Pragmatic Markers in the Corpus "One Day of Speech": Approaches to the Annotation

Kristina Zaides[1], Tatiana Popova[2], and Natalia Bogdanova-Beglarian[3]

[1-3] Saint Petersburg State University, Saint Petersburg, Russia
kristina.zaides@student.spbu.ru, tipopova13@gmail.com,
n.bogdanova@spbu.ru

Abstract. The article describes the scheme of the annotation of pragmatic markers in the corpus of Russian everyday speech "One Day of Speech". Pragmatic markers are defined as special units in the speech that have only pragmatic function without any (or with 'bleached') lexical meaning. The annotation of pragmatic markers is usually performed manually due to the existing ambiguity of markers in different contexts. The typology of pragmatic markers includes different groups marked with special annotation tags. The annotation process was split into two stages since several issues of tagging of PMs arose. The main problems, which occurred during the annotation process, and the possible ways of their solution are also discussed in the research. The paper propose the improved methods of problem solving during the annotation of pragmatic markers applied to the corpus of oral speech, which can be useful for the linguistic annotation of any other levels of oral speech.

Keywords: Pragmatic Marker, Spoken Speech, Corpus of Everyday Speech, Corpus Linguistics, Corpus Annotation.

1 Introduction

The annotation of any corpus is the main linguistic tool in the corpus structure used for receiving correct search results and meta-information about texts and authors (speakers). Nowadays, the number of corpora of oral speech is growing exponentially around the world, so that an important and relevant issue in modern linguistics is being stated—to develop the basic principles of speech annotation, including such its units, which have never been described in the scientific literature before. Besides the well-known widespread levels of annotation, such as the marking of prosodic units, the part-of-speech tagging, the syntactic and semantic parsing, certain linguistic information should be tagged for some modern research tasks in communication studies, in particular, the discourse and pragmatic annotations. While the automatic annotation of a corpus material is implemented by the number of special parsers, the pragmatic annotation is still carried out manually because the instruments for such annotation are awaited to be produced in the near future [1, 2]. Moreover, many kinds of pragmatic annotation involves such patterns and details of speech that cannot be fulfilled by the automatic device, e.g., speech acts analysis or pragmatic markers re-

vealing. This paper presents the results of two stages of pragmatic markers annotation; therefore, we focus on the definition of the term *pragmatic marker* and its characteristics below.

A *pragmatic marker* (PM) is a relatively new term in the linguistics, introduced in this meaning by N.V. Bogdanova-Beglarian [3], which is used towards the particular speech units: words, expressions and phrases fulfilling different pragmatic functions in the discourse. The meaning of a term *discourse marker* (DM) do not coincide with the content of the term *pragmatic marker* since they describe different groups of discourse/pragmatic units, although both of them demonstrate the ability to structure the discourse but by different means. Discourse markers usually either navigates the paragraphs of a text or reveal time, causal, conditional and numerous other relations between the fragments being meaningful content words with a certain lexical meaning. A brief literature review, based on different researchers' understanding of DMs, can identify the specificity of these units more narrowly.

B. Fraser defines the DM as "a pragmatic class, lexical expressions drawn from the syntactic classes of conjunctions, adverbials, and prepositional phrases" [4]. The representatives of this class mainly "signal a relationship between the segment they introduce, S2, and the prior segment, S1" [Ibid.]. Basically, according to B. Fraser, they fall into two types: "those that relate aspects of the explicit message conveyed by S2 with aspects of a message, direct or indirect, associated with S1; and those that relate the topic of S2 to that of S1" [Ibid.]. The researcher characterizes the DM as "a linguistic expression only which: (i) has a core meaning which can be enriched by the context; and (ii) signals the relationship that the speaker intends between the utterance the DM introduces and the foregoing utterance" [Ibid.]. As it is explained, "they function like a two-place relation, one argument lying in the segment they introduce, the other lying in the prior discourse" [Ibid.]. Syntactically, DMs do not form a separate syntactic category. So-called *pragmatic markers* B. Fraser earlier identified as "structures and expressions which linguistically encode aspects of the speaker's direct communicative intention" [5] that "do not contribute to the propositional content of the sentence but signal different types of messages" [4].

D. Schiffrin argues that DMs do not fit completely into some linguistic category since their main function lies in adding to discourse coherence and providing "contextual coordinates for ongoing talk" [6]: DMs are "sequentially dependent elements which bracket units of talk" [Ibid.] which can be sentences, prepositions, speech acts, tone units, etc.

L. Schourup describes as DMs "conversational particles such as *well* and *oh*, parenthetical lexicalized clauses such as *y'know* and *I mean*, and a variety of connective elements in speech and writing, including *so*, *after all*, and *moreover* [7]. L. Schourup pointed out that "DMs are more often regarded as comprising a functional class that draws on items belonging to various syntactic classes" [Ibid.].

E. Traugott notices that DMs "allow speakers to display their evaluation not of the content of what is said, but of the way it is put together, in other words, they do metatextual work". [8]. The author supposes that DMs (in this work, the markers *indeed*, *in fact*, *besides* are investigated) go the grammaticalization path from the clauseinter-

nal adverbial through the sentence adverbial to the discourse particle, the subtype of the class of discourse markers [Ibid.].

In case of the annotation, the hesitation disfluencies sometimes are classified as discourse markers [9]. We suppose that such approach is not very productive since the hesitations can be detected automatically and usually treated as phonetically filled hesitation pauses and not as markers.

