Pragmatic Markers of Russian Everyday Speech: the Revised Typology and Corpus-Based Study

Natalia V. Bogdanova-Beglarian[1], Olga V. Blinova[1], Tatiana Y. Sherstinova[2,1],
Ekaterina V. Troshchenkova[1], Daria Gorbunova[1], Kristina D. Zaides[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; {st058398, st017075}@student.spbu.ru

***Abstract*—Pragmatic markers (PMs) mainly have an influence on a pragmatic aspect of communication and are mostly devoid of their own referential meaning. These markers are indispensable elements of oral communication in any language. The article suggests a typology of pragmatic markers for Russian everyday speech that includes 10 basic types. The frequency study for the use of various marker types is carried out on the basis of two representative speech corpora – a corpus of Russian Everyday Speech "One Speech Day" (ORD) and "Balanced Annotated Collection of Texts" (SAT). Preliminary data about PM distribution in dialogues and monologues was obtained and the article describes the main difficulties one comes across while annotating PMs according to our methodology. The main requirements for creating a Dictionary of Pragmatic Markers are enumerated. The paper indicates the scope of pragmatic markers and further prospects for their use, which includes (but not limited to) datasets labelling for voice assistants and speech recognition systems development.**

I. Introduction

Pragmatic markers are a mandatory component of oral communication in any language [1] – they allow to break the discourse flow into fragments, help the speaker to establish correct relationship with the interlocutor, help to convey the speaker's stance and perform many other pragmatic functions. Therefore, it is largely PMs that are responsible for the effectiveness of communication. However, unlike other lexical units that are well-represented in numerous dictionaries, lexicographic description of pragmatic markers for many languages leaves much to be desired. This tend to be a problem for parsing natural speech in everyday communication, because the units parsed can't get the right interpretation, as there are no such patterns inside speech recognition systems.

This article focuses on the study of pragmatic markers in Russian everyday speech. The results obtained can be useful not only for theoretical studies of PMs, including comparative ones, but can also be applied in the development of didactic materials and textbooks on Russian oral discourse, as well as find numerous applications in the sphere of the speech synthesis and recognition. On the other hand, using different types of ML approaches, the results can be the first stage on the way of dealing with the most common problems in recognition of pragmatic units: incoherence, less grammar restrictions, unclear speech interpretations.

II. Some Theoretical Issues on Pragmatic Markers: the Case of Russian

One usually talks about pragmaticalization when in natural speech certain grammatical forms or individual lexemes change their status by going to the communicatively-pragmatic level of language and becoming purely pragmatic units that can take the form of independent utterances [2]; see also [3]. A common word is converted into a pragmalexem (for the term *pragmalexeme* see, for example: [4]), or a pragmateme, or a pragmatic marker. Often, these units are not one word, but a construction that, when used, can have several variants: *eto samoye, kak skazat', (nu) (ty) znayesh, vot (etot) vot, tuda-syuda, kak yego (yeye, ikh), kak eto, (ya) ne znayu* (*well, erm, you know, how shall I put it, what d' ya call it, like*) and others.

In linguistics, more than once, attempts were made to find a suitable term and to describe such units of oral discourse. For example, the concept of Discourse Markers [5] can be viewed as something similar to a "pragmateme". This term – *Discourse Marker* – now seems to be the most common. Besides, in Russian linguistic tradition there is also a term "discourse words". A. N. Baranov, V. A. Plungian and E. V. Rakhilina in their "Guide to the discourse words of Russian", one of the first Russian works specifically devoted to these elements of spoken speech, give the following interpretation to this concept: "In fact, there are units which, on the one hand, ensure the coherence of the text and, on the other hand, directly reflect the process of interaction between the speaker and the listener " [15].

The study [6],[7] ed. by K. L. Kiseleva and D. Payar can be considered one of the most fundamental research on discourse words. The authors argue that concept "discourse words" should be defined primarily on the basis of the functional criteria, the main aim of which is establishing the relationship between two (or more) discourse components. Thus, the meaning of discourse words that have no denotation can be studied only through their usage. The class of discourse words is open, and according to functional criteria it includes

relational or syntactic words, that is, modal words, as well as particles and some adverbs. In the article collection edited by K. L. Kiseleva and D. Payar such units are also called *logical particles*, *modal particles*, or *connectors*.

It should be mentioned that there were other terms suggested for some of the unit types we call pragmatic markers — such as *speech automatism* [8], a *discursive unit*, and *verbal hesitative*. The terms were to cover, in particular, polyfunctional units (often – expanded constructions), see, for instance, [9], which are verbal fillers for hesitation pauses. In the study of S. Brizer [10], such units are referred to as discourse structuring elements.Choosing *pragmatic marker* to be the main term, we proceed from an assumption that during pragmaticalization process there occur changes in the semantics of lexical units, the role of the pragmatic component increases, while significance of the denotative and significative elements decreases. This process may be accompanied by changes in usage (for example, unrealized valence, non-standard word order, etc.). As a result, the unit largely loses its lexical and often grammatical meaning, and its main function becomes the function that it realizes in the structure of oral text (discourse) and which can be called the pragmatic meaning of this unit [11].

