikh) 'whatchamacallit', kak eto? 'whatchamacallit?' kak skazat'? 'how can I say?' kak eto nazyvaets'a? 'what am I call it?' chto jeshcho? 'what else?' kak (by) skazat'? 'how can I say it?'*

2. combination of two or more PM, which consist of one word*: nu vot 'well er', nu tam 'well em', nu tak 'well um', vot tak 'er um', nu znaesh' 'well you know', tak skazhem 'let's say', skazhem tak 'let's say', skazhem tam 'let's say em', vot skazhem 'er let's say', znaesh' tam 'you know em', vot kak by 'er kinda', vot skazhem tak 'er let's say', nu ne znaju 'well don't know', tam tipa 'em sort of', nu koroche 'well in short', znachit vot 'well er', v principe vs'o 'basically that's all'*;
3. combination of PM, which consist of one and more than one word: *nu kak skazat'? 'well how can I say?' kak jego tam? 'em whatchamacallit?' nu vot eti vot 'well these ones', nu (ja) ne znaju tam 'well (I) don't know em'*;
4. addition of the personal pronoun with a weakened lexical and grammatical meaning: *(ja) ne znaju '(I) don't know'; (ja) (ne) dumaju (chto) '(I) don't think (that)'; (ty) znaesh', ponimaesh', vidish'… '(you) know'; (ty) predstav', prikin'… '(you) imagine'; chto (tebe) jeshchyo skazat'? 'what else can I say (to you)?'*
5. addition of emphatic particles/conjunctions: *i vse dela 'and all that', i vs'o takoe 'and all that', a vot 'and er', nu i vs'o 'well that's all', i to i s'o 'this and that', ja uzh ne znaju tam 'I even don't know em', ty zh ponimaesh' 'you really know'*;
6. addition of non-personal pronoun: *vs'o takoe prochee 'all that stuff', tipa togo/etogo 'sort of', vrode togo 'like', takoj kakoj-to 'like that one', kak (by) eto skazat'? 'how can I say that?'*
7. addition of the conjunction CHTO (CHEGO) 'that': *dumaju chto 'think that', bojus' chto 'am afraid that', tipa togo chto 'sort of that', vrode togo chto 'like that', znaesh' chto/chego 'you know that'*;
8. addition of parentheses: *vs'o navernoe 'that's all probably', vs'o pozhaluj 'that's all perhaps'*;
9. addition of interjection: *oj slushaj 'ooh listen'*;
10. loss of the gerund GOVORYA 'speaking': *koroche 'in short', sobstvenno 'strictly', voobshche 'generally'*;
11. reduplication: *da-da-da 'yeah-yeah-yeah', na-na-na, shchas-shchas-shchas 'one moment-one moment-one moment', te-te-te, op-op-op, bla-bla-bla, tak-tak-tak 'em-em-em', eto-eto-eto 'what-what-what'*;
12. ILI 'or' + (more often) the rhetorical question: *ili kak jego? 'or whatchamacallit?' ili kak tam? 'or whatchamacallit?' ili chto? 'or what?' ili kak skazat'? 'or how to say?' ili etot? 'or what?' nu ili ne znaju 'well or I don't know'*.

Conclusion

Previous works have shown that in the speech recognition process POS-tagging of some markers (excluding multi-word markers and phrases) can be useful for the task of prediction of the following words [Heemant et al. 1998].

However, the automatically derived classification algorithm of DM POS-tagging showed an error rate of 37,3%, in comparison to, for instance, the error rate of 45,3% for the algorithm of J. Hirschberg and D. Litman [Hirschberg, Litman 1993]. In other words, using this automatic algorithm (decision tree), only for 4 from 10 particular pragmatic markers the correct tag could be assigned. Presumably, this POS-tagging

heuristic may be improved by the expansion of data, and only after implemented for the objectives of this investigation.

Our paper provides the theoretical basis of the relevant PM POS-tagging and the classification of PM-models for further linguistic elaboration. Anyway, the result of speech corpora annotation at the level of pragmatic markers can become a systematic description of PM as the inherent structural components of oral discourse. The description should be done considering PM functions, polyfunctionality, and possible "synonymic" relations, their formal grammar (and not only at the parts-of-speech level, but also, for example, at the level of predicative units [Bogdanova-Beglarian, Zaides 2019]), the specificity of their usage, and the possible correlation with speaker's characteristics, type of speech (monologue/dialogue) or communicative situation.