To the contrast, pragmatic markers derive from both content and functional words (nouns, verbs, adverbs, prepositions, etc.), and, during the process not only of grammaticalization, but also of pragmaticalization, they lose (in whole or in part) their lexical and/or grammatical meaning and get pragmatic one in some of their everyday speech usages. A content or functional word becomes a PM in a process of pragmaticalization: as a result, the role of its pragmatic component increases and a role of significant component decreases. The pragmatic function of a PM turns to be the leading one for a certain word, wherein the grammatical component can be still presented (for example, Aijmer reports that some units like *I think* are pragmaticalized, but they still have tense, aspect, and mood [10]). In this understanding, pragmatic markers such as *you know*, *I think*, *sort of*, *actually*, and *that sort of thing*, "have the function of checking that the participants are on the same wavelength or of creating a space for planning what to say making revisions, etc." [Ibid.]. PMs in the discourse approach "express speaker attitude to what has gone before, what follows, the discourse situation, and so forth" [8]. The further development of a pragmatic marker includes the lexicalization of a new meaning in everyday speech through its usage as the speech automatism and the assignment the special function to this marker in a certain communicative context [3].

The group of various discourse markers is formed by the words and phrases which are grammatically parts of speech, and the presence of this term, for the most part, points at the new approach of discourse analysis and constitute the opportunity to investigate relations of discourse more precisely. The words belonging to the group of discourse markers are different parts of speech, however, all of them have the ability to structure the pronounced speech or the written text. The range of pragmatic markers, as it is supposed here, consists of functionally "new" words – pragmatic markers, which have as their sources the full meant already existed lexemes, but for now are related to original words as homonyms. Thus, the class of discourse markers is largely the way of analyzing the text considering the functions of markers which manage it, whereas the group of pragmatic markers, it can be said, actually forms a new independent circle of functional words through their usages as speech automatisms, see examples below:

1. 'vidish/-te' (V, 2, Sing./Plur.) (*you see*) is used to attract the listener's attention to the subject of speech, but not to point at the item that both the speaker and the listener see (e.g., it is used during telephone conversation);

2. 'sejchas-sejchas-sejchas' (*one moment*) or 'minutochu-minutochku' (*wait a minute*) appear in the speech as hesitation pragmatic markers which forces the listener to wait a moment until the word, that is looking for by the speaker, is found.

The distinction between pragmatic and discourse markers is formed by the following points [11], [12]:

a) PMs are used in speech unconsciously, without any reflection, at the level of speech automatisms; DMs are put in text consciously, in order to structure its parts in a certain order;

b) PMs do not have (or have weakened, slightly vanished) lexical and/or grammatical meaning; they are almost completely "agrammatical"; DMs have full lexical meaning and grammatical paradigm;

c) PMs are not content or valuable units of speech, they have only functions; DMs have their own definite meaning as content words;

d) PMs are used essentially only in oral spontaneous speech and cannot be found in written texts (except for oral speech imitations, e.g., in modern plays or movies); DMs are presented both in written and oral texts equally;

e) PMs usually express speakers' attitude to the very process of speech production with all related difficulties being sometimes meta-communicative [13]; DMs always convey only speakers' evaluation of the subject discussed and its characteristics, but not of the text that they produce;

f) PMs are not included in the dictionaries in their functional diversity; DMs are the integral part of traditional lexicography as words, from the one hand, and are the subject of discourse related studies, from the other hand.

The typology of pragmatic markers is discussed in details in the section of presented paper which concerned the annotation of material and the system of tags.

2 Practical Significance of the Annotation of Pragmatic Markers

The results obtained by means of analysis of large corpus material allow clarifying traditional views of communication act using the identifying such discourse units—different types of pragmatic markers—which are uttered in speech in order to solve the particular communicative tasks. With the help of PMs, a speaker explicitly verbalizes his/her communicative intensions, attitude to the addressee, and appeals to the common with his/her interlocutors' perceptual basis. Because of the presence of PMs, the hearer can percept not only truth-conditional, informative level of speech, but also its structural level, as well as can understand how the communication itself functions: the beginning and the end of a speech act or an utterance, the search for words and omissions of lexemes, stressing of the important parts, any disfluencies and call to continue the interaction are marked.

The detailed elaboration of the spontaneous speech pragmatic annotation permits to create the algorithms of automatic checking of the annotation. Approximately each PM has its homonymic analogue which has a full meaning in sentence and is a part of speech, so that the distinction based on hesitation pause after the PM, e.g. 'sejchas', cannot be used since the hesitation break can follow the pronoun 'sejchas', as well as the homonymic PM, too. Each decision about the marking of the PM should be made taking into account the context near PM-"candidate". However, further annotation steps, for sure, will show that some kind of automatism can be presented in the tagging. The ability to implement in the natural language processing system the analysis

of functional and structural sides of language, for its part, will contribute to the artificial perceptual basis forming. The modeling of realistic speech dialogues "human–computer/robot/machine" interfaces, that is the most relevant issue in robotics and artificial intelligence development, will be also possible to improve.

The receiving of a full inventory of pragmatic markers of oral speech is also important in such applications as linguodidactics and translation practice. In particular, the introducing of the natural spoken speech materials into textbooks for the foreign students is essential for training them to understand Russian fluent speech and to avoid plenty of communicative failures. PMs that are used by the native speakers easily and naturally, at the level of speech automatisms, do not prevent to perceive the meaning of a message, and leave beyond the frame of their perceptual field [14]. These markers fall into the perceptual field of foreign speakers and can cause great challenges in communication using a non-native language.