At the next stage, the pragmatic markers are lexicalized in everyday communication which is associated with the general automatism of spontaneous speech and the fact that a pragmatic function is associated with the construction as a whole in certain communicative situations. This poses the task of creating a typology and further lexicographical description of PMs in our everyday speech, as well as of developing a methodology for annotating these units on corpus data.

Discourse markers vs. Pragmatic markers. To begin with, it is necessary to distinguish between the concepts of "pragmatic marker" (PM) and"discourse marker" (DM). These notions are in many ways close and DMs are quite well-described both in Russian and worldwide studies (see, for example: [12]). Both PMs and DMs are frequently used and participate in discourse creating and structuring. The specific features of PMs are as follows:

1) DMs (*v samom dele, pryamo, pochti, kstati*, etc. / *in fact, really, almost, by the way*) function in any speech, both oral and written. PMs (*eto samoye, nu vot, kak yego* (*yeyo/ikh), skazhem tak, znachit*, etc. / *um, well, what d' ya call it, so to speak, I mean*) are units of mostly oral speech or its stylizations in the written text, as well as in the texts of computer-mediated communication which is considered to combine the features inherent to both oral and written texts [13].

2) DMs that differ from PMs are generated by the speaker consciously and on purpose. They are fully functional discourse units that take part in the formation of its content structure (cf. such units as *vidimo, deystvitel'no* / *apparently, really*). PMs are generated at the level of speech automatisms and are practically uncontrolled by the speaker.

3) DMs, which are different from PMs, have both lexical and grammatical meaning (often they are adverbs, introductory words / phrases). The lexical meaning of the PM is either completely absent or considerably weakened. The grammatical meaning is either absent or it remains as an "atavism" [14]. The semantics and grammar of the original forms, from which the PM actually derives, is replaced by a pragmatic function.

4) DMs, which are different from PMs, express the speaker's conscious attitude to the subject of speech, "control the process of communication", "directly reflect the process of interaction between the speaker and the listener, the speaker's stance" [15]. The PMs can verbalize the speaker's attitude to the process of speech generation and perform many other functions (see section 3 below).

III. The Inventory of Pragmatic Markers in Everyday Spoken Russian

In this article the study of PMs is based on the corpus approach that has recently become very popular. Two corpora of spoken Russian (ORD and SAT) were used for finding out an inventory of the PMs in Russian oral speech and creating their typology [16].

Firstly, this is the corpus of Russian Everyday Speech "One Day of Speech" (ORD), which is today one of the most representative resources for analysis of Russian oral discourse [17]–[21]. An important feature of this resource is the fact that on the principle of "Holter monitoring" volunteer informants recorded their entire speech communication during the day. Thereby, sound recordings of everyday speech in a natural situation were obtained [22].

Secondly, this is the so-called "Balanced Annotated Collection of Texts" (SAT) which includes monologue speech recordings received from different professional groups of native speakers. All texts in SAT were obtained in 4 experiments – reading, rendering, image description, storytelling [9],[23–25].

A hypothesis was formed that the inventory and functions of the PMs in dialogical speech differ significantly from those of the PMs in monologue speech. Further, this hypothesis was confirmed [26].

For manual PM annotation in ORD and SAT the following operational typology was used.

- **A** – marker-approximator showing the speaker's uncertainty about what he is talking about:

  *- no u Barchukova%-to **kak by** / no i mashinu *N / i sama zarabotala vosem'desyat tysyach* (ORD) — *but with Barchukov **as it were** ... but also the car *unclear / and she earned herself 80000*;

- **G** – boundary/limit-setting marker (starting, finalizing and navigational) marks the borders within the text:

  *- **znayesh** / vot tozhe slaboye mesto b**d' a? tak stalo byt' ya tam / vot za kordonom / oni zh ne privykli remontirovat' // oni k etomu ne privykli kak u nas* (ORD) (starting + metacommunication) — ***y'know** / here is a f**cking weak place as well, yeah? So I'm there / well abroad / they are not used to repairing // they are not used to it as with us*;
  - a yeshche zhe yest' teoriya / chto my zhivom / (e...e) na

*vnutrenney poverkhnosti sfery // **nu vot da** # a v tsentre visit solntse // eto polaya zemlya // nazyvayetsya teoriya poloy zemli // nu eto zh nado bylo takoye pridumat' // *P eto kak (...) ya () odin raz / kogo-to sprosila chto takoye inversiya // *P mne skazali / *P vot predstav' sebe yaytso // *P vot inversiya / *P eto togda kogda / *P v yaytse budet ves' mir / a ves' mir vokrug budet sostoyat' iz *N zheltka* (ORD) (navigation) / *there is also a theory / that we live / uh, eeh at the inner surface of a sphere // **well yeah** # and at the centre there is a sun // this is hollow earth // called the theory of hollow earth // how could one have invented it //*P is like (...) I () once / asked someone what inversion is// * P told me / *P imagine an egg // *P here is inversion / *P it is when / *P the egg will contain the whole world / and the whole world will consist of *unclear yolk*;
*- i v kinoteatr tozhe khodim // teatr tak boleye red… / nu porezhe n-no / vse zhe byvayet // **v-o-ot** // letom // ya-ya / tak obychno byvayet chto-o-o / vse vykhodnyye* (CAT) (navigation) / *we also go to the theatre// less oft…/ well less often b-but/ it happens // **so-o** // in summer // I – I / it usually happens tha-at / every weekend eh*;
*- i ona prosto / u neye tam na na urovne podsoznaniya srabatyvayet / net / ne khochu / potomu chto // ya ne znayu pochemu / **dumayu chto*** (ORD) (finalizing) / *and she is simply/ it works with her subconsciously/ no/ I don't want/ because// I don't know why/ **I think so***;

