Pragmatic Markers Distribution in Russian Everyday Speech: Frequency Lists and Other Statistics for Discourse Modeling

Natalia Bogdanova-Beglarian[1], Tatiana Sherstinova[1,2(✉)], Olga Blinova[1], and Gregory Martynenko[1]

[1] Saint Petersburg State University, 7/9 Universitetskaya nab., St. Petersburg 199034, Russia
{n.bogdanova, t.sherstinova, o.blinova, g.martynenko}@spbu.ru
[2] National Research University Higher School of Economics, 190068 St. Petersburg, Russia
tsherstinova@hse.ru

Abstract. Pragmatic markers (PMs) are discourse units (words and multiword expressions) with a weakened referential meaning, which perform a variety of pragmatic tasks. For example, in English the common PMs are "well", "you know", "I think", and many others. PMs are integral elements of spoken discourse in every language. According to the results obtained from the ORD corpus of everyday Russian, their share can reach up to 6% of the total number of words in speech of individual speakers. More than that, in some speech fragments, PMs may even exceed the share of significant units (i.e., standard words). However, despite their frequency and usualness, PMs are still poorly understood. Current NLP and discourse modeling systems lack information on PMs distribution and usage, this fact leads to noticeable shortcomings in work of these systems when they face spontaneous speech of everyday spoken discourse. In this paper we present top frequency lists of PMs for Russian dialogue and monologue spoken speech in general, and also for separate sociological groups of informants (by gender and by age). Our current list of PMs for Russian contains 450 units which are the variants of 50 main structural types. Besides, we consider the most frequent functions of PMs in spoken Russian. The presented quantitative data may be used for improvement of NPL and discourse modeling systems.

Keywords: Spoken Russian · Everyday discourse · Pragmatics · Pragmatic markers · Spoken dialogue · Spoken monologue · Speech corpus · Statistics · Frequency lists · Sociolinguistics · NLP

1 Introduction

Pragmatic markers (PMs) are discourse units (words and multiword expressions) with a weakened referential meaning, which perform a variety of pragmatic tasks [1]. They are used to express speaker's attitude to speech content, help to introduce new topics and to structure the discourse as the whole, provide the speaker time to find the proper word,

etc. [2]. In earlier papers on spoken discourse, PM were considered within a wider class of *discourse particles* or *discourse markers* [3–7, etc.]. In this paper we adhere to understanding PM as it was proposed in [8], and the distinction between pragmatic markers (PMs) and discourse markers (DMs) is presented in [9].

It can be assumed that PMs may be observed in every natural spoken language. And more than that, they are integral elements of every spoken discourse. For example, in English the common PMs are "well", "you know", "I don't know", and many others, and in Russian the similar PMs are "tak" ("well"), "znaesh" ("you know"), "ne znaju" ("I don't know"), etc.

According to the results obtained from the ORD corpus of everyday Russian [10, 11], their share can reach up to 6% of the total number of words in speech of individual speakers. More than that, in some speech fragments, PMs may even exceed the share of significant units (i.e., standard words). However, despite their frequency and usualness, PMs are still poorly understood. Current NLP and discourse modeling systems lack information on PMs distribution and usage, this fact leads to noticeable shortcomings in work of these systems when they face spontaneous speech of everyday spoken discourse [9].

In this paper we present frequency lists of PMs for Russian dialogue and monologue spoken speech in general, and also for separate sociological groups of informants. Besides, we consider the most frequent functions of PMs in spoken Russian.

2 Research Data

2.1 Dialogue Everyday Speech

For the analysis of dialogue everyday speech, 149 macroepisodes of everyday communication [12] were selected from the ORD corpus, obtained from 98 volunteer participants. The total sample size is 308905 words. To form a balanced subcorpus, the representatives of different gender, age and professional groups were included in the sample, and the information from participants' psychological tests concerning their psycho type and temperament was used as well. Moreover, when forming the sample, different conditions of everyday communication (locus, participants, type of communication) were taken into account.

As a result, the research subcorpus contains speech samples from participants of different professional groups—service workers, IT-specialists, representatives of engineering specialties, "office" workers, representatives of power structures, university professors, etc. The presented communicative settings relate to both formal and informal communication (with colleagues at work, with classmates, friends or parents, and communication in the family). Among the sample participants are 45 women (46%) and 53 men (54%). Speakers' age varies in a wide range (from 17 to 83 years).

2.2 Monologue Speech

For the study of monologue speech, texts of various types, obtained from 34 informants, were selected from the SAT corpus, known as "Balanced Annotated Text

Library" [13]. The sample consists of 50128 words. When selecting spoken texts for a balanced subcorpus, we tried to balance texts reflecting different scenarios of monologue speech generation, as well as speech samples from participants belonging to different gender, age and professional groups.