Besides, the typical range of pragmatic markers could be individual for the particular speaker; consequently, this information may be used for the identification of diagnostic features of some age, gender, social or psychological group during conducting linguistic or forensic expertise of oral speech audio recordings.

As one could see, the annotation of the pragmatic markers is required for different linguistic, scientific, and practical needs. This study presents one of the possible ways to organize the process and to develop the methods of the pragmatic annotation that can be applied to analysis of different corpora data.

3 Research Material

The research was carried out on the material from the corpus of Russian everyday speech "One Day of Speech" (ORD), which is one of the most representative resources for the analysis of Russian oral spontaneous dialogic and polylogic speech. The ORD corpus contains 1,250 hours of speech files recorded from 128 informants, which are native speakers of Russian, living in St. Petersburg, and more than 1,000 of their interlocutors, all of them represent various social groups [15, 16]. The records were made using a method of the 24-hours recording of speech day [17] and, after recording, received material were transcribed in the ELAN linguistic annotator. The ELAN files contain several main levels of annotation: transcribed phrase, speaker who pronounced the particular phrase, his/her voice characteristics, events in real life that accompanied the recording, phonetic and phrase commentaries, notes, and episode to which this communicative situation belongs [18].

The pilot subcorpus balanced by gender and age was created for the first annotation of pragmatic markers. The annotation of 12 episodes of corpus speech taken from 12 recordings of different speakers was performed by the group of four annotators independently one from another; total duration amounts 1 hour 46 minutes, 10259 word tokens. For the annotation, additional levels in the ELAN files were made:

- ***PM***, which contains the pragmatic marker in its orthographical form;
- ***Function PM***, that indicates the functions of the PM;
- ***Speaker PM***, which marks the speaker's code;

- ***Comment PM***, that reflect other commentaries connected with the specific PM usage.

4 Development of the System of Tags and Stages of the Annotation

For the annotation, the special system of tags was elaborated that included references to the groups of pragmatic markers already described in the scientific literature [3], [12], [19]. Briefly, for the marker from each group the function manifested in its name is main, but there are plenty of markers that have several functions, i.e., share the common feature of multifunctionality. In the typology of tags below that was developed matching with the system of pragmatic markers itself, the cases of marker polifunctionality are specially commented.

1. APPR — marker-approximator that expresses speaker's uncertainty and hedge:
   - *ne znayu // *P vidish' / chego-to Kirill% govorit / chto gips luchshe / yesli (e-e) / tsement bystro vysokhnet / v malen'kikh dyrkakh* ***kak by*** */ yesli tsement bystro vysokhnet / to (:) on ne budet prochnym* [S1];

2. DEICT — deictic marker that points at something vague and consists of 3 elements, two of which are 'vot':
   - *nu v obshchem defekt kishki / kogda (e) na nej takoj otrostochek / kak byvaet vot (...) (e-e) v venakh / kak appendiks /* ***vot takoj vot*** *kakoj-to tam* [S130];

3. ZAMEST-PR — replacement marker for the whole set of enumeration or its part:
   - *Natasha% / vy uzhe otpustili etogo / () Alekseya%(:) / Maksima% / i* ***vsego prochego*** *? *P vot* [S19];
   - *ya govoryu ya togda v devyati tri... tam k devyati pyatnadtsati pridu / poka* ***to syo****...* [S124];

4. ZAMEST-CHR — replacement marker for someone's speech, e.g., 'bla-bla-bla':
   - *a / my s toboj zhe byli / pomnish' / Nastya% i Katya%. Aaaa… Kat'ku% ya videla paru raz v universitete / nu / my s nej poskol'ku ne obshchalis' / postoyali / «privet-privet» tam /* ***bla-bla-bla*** [this example is borrowed from the Russian National Corpus];

5. XEN — quotational marker which marks someone else's speech before its appearance in the utterance:
   - *nikto poka nichego ne mozhet vnyatnogo skazat' / vse tol'ko razvodyat rukami / (e) i* ***govoryat*** */ nu / sochuvstvuyu tipa mol / *P namekayut chto(:) prosto da / oforml... oformlyaj novuyu strakhovku i(:) (...) zhivi spokojno* [S110];

6. MET — meta-communicative marker which fulfills meta-communicative function: the establishment of a contact and understanding between speakers and the speaker's reflection on his/her own speech:

- *nu i Vadik% priezzhaet / *P i oni yemu govoryat **slushaj** chuvak my tebe vsyo otremontirovali / *P tol'ko my tebe koroche (...) (e-e) v bak (...) vmesto(:) (e) dizelya devyanosto vos'moj zalili* [S72];
- *nu Andrej% / togda **vy smotrite** / znachit ya do devyati budu (...) nu (e) telefon vyklyuchu / i otvechat' ne budu / to est' ya prosnus' gde-to v devyat' s kopeechkami / budu uzhe (e) min... vy uzhe v eto vremya budete ekhat'* [S123] (during telephone conversation);

7. NAVIG — navigational marker which serves as structuring device;
   - *nu i (...) a do etogo proverili / zheludok vsyo khorosho / a tut polosnaya operatsiya / **vot** eto ya vsyo ... / **vot** eto pervaya chast' Kazani u menya byla normal'naya / a vtoraya chast' (...) **vot** ya vot na etikh samykh zvonkakh nepreryvnykh* [S130] (the marker 'vot' also fulfill the hesitative function here);