- **D** – deictic marker that primarily has a demonstrative function:

*- eeh v kachestve cheloveka kotoryy tam rabotayet / eeh i kak by zhivet naverno // **vot tak vot** // strana izumitel'naya potomu chto oni drugiye // oni po-drugomu myslyat* (CAT) (+ navigation) / *eeh as a person who works there/ eeh and sort of lives probably// **so like that** // the country is wonderful because they are different// they think differently*;
*- eto dovol'no smeshno vyglyadit so storony // nu / naverno vso // **vot tak vot*** (CAT) (+ finalizing) / *it looks pretty funny to onlookers // well / probably everything // **like this***;

- **Z** – all kinds of replacement markers (somebody else's speech, enumeration line or its parts):

*- u neyo ... # a ya **i to i drugoye** (eh) to yest' ... # vy s ney ochen' ostorozhno* (ORD) / *she has… # and I tried **this and that** (eh) that is ... # you should be very careful with her;*
*- snayperka prichom prilichno strelyayet // *P s etim / s optikoy / **so vsemi delami** // *P a(:) / p... protiv kogo oni voyevali* (ORD) / *a sniper-woman and shoots quite well //*P from that/ with optics/ **all that stuff** //*P whom did they fight with*;
*- i ona kak na nas naletela ! vot tam **ty-ty-ty-ty-ty-ty** / da my alkashi tam / nu chto-to tam takoye / ya ne pomnyu* (ORD) / *and she attacked us! / like blah-blah-blah / yeah we are alcoholics/ well something like this/ I don't remember*;

- **X** – xenomarker that introduces someone else's speech into the narration:

*- ya vchera ikh vstrechayu / na ulitse / nu v «Pyatorochku» / ya shla kak raz v magazin / a byl vecher / poldesyatogo // ya **takaya** / o-o / vy priyekhali // a u vas zavtra zanyatiya budut? / oni **takiye** / budut / ya govoryu / da-a / ne povezlo mne // a chto takoye ? // a u menya zavtra u vas tri ... dva seminara* (ORD) — *yesterday I met them / in the street / well to "Pyaterochka" shop / I was just going to the shop / it was in the evening / half past nine // I'm **like** / oh / you've come // are you going to have classes tomorrow?/ they are like / yeah / I say / well / I'm unlucky // why? // tomorrow I have three… two seminars*;
*- i kiska nachinayet nazvanivat' etoy (e...e) staroy deve / **tipa togo chto** za kh**nya tam ? moy tipa muzhik prishol k vam ? *P a ta tam (...) tipa otsylayet / no postoyanno kucha prikolov vsevozmozhnykh* (ORD) — *and this puss starts calling this (eeh) spinster / **like** what the f...ck ? my like man has come to you ? *P and there (...) like tells her to bug off / lots of crackers of all sorts*;

- **M** – metacommunicative marker, showing there is "communication about communication" going on – speaker to the listener or speaker to herself:

*- a seychas / a seychas oni vot / (e...e) strakhovuyu **da** [a] sperva ? *P (e...e) / nu u kogo kakaya strakhovaya / **ponimayesh** [b] ? u kogo bol'shaya / tomu vygodno // a u kogo (...) ona ne povyshalasya* *P vot* (ORD) (+ navigation[b]) — *and now / and now they are / eeh insurance yes an(d) at first? *P eeh / well it depends on the insurance you have/ **y'see** ? some have a large one / for those it is profictable // and those who have (...) it hasn't been raised up*P well*;
*- da tam kakiye-to / eti samyye / i (yeshcho vot) / chto-to po-moyemu / ona kakiye-to protokoly raznoglasiya pishet // **ya ne znayu*** (ORD) (+ finalizing) — *well there are some / well these / and also / something like / she writes some protocols of disagreement // I **dunno***;

- **F** – Reflexive marker that shows the speaker's attitude to what has been said:

*- byli (e) kak-to vot / (e) (...) vot eti / kak ikh ? lyamblii ? **ili kak eto ?*** (ORD) — *there were eeh like / eeh these / how are they ? lamblia ? Or **what are they called ?***
*- s drugimi // *P nu (...) nespetsialistami **tak skazhem** // *P v toy oblasti / v kotoroy ya rabotayu* (ORD) — *with others // *P well (...) non-experts **so to say** // *P in the area / where I work.*

- **R** – Rhythm-setting marker:

*- slushay / gde-to (...) / berut **eti** (...) / vzryvnyye / veshchestva* (ORD) — *listen / one can get somewhere (...) / get **those** (...) / explosion / explosives;*
*- devyat' tysyach **tam** / s kopeykami* (ORD) — *nine thousand rubles **like** with kopeks*;