References

1. Baranov, A. N., Plungian, V. A., Rakhilina, E. V.: Guide to Discursive Words of Russian. Pomovskij i Partnery, Moscow (1993). [In Rissian].
2. Bogdanova-Beglarian, N. V.: On the Possible Communicative Interference in Cross-Cultural Oral Communication. Mir russkogo slova, 3, 93–99 (2018). [In Rissian].
3. Bogdanova-Beglarian, N. V.: Grammatical "Atavisms" of Pragmatic Markers of Russian Oral Speech. In: Glazunova, O. I., Rogova, K. A. (eds.). Russian Grammar: Structural Language Organization and Processes of Language Functioning, pp. 436–446. Moscow (2019). [In Rissian].
4. Bogdanova-Beglarian, N., Blinova, O., Martynenko, G., Sherstinova. T., Zaides, K.: Pragmatic Markers in Russian Spoken Speech: an Experience of Systematization and Annotation for the Improvement of NLP Tasks. In: Balandin, S., Cinotti, T., Viola, F., Tyutina, T. (eds.). Proceedings of the FRUCT'23. Bologna, Italy, 13–16 November 2018, pp. 69–77. FRUCT Oy, Finland (2018).
5. Bogdanova-Beglarian, N. V., Blinova, O. V., Martynenko, G. Ja., Sherstinova, T. Ju., Zaides, K. D., Popova, T. I.: Pragmatic Markers Annotation in Russian Speech Corpus: Research Problem, Approaches and Results. In: Selegej, V. P. (ed.). "Dialogue" Conference Proceedings "Computational Linguistics and Intellectual Technologies". Moscow, 29 May – 1 June 2019, 18(25), pp. 72–85 (2019a). [In Rissian].
6. Bogdanova-Beglarian, N. V., Blinova, O. V., Sherstinova, T. Ju., Troshchenkova, E. V., Gorbunova, D. A., Zaides, K. D.: Pragmatic Markers of Russian Everyday Speech: The Revised Typology and Corpus-Based Study. In: Balandin, S., Niemi, V., Tuytina, T. (eds.). Proceedings of the 25th Conference of Open Innovations Association FRUCT, pp. 57–63. Helsinki, Finland (2019).
7. Bogdanova-Beglarian, N. V., Maslova, Je. R.: Russian Contact Verbs in Oral Spontaneous Speech: Dictionary Volume and Functional and Semantical Diversity. In: Acta Linguistica Petropolitana. Proceedings of the Institute of Linguistics RAS, 3(15), pp. 115–135. St. Petersburg (2019). [In Rissian].
8. Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Baeva, E., Martynenko, G., Ryko, A.: Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech. In: SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811, pp. 659–666. Springer, Switzerland (2016).
9. Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G.: An Exploratory Study on Sociolinguistic Variation of Spoken Russian. In: SPECOM 2016. Lecture Notes in Artificial Intelligence, LNAI, vol. 9811, pp. 100–107. Springer, Switzerland (2016).