8. SEARCH — searching marker that helps the speaker to find the word or expression he/she is looking for:
   - *no pri etom b***d' / *P chuvstvuyesh' takoe na***j opustosheniye ! vnutri katarsis chuvstvuyesh' // **kak eto** b***d' () Gracheva% govorila nado // *V ochishcheniye cherez stradaniye* [S15];

9. REFL — reflexive marker which express speaker's reaction to what is said:
   - *v itoge my vyzyvali kakogo-to traktorista // *P # khorosho chto nashli vy traktorista // # ugu // *P **ili yeshchyo chego-to takoye** / i koroche vytaskivali Vadika% ottuda // @ ugu* [S72 and W1];

10. RHYTHM — rhythm-forming marker that attaches rhythm to the utterance:
    - *vot sejchas uzhe batarei dali / uzhe on bystro vysokhnet // a tak by vot / vot kogda dozhdi shli / **vot** khorosho bylo by zadelat'* [S1];

11. SELFCORR — marker of self-correction:
    - *yarkaya solnechnaya pogoda // govorit' mozhno? tak byl yark… ∫ **eto samoe** ∫ byl ∫ iyul'skij den' / vot / nebo bylo chistym / bezoblachnym / solntse ∫ svetilo* (this case is taken from the corpus "Balanced Annotated Collection of Texts", another corpus of oral speech, created by the group of the same linguists as creators of the ORD-corpus);

12. START — marker of the beginning of an utterance or the process of speech production:
    - *ditya moyo / **znachit tak** // *P ta(:)k ? // v etom (...) (m-m) v sentyabre / budet tut vsyo vot tak / *V a v oktyabre / a ... # analogichnaya situatsiya budet na sleduyushchej nedele // # da // @ a ... / a ...* (the marker 'znachit tak' also fulfill the hesitative function here);

13. FIN — marker of the end of an utterance or the process of speech production:
    - *nu ponyatno delo / nu y**ta / a(:) da tebe voobshche / dazhe zakonnyje vykhodnyje mogut ne dat' / da ? **ya dumayu*** [S110] (the marker 'ja dumaju' also fulfill the hesitative function here);
    - *tak / **nu vsyo** / ya ostanavlivayu zapis' / potomu chto eto pustoye / slushat' eti kliki / vsyo ravno ya nichego bol'she ne skazhu / vse uzhe spyat* [S123];

14. HES — hesitation marker:

- ***nu tam*** *(...) sil'no deshevle ne bylo / potomu chto ya () zdes' kak by / oni vsyo ravno ekhali* [S103].

The special guideline for the annotators was elaborated. At the first stage of the annotation process, the guideline included the tags consisted of several first letters of particular function (named, as it was showed above), the instructions, such as to write the marker orthographically, to put the tags in the alphabetic order, noting first the main function(-s) of PMs and second the additional function(-s), to separate the repeated markers one from another (do not place them using the hyphen) as well as the description of the process of new level creation in the ELAN program. The possibility to point the new function of a marker was also provided to the annotators. Moreover, before the first try of the annotation, already revealed and described markers were illustrated with an examples from the corpus with an indication of possible functions they can perform. Fig. 1 shows a fragment of the table which was made to help the annotators. The table includes the marker, its structure (one or more words form the marker), examples of usage in speech in the main and additional functions, the tag, items per million value counted in previous researches, the tendency to use it in dialogues or in monologues. In addition, this table contains the link to the document with so-called "described in dictionaries" usages of homonymic to the pragmatic markers expressions. We believed that by producing such table we assisted the annotators to detect the possible pragmatic functions of markers faster and easier.

PM	N	Example	Function	Tag	Add. functions	Examples	IPT	Dialogue/monologue	Link to the dictionary
vidish/vidite	yes	ne znayu / **vidish** / chego-to K	MET	M	REFL	REFL: nam obyasnyajut / **ne vidi**	329(450 615)	D	https://drive.google.com/ope
<po>smotri/<po>smotrite	yes		MET	M	START NAVIG	START: **smotrite** / vot / vot eta shtuchka / da / vot eta vot; NAVIG: Dasha% / ty voobshche prostranstvenno myslish ili net ? **nu vot smotri** // ja v plane tebe risuyu / vot zritel'nyj zal	423(450 615)	D	https://drive.google.com/ope
<po>slushay/<po>slushayte	yes	**slushay** / davay ya tebe perez	MET	M	XEN START NAVIG FIN HES	XEN : i on blin mne napisal **slushay** sorri ne mogu; START: Gena% / Gen% // **vot poslushay** menya vnimatel'no pozhaluysta; NAVIG: nu eto khorosho / **slushayte** ! *P ya zavtra vecherom vozmozhno budu v tekh krayakh ... FIN : bez ochkov strigyot / vot khrabraya / **slushay** ! HES: a nu da / pust' zvonit // *P znachit (e) **slushay** smotri ... *V	374(450	D	https://drive.google.com/ope

Fig. 1. A fragment of the table of described pragmatic markers.

After the first stage of the annotation, it turned out that the inter-annotator agreement counted with the help of Kohen's Kappa coefficient (the formula see in [1]) was very low. The best agreement between experts was achieved only for three groups of PMs, i.e., quotational markers, meta-communicative markers, and reflexives. Therefore, the decision to improve the guideline for the annotators was made. Fig. 2 presents a fragment of the table with all possible variants of one marker that can be united by its main type.