- **S** – Self-correction marker:

*- yarkaya solnechnaya pogoda // govorit' mozhno ? tak byl yark… / **eto samoye** / byl / iyul'skiy den' / vot / nebo*

bylo chistym / bezoblachnym / solntse / svetilo (CAT) — *bright sunny weather // can I speak ? Well it was a bright / **oh umm** / was July day / yeah / the sky was clear / no clouds / the sun / was shining*;
*- moy khoroshiy ! pozvonit' mne / i uznat' u menya ! *P ty mne zvonish i sprashivayesh o chem ugodno / no ob etom *V sprosila by / ya by tebe ob"yasnila by / *V i ty by () uzhe davno by sdelala / i mne by / v poldesyatogo / nervy ne trepala by / s etoy yerundoy / duratskoy ! s gektarami ! *P chto oni iz vas / zhivotnovodov khotyat () etikh (...) fu ty () pakharey (...) chertovykh vyrastit' / **chto li** ?* (ORD) — *my dear! Call me / and ask me! *P you call and ask me about whatever / but not about this *B have you asked / I'd explained / *B and you would () have made it long ago / and you wouldn't have / half past nine get on my nerves / with this nonsense / stupid nonsense! With hectars! *P what do they want / make those (...) stock-breeders of you my goodness () bloody ploughmen they want to bring up / **or what** ?*;

- **H** – Hesitation marker:

*- pokhozhe na kartiny Shishkina mne pochemu-to srazu vspomnilos' "Utro v sosnovom lesu" samaya moya / **ne znayu** samaya primitivnaya khranyashchayasya u menya v golove kartina iz detstva / vot* (CAT; description + hesitation + metacommunication) — *it looks like Shishkin paintings somehow I remembered at once "Morning in a pine forest" my most / I dunno the most primitive picture I've kept in my head since childhood*;
*- nu u neyo vral / (...) **etot** (...) pribor navernoye* (ORD) — *well it was wrong/ (...) **that** (...) device I dunno*.

IV. Pragmatic Markers Annotation and its Difficulties

To obtain statistical information on the frequency of PMs use in oral Russian, continuous marking of PMs was performed on a pilot corpus data (60.000 tokens for the ORD corpus and 15.000 tokens for SAT). Annotation was made in the program ELAN. For details on the method, see [27]. The developed annotation technique takes into account structural variability and polyfunctionality of PMs.

While doing the pilot annotating, we encountered the following main problems:

1) It is not always clear how to determine the stages of the pragmaticalization process that takes place with PMs, when from a fully notional word or construction, through the process of grammaticalization and pragmaticalization they turn into a lexicalized expression used as speech automatism, cf.:

*- my vozili semechki v obmen na maslo takoye **znayesh** () **nu** aromatnoye — we were bringing sunflower seeds and in return we getting oil like **y'know well** the one that has strong flavor.*

2) It is quite difficult to determine the main and secondary functions for individual PMs. Thus, in the following example, it is impossible to unequivocally say whether the main function of the PM is to hesitative or approximative:

*- nu **tam** v osnovnom sovetskuyu chital / **znayesh** literaturu // nashu tam / a(:) ! vpered k kommunizmu ! — **well** he mostly read Soviet / **y'know** fiction // our like / for the communism !*

3) It is difficult to determine the PM borders, i.e. to decide whether an expression is one multi-word marker or a chain of several markers, cf.:

*- vchera my s na... s Nadey% vykhodim s raboty // *P ona menya prosit / u vas yest' **tam** telefon (eh) Glukharevoy% ? ya govoryu da // *P nu i **znachit tam** (...) nakhozhu / diktuyu yey* (ORD) — *Yesterday me with na.. Nadya go out from work // *P she asks me/ do you have **urm** a telephone number (eh) of Glukhareva ? I say yes// *P well **and so** (..) I find it/ dictate to her*.

4) Practical work on PM annotating is made more difficult by the fact that splitting spontaneous speech into minimal units (syntagmes, in our case) cannot always be fulfilled univocally, cf.:

*- ya seychas 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 (:) (eh) **vot** / ya vyyasnyu / poyedet ona segodnya ili zavtra k vam (ORD)/ I'm going to call Marina now/ and I will find out// the thing is/ Marina Zhdanova was going to visit you// no no no// Marina Glukhareva// *unclear **ok**/ *P a-and eeh **ok**/ I will find out / if she comes to you today or tomorrow.*

5) Corpus analysis has shown that PM class contains a variety of units, in particular, "lexicalized constructions with a pronominal component" [28]); verbal PMs. All of them are discourse units that have undergone pragmaticalization: their lexical meaning in actual cases of usage has been largely weakened or completely lost and has been replaced by a pragmatic meaning or function in speech.