10. Bogdanova-Beglarian, N. V., Zaides, K. D.: Corpus of Monological Speech: Real and Formal Predictivity of Russian Oral Discourse Units. In: Kocharov, D. A., Skrelin, P. A. (eds.). Analysis of Spoken Russian Speech (AR3-2019): Proceedings of the 8th Interdisciplinary Seminar, pp. 11–16. St. Petersburg (2019). [In Rissian].
11. Bolly, C. T., Crible, L., Degand, L., Uygur-Distexhe, D.: Towards a Model for Discourse Marker Annotation: From Potential to Feature-based Discourse Markers. In: Fedriani, Ch., Sansó, A. (eds.). Pragmatic Markers, Discourse Markers and Modal Particles: New perspectives. Pp. 71–98. John Benjamins, Amsterdam (2017).
12. Burnard, L. (ed.): Reference Guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by Oxford University Computing Services, [Electronic resource], http://www.natcorp.ox.ac.uk/docs/URG/, last accessed 01/05/2020.
13. Campbell, N.: Speech & Expression; the Value of a Longitudinal Corpus. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation LREC 2004, pp. 183–186. ELRA, Lisbon, Portugal (2004).
14. Crible, L.: Discourse Markers and (Dis)fluency. Forms and Functions across Languages and Registers. John Benjamins, Amsterdam (2018).
15. Crible, L., Cuenca, M.-J.: Discourse Markers in Speech: Characteristics and Challenges for Corpus Annotation. Dialogue and Discourse, 8(2), 149–166 (2017).
16. Crible, L., Zufferey, S.: Using a Unified Taxonomy to Annotate Discourse Markers in Speech and Writing. In: Proceedings of the 11th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, pp. 14–22. London, UK (2015).
17. Kiseleva, K., Paillard, D. (eds). Discursive Words of Russian: Experience of Contextual and Semantic Description. Moscow, Metatext (1998) [In Rissian].
18. Kiseleva, K., Paillard, D. (eds). Discourse Words of Russian: Contextual Variation and Semantic Units. Moscow, Azbukovnik (2003) [In Rissian].
19. Fraser, B.: Pragmatic Markers. Pragmatics, 6(2), 167–190 (1996).
20. Heemant, P. A., Byron, D., Allen, J. F.: Identifying Discourse Markers in Spoken Dialog. In: AAAI 1998 Spring Symposium on Applying Machine Learning to Discourse Processing, pp. 44–51. The AAAI Press, California, Menlo Park (1998).
21. Hirschberg, J., Litman, D.: Empirical Studies on the Disambiguation of Cue Phrases. Computational Linguistics, 19(3), 501–530 (1993).
22. Lenk, U.: Marking Discourse Coherence: Functions of Discourse Markers in Spoken English. Narr, Tuebingen (1998).
23. Mann, W. C., Thompson, S. A.: Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text, 8(3), 243–281 (1988).
24. Plungian, V. A.: Why We Need the Russian National Corpus: Informal Introduction. Russian National Corpus, 2003–2005, pp. 6–20. Moscow (2005). [In Rissian].
25. Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A.: The Penn Discourse Treebank 2.0 annotation manual. Technical report, Institute for Research in Cognitive Science, 2007, [Electronic resource], https://repository.upenn.edu/cgi/viewcontent.cgi?article=1203&context=ircs_reports, last accessed 01/05/2020.
26. The Everyday Russian Language: Functioning Features in Different Social Groups. Bogdanova-Beglarian, N. V. (ed.). Collective Monograph. St. Petersburg (2016). (in Russian)
27. Samy, D., Gonzalez-Ledesma, A.: Pragmatic Annotation of Discourse Markers in a Multilingual Parallel Corpus (Arabic-Spanish-English). In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, Marrakech, Morocco, [Electronic resource], http://www.analedesma.es/wp-content/uploads/2010/04/doaingles.pdf, last accessed 01/05/2020.

28. Sanders, T. J. M., Spooren, W. P. M. S., Noordman, L. G. M.: Toward a Taxonomy of Coherence Relations. Discourse Processes, 15(1), 1–35 (1992).
29. Stojka, D. A.: The Dictionary of Reduced Forms of Russian Speech. Bogdanova-Beglarian, N. V. (ed.). St. Petersburg (2019). [In Rissian].
30. Shershneva, D. M.: Da as a Lexical and Functional Unit of Russian Speech. Studia Slavica XVIII, Tallinn, 270–278 (2015). [In Rissian].
31. Shibata, T.: Investigation of Language Living within 24 Hours. In: Linguistics in Japan, pp. 134–141. Moscow (1983). [In Rissian].
32. Shiffrin, D.: Discourse Markers. Cambridge University Press, Cambridge (1996).
33. Schourup, L.: Discourse Markers. Lingua, 107, 227–265. Elsevier, The UK (1999).
34. Turchanenko, V. V.: "Tam"-analysis: Functioning of a Unit in Russian Oral Spontaneous Speech. In: Proceedings of the XX Open Conference for Students-Philologists. St. Petersburg, 17–21 April 2017, pp. 60–65. St. Petersburg (2018). [In Rissian].
35. Université catholique de Louvain official website, MDMA – Model for Discourse Marker Annotation, [Electronic resource], https://uclouvain.be/fr/instituts-recherche/ilc/valibel/mdma-model-for-discourse-marker-annotation.html, last accessed 01/05/2020.
36. Verdonik, D., Rojc, M., Stabej, M.: Annotating Discourse Markers in Spontaneous Speech Corpora on an Example for the Slovenian Language. Language Resources and Evaluation, 41(2), 147–180. The Netherlands (2007).
37. Zakharov, V. P.: Corpus Linguistics: Teaching Aid. St. Petersburg (2005). [In Rissian].
38. Sound Corpus as a Base for Analysis of Russian Speech: Collective Monograph. Part I. Reading. Retelling. Description. Bogdanova-Beglarian, N. V. (ed.). St. Petersburg (2013). [In Rissian].