A	B	C
Invariant	Variant	Key word for search
prikin'	prikin'	priki...
	prikin'te	
zatseni	zatseni	zatseni
	zatsenite	
glyan'	glyan'	glyan'
	glyan'te	
zamet'	zamet'	zamet'
	zamet'te	
chto yeshchyo	chto yeshchyo	yeshchyo, skazat'
	chto yeshchyo skazat'	
	chego-to yeshchyo khotela skazat'	
	chto yeshchyo b skazat'	
	chto yeshchyo by skazat'	
	chto-to yeshchyo khotel skazat'	
	chego-to yeshchyo khotel skazat'	
	chto-to yeshchyo khotela skazat'	
koroche	koroche	koroche
voobshche	voobshche	voobshche
	voobshche govorya	
sobstvenno	sobstvenno	sobstvenno
tam	tam	tam

Fig. 2. A fragment of the table of variants of pragmatic markers.

This step allows annotating markers automatically and to narrow down the variants to one basic construction. Such variety of grammatical forms reflects the process of pragmaticalization without grammaticalization, as well as the ability of markers to combine with other pragmatic or "meaningless" (functional) components of speech (particles, interjections, conjunctions, etc.), and exists for the all the markers considered in the research: 'eto', 'eto samoje', 'kak jego', 'ne znaju', 'sejchas', 'minutu', 'sekundu', 'tipa', 'vrode', 'kak by', 'takoj', 'bla bla', 'lia lia', 'ili kak eto', 'ili kak jego', 'ili chto yeshchyo' and many others.

For prepare the next stage of the annotation, it was determined, first, not to reduce all the variants of one marker to one basic structure, leaving, during the annotation, the PM in the form in which it was presented in speech, which saved the variety of markers structure; and second, to shorten the list of PMs' functions, so that exclude the most ambiguous cases which revealed total annotators disagreement. Third, the opportunity to list the main and additional functions in a free order was given to the annotators, because of mentioned in the introduction of this paper the multifunctionality of PMs.

At the second stage of the annotation process, the new guideline included fewer tags as some of them were grouped (e.g., the group of markers of a boundary (G) unified previous existed start, final and navigational markers), and all the tags were cut to one letter in order to make the annotation process less time-consuming. The annotation of the same files was performed by the same group of annotators inde-

pendently one from another; they also had been asked to use the new instructions and the system of tags. The analysis of inter-annotator agreement showed the increased level of agreement—up to Kappa=0,51, especially for two annotators who are the authors of presented article [20]. It means that the development of the annotation scheme discussed above, the guideline and the tables of variants improves the results of annotation. The elaborated procedure of the annotation of PMs is supposed to be widely used in the investigations involving the similar methods and data.

However, the process of the annotation cannot be lead without any issues. The human factor and the subjectivity cannot be absolutely removed from the language analysis, but there are certain problems of the annotation that corpus linguists might deal with. The ways of solution of this kind of annotation problems are described in the next section.

5 Main Annotation Problems of Corpus Material and Ways of Their Solution

During the process of the manually performed annotation of pragmatic markers, the group of annotators, including the authors of this research, confronted several problems involving the functions of PMs, the difference between a PM and a homonymic expressions (see also: [21]), the human factor, the prosodic features of speech, etc. These problems and the possible methods of their solution will be discussed here one by one.

5.1 The Syntagmatic Division of Spontaneous Speech

One of the most important issue was the syntactic and intonation division of speech in syntagmas that cannot be clearly defined in some cases. The addressing of such ambiguity is relevant for the definition of the PM ‘vot’ functions that performs as a marker of start or final of a phrase or speech part, according to its pre- or postposition:

- *da / poka vot () Marina% ne sde... da / i ne posmotrit i ne otfotografiruyet // *P* ***vot*** *// *P vsyo // pozhalujsta // vsego dobrogo / do svidaniya* [S19];
- *ya sejchas pozvonyu Marine% / i vyyasnyu // delo v tom chto / k vam sobiralas Marina% yekhat' Zhdanova% // ne ne ne ne ne ne // *V Marina% Glukhareva% // *N* ***vot*** */ *P i (:) (e-e)* ***vot*** */ ya vyyasnyu / poyedet ona segodnya ili zavtra k vam* [S19];
- *postoyannye koroche / bunty kakiye-to / sobraniya kakikh-to partij raznykh / politicheskikh / tam vsyakikh // tam b***d' partiya na partiyu / koroche /* ***nu vot*** *// *P zastrelili / odnogo na ulitse / sluchayno // *P (e)* ***vot*** */ *P vtoroj spilsya / a glavnyj geroj / koroche / u nego umerla eta devushka* [S15];
- *moj Seva% byl (...) v techeniye (...) tryokh / chetyryokh dnej v reanimatsii //* ***vo(:)t*** */ sejchas ya yedu / (...) prosto poyedu / net / nu yego uzhe vypisyvayut v chetverg / poyedu povezu / on menya poprosil / chto privezti* [S130].

The pause after the marker means that the topic shift takes its place in the utterance. This unit can be classified as the PM of start due to its position in the beginning of a new phrase. However, it is not defined in these examples, whether the marker attributes to the new topic or discourse fragment itself or the marker closes the previous speech segment with the meaning of conclusion.