6) Finally, PM annotating is complicated by the fact that they seem to be no different from meaningful speech units and only in the context realize their new status, which appears, as a rule, as a result of the process of pragmatization, cf .:

*- **tam** mne kazhetsya blizhe —it's closer **there**, it seems to me* (adverb of place);

*- vsyo ravno vsya eta utilizacija koroche ona **tam** maksimum davala garantiju **tam** na 50 let — this waste disposal, **to put it short,** it gave 50 years guarantee maximum* (PM);

*- **ja ne znaju** / otpravila ona ego ili net — **I don't know** if she has sent it or not* (the main clause in a complex sentence);

*- ili... ili kakoj-to nemeckij ? nu **ja ne znaju** / Brandenburgskie_vorota$ / chto-to takoe — Or... something Germain? well, **I don't know** / Brandenburg gate / something like that* (PM).

Often the status of a PM (as a rule, that of a hesitation marker) is acquired by the unit exclusively in a hesitation context, cf. (additional hesitation in the contexts are underlined):

*- no vot kak-to eshcho kak-to / ja pervyj raz **kak govoritsja** / ja tak pisala kakie-to svoi tam // *P chisto takie / vizual'nye*

posmotrela tam... — but well I also like / for the first time I, ***as they say*** */ I was writing some of my // just some / I looked some at visuals;*

- vot no (...) v etom ***sobstvenno*** *(...) kak skazat' / v etom ... *P zagadka Rossii — well but.. (...) that is* ***exactly*** *so to say... mystery of Russia;*

It is important to note that in the absence of such a hesitation environment in the use of such units, one can say about their unmotivated use by the speaker — for the sole purpose of decorating his speech (for just a manner of speaking), which allows them to be assigned to the PM class of ornaments. For example:

- nu // v principe / ja sejchas smotryu / potomu chto / ***kak govoritsya*** */ on vsjo-taki sostavlyal dva mesjaca nazad — well, actually I'm now looking at / because /* ***as they say*** */ he made it two months ago;*

- nu zdorovo / nu vsjo budet zaviset' ot moego ***tak skazat'*** *novogo grafika — well that's great / well everything will depend on my* ***so to say*** *new schedule.*

All this once again emphasizes the need for analysis in identifying PMs in a wide context, indispensable manual refinement of the results of PMs automatic annotation corpus material, as well as taking these features into account when lexicographically "portraying" such units.

V. Some Statistics of Pragmatic Markers Usage in Everyday Russian

Continuous annotation of speech material allowed us to obtain preliminary statistics on the frequency of use for various PM types (see Tables I–II).

TABLE I. The Distribution of PM Functional Types in Everyday Dialogues

PM Functional Type	%	ipm
H	29.81	4179
M	18.77	2631
K	9.72	1362
G	3.11	436
A	2.83	397
D	1.89	264
Z	1.04	145
F	0.85	119
R	0.57	79
C	0.10	13
Multifunctional PMs	28.96	4059
Uncertain	2.30	304

TABLE II. The Distribution of PM Functional Types in Monologues

PM Functional Type	%	ipm
H	23.70	4251
G	6.30	1129
D	1.85	332
Z	1.85	332
R	1.11	199
A	0.74	133
M	0.74	133
F	0.74	133
K	0.37	66
Multifunctional PMs	61.11	10958
Uncertain	1.48	266

Thus, Table I shows the shares (per cent) and frequencies (items per million) for the main functional PM types in everyday dialogical speech (ORD data), while table 2 shows the same indexes for spoken monologues from SAT corpus. Separate lines provide data on multifunctional PMs, as well as on those pragmatic units for which their functional type was difficult to determine.

One can see that for both types of speech, hesitation marker (H) is the most frequent; in the dialogical speech, the proportion of metacommunicative and xeno-specifying PMs is high, and the limit-setting/boundary markers are relatively frequent. High percentage of polyfunctional PMs in the SAT corpus can be explained, apparently, by a specific recording format of experimental material – the subjects mainly described pictures and rendered the text, which resulted in a high frequency of hesitative PM components. Other statistical characteristics regarding the use of PMs in oral speech can be found in [26], [29].

VI. The New Methods of Corpus Data Investigation: Proof of Concept

It is also necessary to consider a number of specific characteristics of each speaker (gender, age, social status, psych type) and provide a separate analysis of speech of their interlocutors to get representative sample and maximum accuracy of the data obtained. This kind of research recently required a lot of efforts to collect data and a huge amount of time was necessary for the manual data processing. Computer technologies and corpus linguistics offer fundamentally new possibilities in this aspect today.

The new methods for processing corpus data will answer a number of questions related to the establishment of a correlation between the frequency of use of a particular pragmatic marker and various speaker's characteristics informant (gender, age, psychological type), as well as the functional distribution of pragmatic markers without involving additional tools. To obtain the most reliable data about the number of pragmatic markers to the number of words spoken by an informant, it was necessary to obtain summary data on the general distribution of speakers by the number of words. Correspondingly processed preliminary distribution of speakers by the number of words is presented in Fig. 1.