The annotation of the start, navigational and final markers caused disagreement at the first stage of the annotation. It is obvious that all these markers share one common function—the marking of a boundary, with the possible change of topic, the communication strategy, the conditions or a manner of speech producing, etc. However, practically, in speech several markers can serve merely in one definite function, e.g., 'znachit tak' for the marking of start or 'vsyo' for the marking of the end of speech. Despite this, the most commonly used markers of this type—'vot' and 'koroche'—tend to appear in different positions in phrases, not having only one preferable place of occurrence. Therefore, the new annotation rules were implemented at the second stage. As a result of the annotation, the receiving of a complete list of markers, as well as their functions, which all the annotators could agree with, the main goal of the researchers was achieved. The variety of "boundary"-tags resulted in inter-annotator disagreement, which showed the disadvantages of tags system. The reduction of tags by clustering them into groups led to making the functions more identifiable. Thus, one tag "G" was produced to unite different tags of boundary markers: "START", "FIN", and "NAVIG". The specifics of each case of boundary PMs will be described during the qualitative analysis of the material after the annotation of all corpus data. Moreover, the distinctive features of different types of boundary PMs are planned to elaborate.

5.2 Pragmaticalization as a Continuing Process

The annotation of pragmatic markers is complicated by the live processes existing in oral spontaneous speech, i.e. grammaticalization and pragmaticalization. Thus, the different degrees of pragmaticalization, a closeness of a unit to the PM class, can be distinguished, e.g.:

- *nu ya sproshu // yesli tsementa ne budet / togda ya gips voz'mu // # v malen'kikh dyrkakh / *P dlya bolshikh dyrok gips ne podkhodit / a () dlya bolshikh dyrok podkhodit tsement // *P* ***ya dumayu*** *// nu ya ne znayu / *P chto takoye bolshaya dyrka // *P v takom-to vot sluchaye* [W1 and S1];
- *nu ponyatno delo / nu y**ta / a(:) da tebe voobshche / dazhe zakonnyje vykhodnyje mogut ne dat / da ?* ***ya dumayu*** *// *P u menya tam podnakopilos' etikh samykh / neispol'zovannogo otpuska / da / poetomu ya i ispol'zuyu* [S110];
- **P kak to tak ona korotkovata nemnozhko poluchilas' // vrode yeshchyo odin shkaf prositsya // *P kholodilnik ne vkhodit a / tak mesto svobodnoye est' // *P* ***ne znayu*** [W1];
- *ponyatno / ya prosto khochu vam skazat' / ya ne ... / vernej sprosit' / snachala dlya nachala / potom uzhe skazat' / *V po povodu etoj programmy (:) / vot ona (...) nastol'ko zamedlyayet rabotu komp'yutera / *P*

*chto vot (e-e) / nu mne prikhodyat gigantskiye fajly / **ya ne znayu** chto tam / eto samoye / no ...* [S19].

It seems that the first two examples shows already pragmaticalized usages of VP 'ja dumaju' that only marks the end of a sentence and do not contribute anything to the content. These PMs also reflect the speaker's hesitations and serve as means of a hedge, as well as the unit 'ne znaju' in the third case. It should be noted that there is a possible interpretation of these markers as not fully pragmaticalized, but only taken a pragmaticalization path ones, that are mostly potential, than real, PMs.

The last phrase is truncated, but by the presence of the hesitation ('eto samoje') we can conclude that the speaker does not know what to say next and how to describe the problems with the computer in more detail. It leads us to the assumption that 'ja ne znaju' in this case is the hesitation PM used in preparing, after all, unsuccessful tries to continue the speech production. However, this construction can be also examined as a meaningful sentence, just left by the speaker and not extended further. Since that, the annotation of such case is ambiguous, from our perspective. The variability of analysis is not only possible, but also necessary for dealing with PMs. Perhaps, the annotation of a wider data allows solving the issue of annotating of such phenomena; the experts have to create the acceptable limits up to which the meaning of a lexeme is identifiable and the unit is still not a marker, otherwise, it should be considered a pragmatic unit having only function in oral discourse.

5.3 Main and Additional Functions of PMs

The dynamic aspect of producing speech causes certain difficulties in function attributions: the problem of determination of the main and the additional functions of PMs and their difference is also complicated by permanent changing the PM place in phrases. For instance, in phrases:

- *nu tam v osnovnom sovetskuyu chital / znayesh literaturu // nashu **tam** / a(:) ! vperyod k kommunizmu !* [S15];
- *nu ya pytayus // no tam zhe kak prosto kak by () konkurentsiya // *P // to est' **kak by** dazhe yesli ya podnimayu ruku / to yeshchyo ne ... // *V nu ya v printsipe pochti na kazhdom podnimayu / no menya prosto ne vsegda sprashivayut* [S27]

is not possible to identify precisely whether the approximation or hesitation is the main function of PMs 'tam' and 'kak by'. The role of this PM in the discourse lies in the fact that they help the speaker to have a little pause in speech structuring and give him/her an opportunity to express the idea approximately, without further description. To determine which function is predominant seems quite impossible here (see also: [21, 22]).

At the second stage of the annotation, we rejected the difference between the main and the optional functions since the inter-annotator agreement in their annotation was very low. Henceforth, beyond the annotation of all the functional sets of a particular marker, it will be possible to determine the criteria of function domination and increasing prominence.

The tagging of a rhythm-adding function was also uncoordinated and inconsistent. The findings of the investigation [23] shows that there are rhythm-forming markers which organize spontaneous speech into isochronous structures:

- *vot sejchas uzhe batarei dali / uzhe on bystro vysokhnet // a tak by vot / vot kogda dozhdi shli / **vot** khorosho by bylo zadelat'* [S1];
- *nu i (...) a do etogo proverili / zheludok vsyo khorosho / a tut polosnaya operatsiya / vot eto ya vsyo ... / vot eto pervaya chast' Kazani u menya bila normalnaya / a vtoraya chast' (...) vot ya **vot** na etikh samykh zvonkakh nepreryvnykh* [S130].

We suppose that in the cases (in bold) the rhythm-forming function is realized. The first PM 'vot' in the first example functions as the boarder-marker, the second operates in the field of hesitation only, the third presumably is a particle for new information actualization, and the last forms the rhythm and the rate of the utterance, which are supported by the repetition of 'vot'. The second case also shows a frequent usage of 'vot', one of which can be regarded as the rhythm-forming PM in the last position. However, it is possible that all these markers are the individual way of hesitating of the particular speaker.

5.4 Chains of Markers or One Marker?

The cases of neighborhood of pragmatic markers are quite frequent in the spontaneous dialogues and monologues. It raises the question of what should be considered as a chain of markers and what—as a new complex PM with another function. D. Verdonik, M. Rojc, and M. Stabej [9] analyze discourse markers in the corpus of Slovenian telephone conversations TURDIS and try to deal with cases of markers collocation, describing the most widespread chain of markers at the beginning of an utterance. We suppose that the PM which forms one intonation unit and fulfills one function is one integral marker, otherwise it is the chain of different markers following one another with a hesitation graduation. However, in case of hesitation PMs it is difficult to decide whether the function is intensifying or actually is equally shared by the sequence of markers:

- *pod triumfalnuyu_arku$ tam koroche // **vot tipa** (...) Kebern% ch... nu(:) rasskazyval // *P ya nachal chitat' / ya tak_skazat'(?) sovsem drugoye prochital / chem chto on mne rasskazyval* [S15] (hesitation and approximation marker(-s));
- *vchera my s na... s Nadey% vykhodim s raboty // *P ona menya prosit / u vas est' tam telefon (e-e) Glukharevoy% ? ya govoryu da // *P nu i **znachit tam** (...) nakhozhu / diktuyu yej* [S19] (boundary, hesitation and approximation marker(-s));
- *tam to delay / **tam kak by tam** zadaniye // chego-to kak-to ustayu bezumno na samom dele // *P prosto voobshche kak by / v printsipe i *P ne to chtoby ya pryamo tut tak umatyvayus // da ? no vot real'no ochen' ustayu* [S27] (hesitation and approximation marker(-s));
- *nikto poka nichego ne mozhet vnyatnogo skazat' / vse tol'ko razvodyat rukami / (e) i govoryat / nu / sochuvstvuyu **tipa mol** / *P namekayut chto(:) prosto*

da / oforml... [S110] (approximator or quotational marker and quotational marker ‘mol’, probably not the PM since it is used in written texts);

- *nu smotrite / *P v poldesyatogo / **tak znachit** smotrite Andrey% / ya tut pogovoril / (...) yeshchyo s lyud'mi / mne rasskazali sleduyushcheye / chto vot eto staraya tak nazyvayemaya* [S123] (hesitation and boundary marker(-s)).

The examples above show one of the most interesting tendency of spontaneous speech, which opposes the principle of language (and speech) economy—the language redundancy. The repeated markers also present a challenge for the annotators given that they may be interpreted as one marker since they have the same function or as two or more repeated markers as words:

- *u vas segodnya prikhod budet // *P tak / **minutochku minutochku** / Gul'% // *P tak / ya sejchas pozvonyu Marine% / i vyyasnyu // delo v tom chto / k vam sobiralas' Marina%* [S19];
- **P **tak tak** / **tak tak tak** / *P kto(?) *P privetik* [S117].

However, the existence of non-one-word markers cannot allow using the constituent criteria—a word equals a PM—during the annotation. To solve the issue “one or more markers” we plan to investigate the frequency of such series of PMs in the speech corpus, which can clarify their language status. At this stage of the annotation, only minimal structures are annotated, thereafter the cases of markers combination will be examined more precisely.

The inversion in Russian is one more problem for the automatic annotation of PMs:

- *(e-e) eto dejstvitel'no tak... poka ne ponyal / tak kak eto mne rasskazyval chelovek / kotoryj nichego ne ponimayet // nu vot **v samom etom** *N / prosto skazal / kak eto est' // poetomu elektriki mestnyje / vot troye / s kem ya pytalsya cherez tret'ye litso svyazatsya / vse otkazalis' / potomu chto oni skazali tak / *V yesli sdelat' vsyo eto vser'yoz / to eto dorogo* [S123].

This issue is solved by the containing the list of the possible PMs variations, even performed automatically by combinatorial algorithms.

6 Conclusion

The annotation of pragmatic markers is still a great challenge for the researchers since this is mainly manual process, difficult to automation, which creates the theoretical and practical issues concerning the understanding and the typology of PMs, the definition of their functions, and the investigation of oral unstructured human discourse. In the article, the process of the first annotation of pragmatic markers of Russian spoken speech was fully described, including two stages of the annotation, advantages and disadvantages of proposed approach to the pragmatic level analysis. The annotation concerned the pilot subcorpus, but the annotated material will be expanded. The presented problems of the annotation allowed us to elaborate the guideline for the annotators and the list of tags in such way that the inter-annotator agreement became higher. We state that the inclusive automatic tagging of PMs in oral speech cannot be performed for now, however, the automatic check of the annotation, after obtaining

the full list of PMs' variations, to avoid the human factor of missing markers is necessary. The fuzziness and ambiguity of spontaneous speech are significant issues in the NLP-tasks, and the future research might develop to overcome the multifunctionality of some PMs during the annotation process.