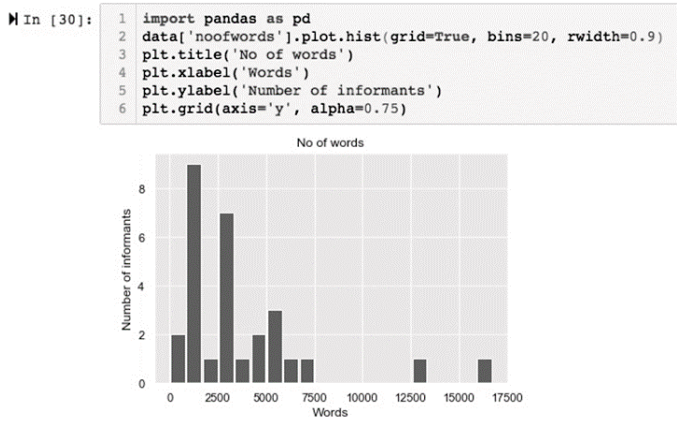

Fig. 1. Distribution of speakers by the number of words

Figure 1 shows that 4 people do not fit the total sample and have a record low (leftmost column) or a record high (rightmost column) number of words. This allows us to measure the total data variation and not take into account the influence of extreme elements. However, in this case, it seems reasonable to weed out only minimal values. It seems to be possible to talk a lot during the day but saying only 5 words during the day seems impossible, so unrealistic entries can be eliminated.

The conducted proof of concept of using automated tools for processing of the everyday Russian speech corpus which was created based on the ORD corpus allowed to improve and visualize some main aspects of the research. The practice of early piloting was considered to be successful and allowed to clarify the requirements and improve usability to develop an effective tool for processing of a large amount of data from the speech corpus. This software toolkit is planned to be carried out at the next stages of this research for the specific search of correlation between pragmatic markers and accentuation of personality.

VII. Conclusion. Towards Compiling the Dictionary of Russian Pragmatic Markers

The study showed that the PMs really are quite frequent units in oral speech and it's important to study it using the most modern approaches [29]. Thereby, it is possible to preserve the progress achieved and additionally solve problems with visualization of results, as well as implement one-time processing and comparative statistical analysis of several pragmatic markers at once using Jupyter Notebook.

Modern voice systems (including voice assistants like Siri, Alisa etc.) have already reached the level at which they invariably have straight interaction with a human. Since this is a two-way interaction, a human almost always uses pragmatic markers in speech. Such functional units are hard to recognize and correctly interpret even for machine learning algorithms. Usually indeed, pragmatic markers can be safely ignored, and a person can adapt to the limitation of the voice system being unable to understand accentuation (e. g., irony or the level of anger). So, a person can simplify the speech and make it easier for the system. This was an acceptable level for the voice systems, but for a better human experience, a voice system should be able to recognize and interpret the pragmatic markers, it can help to clarify the request, find a missing word (which a speaker tries to find out), and make the communication more natural.

Currently, as a consequence of ignoring these natural units of speech in NLP, the speech interpretation loses such important things as — for example — the relation of the speaker to the subject, so the natural interpretation of speech (which contains pragmatic markers) is degraded beyond repair. This loss of pragmatic units is like a two-dimensional movie that as lost its three-dimensional spatial information. Hence this study aims to be a first step in the attempt to solve the deep problems in this field to grasp the deeper meaning expressed by humans.

Therefore, it seems that we still need another sort of a “Guide to Discourse Words,” something like a Dictionary of Pragmatic Markers in Russian Conversation that would include discourse units grouped by their functions, and their detailed description. This dictionary can be used for markup of speech data for machine learning, since for a correct interpretation, the algorithm should identify the function of the pragmatic units as an opposite to the meaning of the regular words. This type of markup would be impossible without a detailed description with lots of examples.

The entry structure in such a dictionary should include several lexicographic zones:

- semantic zone — the definition of the original unit or PM components in dictionaries, a kind of “semantic background” for describing the functioning of this unit in speech;

- functional zone — all possible functions of the PM in everyday oral speech;

- rich illustration material that would accompany the description of the PM functions;

- quantitative ratios for the determined functions;

- correlations with the type of speech (monologue / dialogue, everyday / public speech, academic discourse, etc.) and the speaker's characteristics — through the system of notations (PMs in male / female speech, speech of different age and professional groups, etc.).

Potential users of this dictionary will be linguists, speech technology specialists, researchers of everyday Russian speech, those who describe the grammar of Russian, interpreters of spontaneous oral texts and those who translate stylizations of colloquial speech into other languages (at least, as part of a fiction novel when trying to convey specific features of the characters), teachers of Russian as a foreign language, and all others who are interested in problems of everyday speech.

Another important conclusion of the study is that we could see the need to determine clearer, better formalized features according to which one can identify PMs of different groups (the parallel annotation shows that in some cases the annotators rely on different, intuitive features of PMs, on the assessment of the annotation consistency see [27]).

Optimization of the annotation methodology and PM attribution in corpus material will provide more reliable data about the use of PMs in everyday speech and will help create effective tools for the study of oral communication in the language that one studies.

For example, the successful results of the functional distribution of pragmatic markers can be used in training the interpretation system and will help to find semantic dependencies between words. Apparently, it is advisable to model the behavior of such a “linguistic agent”, which accumulates knowledge about what function a pragmatic marker has in the current context and how it can be interpreted automatically.

ACKNOWLEDGMENT

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

REFERENCES

[1] C. Ghezzi, P. Molinelli (eds.), *Discourse and Pragmatic Markers from Latin to the Romance Languages*. Oxford: Oxford University Press, 2014.

[2] E. Graf, *Interjektionen im Russischen als Interaktive Einheiten*. Frankfurt am Main: Peter Lang-Verlag, 2011.

[3] S. Günther, K. Mutz, "Grammaticalization vs. pragmaticalization? The development of pragmatic markers in German and Italian", in *What Makes Grammaticalization? A Look from its Fringes and its Components*, W. Bisang, N. P. Himmelmann, B. Wiemer (eds.), Berlin: Language Arts & Disciplines, 2004, pp. 77–107.

[4] R. Rathmayr, *Die Russischen Partikeln als Pragmalexeme*. München: Sagner, 1985.

[5] D. Schiffrin, *Discourse Markers*. Cambridge: Cambridge University Press, 1988.

[6] K. Kiseleva, D. Payar (eds.), "Diskursivnye slova kak object lingvisticheskogo opisanija" ["Discourse words as the object of linguistic description"], in *Diskursivnye slova russkogo yazyka: opyt kontekstno-semanticheskogo opisanija* [*Discourse words of Russian: experience of contextual-semantic description*], Moscow: Metatext, 1998, pp. 7–11.

[7] K. Kiseleva, D. Payar, *Diskursivnye slova russkogo yazyka: kontekstnoe var'irovanie i semanticheskoje yedinstvo.* [*Discourse words of the Russian language: context variation and semantic integrity*]. Moscow: Azbukovnik, 2003.

[8] E. Ju. Verkholetova, *Strukturno-dinamicheskij podkhod k social'noj stratifikacii ustnoj rechi.* Avtoref. diss. … kand. filol. nauk [*Dynamic Structure Approach to the Social Stratification of Speech*. Thesis synopsis of PhD philol. sci. diss.]. Perm: Perm State University, 2010.

[9] *Zvukovoj korpus kak material dlja analiza russkoj rechi. Kollektivnaja monografia. Chast' 2. Teoreticheskie i prakticheskie aspekty analiza. Tom 1. O nekotorykh osobennostyakh ustnoj spontannoj rechi raznogo tipa. Zvukovoj korpus kak material dlya prepodavanija russkogo yazyka v inostrannoj auditorii [Speech Corpus as a Base for Analysis. Part 2. Theoretical and Practical Aspects of Analysis. Vol. 1. On Some Features of Different Types of Oral Spontaneous Speech. Speech Corpus as a Base for Teaching Russian in a Foreign Audience*]. N. V. Bogdanova-Beglarian (Ed.) St.Petersburg: St. Petersburg State University, 2014.

[10] S. Brizer, "From subject to subjectivity: Russian discourse structuring elements based on the adverbial participle *govorya* 'speaking'", *Russian Linguistics*, no. 36, pp. 221–249, Nov. 2012.

[11] N. V. Bogdanova-Beglarian, "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*], iss. 3 (27), pp. 7–20, 2014.

[12] N. V. Bogdanova-Beglarian, Yu. A. Filyasova, "Discourse vs. pragmatic markers: a contrastive terminological study", in *5th Int. Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2018*, *Vienna ART Conference Proceedings*, vol. 5, iss. 3.1, 2018, pp. 123–130.

[13] N. V. Bogdanova-Beglarian, "Grammaticheskije "atavizmy" pragmaticheskikh markerov russkoj ustnoj rechi" ["Grammatical "atavisms" of pragmatic markers of Russian oral speech"], in *Russkaya grammatika: strukturnaya organizaciya yazyka i processy yazykovogo funkcionirovaniya* [*Russian Grammar: Structural Organization of Language and Processes of Language Functioning*], Moscow: URSS, 2019, pp. 436–446.

[14] M. A. Krongauz, "'Lytdybr' ot blogera ili kak internet-yazyk delaet pis'mennuyu rech formoj sushchestvovaniya razgovornogo yazyka" ["Lytdybr" from a blogger or as an Internet language makes writing a form of the existence of a spoken language"], *Russkij mir.ru* [*Russian World.ru*], no. 6, 2009, pp. 40–43.

[15] A. N. Baranov, V. A. Plungian, and E. V. Rakhilina, *Putevoditel' po diskursivnym slovam russkogo jazyka* [*The Guidebook on Discourse Words of Russian*]. Moscow: Pomovskij i partn'ory Publ., 1993.

[16] N. Bogdanova-Beglarian, E. Baeva, O. Blinova, G. Martynenko, T. Sherstinova, "Towards a description of pragmatic markers in Russian everyday speech", in *Speech and Computer. SPECOM 2018. LNCS*, vol. 11096. Switzerland: Springer, 2018, pp. 42–48.

[17] *Russkij jazyk povsednevnogo obshchenia: osobennosti funkcionirovania v raznykh social'nykh gruppakh. Kollektivnaja monografia* [*Everyday Russian Language: Functioning Features in Different Social Groups. Collective Monograph*], N. V. Bogdanova-Beglarian, Ed. St. Petersburg: LAIKA, 2016.