Acknowledgement. This research was supported by the Russian Science Foundation, project № 18-18-00242 "Pragmatic Markers in Russian Everyday Speech".

References

1. Leech, G.: Adding linguistic annotation. Wynne, M. (ed.). Developing Linguistic Corpora: a Guide to Good Practice. Oxbow Books, Oxford (2005).
2. Archer, D.: Corpus annotation: A welcome addition or an interpretation too far? Tyrkkö, J., Kipiö, M., Nevalainen, T., Rissanen, M. (eds.). Outposts of historical corpus linguistics: from the Helsinki corpus to a proliferation of resources. Studies in variation, contacts and change in English eSeries (2012). URL: http://www.helsinki.fi/varieng/series/volumes/10/archer/.
3. Bogdanova-Beglarian, N. V.: Pragmatemy v ustnoj povsednevnoj rechi: opredelenie ponyatia 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], 3(27), 7–20 (2014). (in Russ.).
4. Fraser, B.: What are discourse markers? Journal of Pragmatics, 31(7), 931–952 (1999).
5. Fraser, B.: Commentary pragmatic markers in English. Estudios ingleses de la Universidad Complutense, 5, 115–127 (1997).
6. Schiffrin, D.: Discourse markers. Cambridge University Press, Cambridge (1987).
7. Schourup, L.: Discourse markers. Lingua, 107(3–4), 227–265 (1999).
8. Traugott, E.: The role of the development of discourse markers in a theory of grammaticalization. Paper presented at the 12th International Conference on Historical Linguistics, University of Manchester, August 1995. URL: https://www.researchgate.net/publication/228691469_The_role_of_discourse_markers_in_a_theory_of_grammaticalization.
9. 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 (2007).
10. Aijmer, K.: Pragmatic markers in spoken interlanguage. Nordic Journal of English Studies, 3(1), 173–190 (2004).
11. Bogdanova-Beglarian, N. V., Filyasova, Yu. A.: Discourse vs. pragmatic markers: a contrastive terminological study. In: 5th International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2018, SGEM2018 Vienna ART Conference Proceedings, 19-21 March, 2018, vol. 5, pp. 123–130 (2018).
12. Bogdanova-Beglarian, N. V.: O vozmozhnykh kommunikativnykh pomekhakh v mezhkul'turnoj ustnoj kommunikacii [On the possible communicative barriers in intercultural oral communication]. Mir russkogo slova [The World of Russian Word], 3 (2018) (in print). (in Russ.).
13. Zaides, K. D.: Metakommunikativnyje vstavki v russkoj ustnoj spontannoj rechi na rodnom i nerodnom jazyke [Meta-communicative insertions in Russian oral spontaneous speech of native speakers and foreigners]. Kommunikativnyje issledovanija [Communication Studies], 3(9), 19–35 (2016). (in Russ.).

14. Riehakainen, Ye. I.: Vzaimodejstvie kontekstnoj predskazuemosti i chastotnosti v processe vospriyatia spontannoj rechi [The Interaction between Context Predictability and Frequency in the Process of Perception of Spontaneous Speech (on the Material of the Russian Language)], doctorate thesis, St. Petersburg. (2010). (in Russ.).
15. Bogdanova-Beglaryan, N., Asinovskiy, A., Blinova, O., Markasova, Ye., Ryko, A., Sherstinova, T.: Zvukovoj korpus russkogo yazyka: novaja metodologia analiza ustnoj rechi [Sound Corpus of the Russian Language: a new methodology for analyzing the oral speech]. In: Shumska, D., Osga, K. (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, pp. 357–372, Kraków (2015). (in Russ.).
16. Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G.: Linguistic features and sociolinguistic variability in everyday spoken Russian. In: SPECOM 2017, LNAI, vol. 10458, pp. 503–511. Springer, Cham (2017).
17. Russkij jazyk povsednevnogo obshhenija: osobennosti funkcionirovanija v raznyh social'nyh gruppah [Everyday Russian Language in Different Social Groups]. Collective monograph. Bogdanova-Beglaryan, N. V. (ed.). LAJKA, SPb (2016). (in Russ.).
18. Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., Sherstinova, T.: 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. 57292009, pp. 250–257. Springer, Berlin-Heidelberg (2009).
19. Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G., Baeva, E.: Towards a description of pragmatic markers in Russian everyday speech. In: LNAI, vol. 11096: Speech and Computer. 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18-22, 2018, Proceedings, pp. 42–48. Springer, Leipzig (2018).
20. Bogdanova-Beglarian, N., Blinova, O., Sherstinova, T., Martynenko, G., Zaides, K.: Pragmatic markers in Russian spoken speech: an experience of systematization and annotation for the improvement of NLP tasks. In: Balandin, S., Salmon Cinotti, T., Viola, F., Tyutina, T. (eds.). Proceedings of the 23rd Conference of Open Innovations Association FRUCT. Bologna, Italy, 13–16 November 2018, pp. 69–77. FRUCT Oy, Finland (2018).
21. Crible, L., Cuenca, M.-J.: Discourse markers in speech: characteristics and challenges for corpus annotation. Dialogue and Discourse, 8(2), 149–166 (2017).
22. 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, London, UK, pp. 14–22 (2015).
23. Bogdanova-Beglarian, N. V., Kisloshchuk, A. I., Sherstinova, T. Ju.: O ritmoobrazujushchej funkcii diskursivnykh jedinic [On rhythm-forming function of discourse markers]. Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologija [Perm University Herald. Russian and Foreign Philology], 2(22), 7–17 (2013). (in Russ.).