[18] N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, G. Martynenko, "An exploratory study on sociolinguistic variation of spoken Russian", in *SPECOM 2016, Lecture Notes in Artificial Intelligence*, *LNAI*, vol. 9811. Switzerland: Springer, 2016, pp. 100–107.

[19] N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, E. Baeva, G. Martynenko, A. Ryko, "Sociolinguistic extension of the ORD Corpus of Russian Everyday Speech", in *SPECOM 2016, Lecture Notes in Artificial Intelligence*, *LNAI*, vol. 9811. Switzerland: Springer, 2016, pp. 659–666.

[20] N. Bogdanova-Beglarian, T. Sherstinova, O. Blinova, G. Martynenko, "Linguistic features and sociolinguistic variability in everyday spoken Russian", in *SPECOM 2017, Lecture Notes in Artificial Intelligence*, *LNAI*, vol. 10458. Switzerland: Springer, 2017, pp. 503–511.

[21] N. Bogdanova-Beglarian, O. Blinova, T. Sherstinova, G. Martynenko, "Corpus "One Speaker's Day" in studies of sociolinguistic variability of Russian colloquial speech" ["Korpus "Odin rechevoj den'" v issledovaniakh sociolingvisticheskoj variativnosti russkoj razgovornoj rechi"], in *Analysis of Spoken Russian (AR3-2017). Proceedings of the seventh interdisciplinary seminar* [*Analiz russkoi razgovornoi rechi. Trudy sed'mogo mezhdisciplin. seminara*], St. Petersburg, 2017, pp. 14–20.

[22] 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 *TSD 2009, LNAI*, vol. 5729. Berlin-Heidelberg: Springer, 2009, pp. 250–257.

[23] *Zvukovoj korpus kak material dlya analiza russkoj rechi. Kollektivnaja monografia. Chast' 1. Chtenie. Pereskaz. Opisanie [Speech Corpus as a Base for Analysis of Russian Speech. Collective Monograph. Part 1. Reading. Retelling. Description*], N. V. Bogdanova-Beglarian, Ed. St. Petersburg: St. Petersburg State University Publ., 2013.

[24] *Zvukovoj korpus kak material dlya analiza russkoj rechi. Kollektivnaja monografia. Chast' 2. Teoretichskie i prakticheskie aspekty analiza. Tom 2. Zvukovoj korpus kak material dlya novykh leksikograficheskikh proektov* [*Speech Corpus as a Base for Analysis of Russian Speech. Collective Monograph. Part 2. Theory and Practice of Speech Analysis. Vol. 2. Speech Corpus as a Base for New Lexicographical Projects*]. N. V. Bogdanova-Beglarian, Ed. St. Petersburg: St. Petersburg State University, 2015.

[25] N. V. Bogdanova-Beglarian, T. Yu. Sherstinova, K. D. Zaides, "Korpus "Sbalansirovannaja Annotirovannaja Tekstoteka": metodika mnogouronevogo analiza russkoj monologicheskoj rechi" ["Corpus "Balanced Annotated Text Library": Methodology multi-level analysis of the Russian monological speech], in *Analysis of Spoken Russian (AR3-2017). Proc. of the 7th Interdiscipl. seminar* [*Analiz russkoi razgovornoi rechi. Trudy sed'mogo mezhdisciplin. seminara*], St. Petersburg, 2017, pp. 8–13.

[26] N. Bogdanova-Beglarian, O. Blinova, T. Sherstinova, G. Martynenko, K. Zaides, T. Popova, "Annotirovanie pragmaticheskikh markerov v russkom rechevom korpuse: problemy, poiski, resheniya, rezul'taty" ["Annotation of pragmatic markers in the Russian speech corpus: problems, searches, solutions, results"] in *Kompjuternaia lingvistika i intellektual'nye tekhnologii: Po materialam yezhegodnoi mezhdunarodnoj konferencii «Dialog–2019»* [*Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference «Dialog–2019»*], iss. 18(25), Moscow 2019, pp. 72–85.

[27] N. Bogdanova-Beglarian, O. Blinova, G. Martynenko, T. Sherstinova. K. Zaides, "Pragmatic markers in Russian spoken speech: an experience of systematization and annotation for the improvement of NLP tasks", in *Proceedings of the FRUCT'23*, *FRUCT Oy*, Finland, 2018b, pp. 69–77.

[28] V. I. Podlesskaja, "Nechotkaja nominacyja v russkoj razgovornoj rechi: opyt korpusnogo issledovanija" ["Vague reference in Russian: Evidence from spoken corpora"], in *Kompjuternaia lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnoj konferencii «Dialog–2013». Tom 1. Osnovnaja programma konferencii* [*Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference «Dialog–2013». Vol. 1. The Main Program of the Conference*], iss. 12(19), V. P. Selegej, Ed. Moscow: RSUH Publ., 2013, pp. 631–643.

[29] N. Bogdanova-Beglarian, O. Blinova, T. Sherstinova, G. Martynenko, "Pragmatic markers distribution in Russian everyday speech: Frequency lists and other statistics for discourse modeling", In: *Speech and Computer. SPECOM 2019. LNCS*, vol. 11658. Springer, Cham, pp. 433–443.