The outcome subcorpus contains monologues of 5 communicative scenarios types: retelling of both narrative and non-narrative texts, description of both narrative and non-narrative images, and a free story on some definite topic. The sample contains 34 texts of each type, so the sample is balanced, first of all, according to the types of monologue texts included in it.

The subcorpus presents the monologue speech of informants belonging to two professional groups—lawyers and doctors (JUR and MED subcorpora), in total there are 170 monologue texts in the sample. Women predominate among participants, since the "medical speech" of MED subcorpus was recorded only from women. The sample of lawyers-participants is gender-balanced: it includes monologues from 8 women and 9 men. In the SAT sample, the age of speakers ranges from 23 to 49 years. The majority of speakers (22 informants) have a high level of speech competence, 12 informants belong to groups with an average and low level of speech competence.

2.3   Data Annotation

The both subsamples were annotated in ELAN [14] using the scheme described in [9]. Thus, the annotators had to fill in four additional levels: (1) PM, (2) PM function, (3) speaker's code, and (4) comments.

For PM annotation the list of Russian PMs containing 450 units being the variants of 50 main structural types was used. This current list of PMs main structural types is given in Table 1.

For each PM, its main function was assigned. The list of these functions is the following [9]:

(1)   A—marker-approximator ("tipa", "kak by", etc.);
(2)   G—boundary marker, including *starting*, *final*, and *navigational* markers ("vot", "koroche", etc.);
(3)   D—deictic marker ("vot etot vot", "vot takoj vot", etc.);
(4)   Z—replacement marker referring to some whole set or its part ("i tak dalee", "i vs'o takoe", "to-sio"), as well as for imitating someone else's speech ("bla-bla-bla");
(5)   K—"xeno" marker that introduces someone's speech ("tipa", "govorit", etc.);
(6)   M—meta-communicative marker that refers to "communication about communication" ("znaesh", "vidish");
(7)   F—"reflexive" marker that expresses reflection on what is said ("tak skazat'");
(8)   R—rhythm-forming marker ("vot", "tam", etc.);
(9)   C—marker of self-correction ("v smysle", "vernej", etc.);
(10)   H—hesitation markers ("eto", "vot", "tam", etc.) [9].

Apart from the "pure" types, the "mixed" (or polyfunctional) functions could be registered (e.g., AG, AGH, GRH, AF, etc.), reflecting the overall polyfunctionality of PMs, which is very typical in oral speech [6].

Table 1. The main structural types of PM in spoken Russian with the number of correspondent variants

Structural type	The number of variants	Structural type	The number of variants
vot takoj vot	109	tak	5
pyatoe desyatoe	30	tam	5
eto	30	znachit	4
vsyo takoe	25	i to i drugoe	4
eto samoe	18	predstav'	4
znaesh'	17	ne znayu	4
takoj	13	da	4
to syo	11	minutu	3
vrode	10	i tak dalee	3
slushaj	9	ili kak ego/eto	6
dumayu	9	sekundu	3
tipa	8	na samom dele	2
smotri	8	i vse dela	2
tak i tak	8	prikin'	2
ponimaesh'	8	zaceni	2
predstavlyaesh'	8	glyan'	2
boyus' chto	8	voobshche	2
chto eshchyo	8	zamet'	2
skazhem	7	vot	2
vidish'	6	kak by	2
vsyo	6	ili chto	2
ili eshchyo chto-to takoe	6	kak ego	2
govorit	6	koroche	1
kak eto	6	sobstvenno	1
to to	5	kak ih	1

Processing the results of PMs annotation based on corpus material allowed us to obtain data on the frequency of occurrence of individual pragmatic markers and their functions, as well as on the dependence of PMs use on speaker's characteristics. Below are some of the data obtained.

3   Frequency Lists of Pragmatic Markers in Monologue and Dialogue Speech

Two top frequency lists of PMs are presented in Tables 2 and 3, which include: the ranks, the frequency of PMs in absolute numbers, the share of specific PMs from all the PMs in the sample (in %), and the share of specific PMs from all words in the sample (ipm). The data presented here were calculated on the subsample of 60000 words for dialogue speech, and on that of 15000 words for monologue speech.

It is seen from the Tables 2 and 3, that the most commonly used PM in the both types of Russian speech turned out to be "vot", which is usually used as a "boundary marker" (G), a hesitation (H) and rhythm-forming (R) marker. Besides, among the frequent PMs of both types are "tak", "tam", "kak by", and "nu vot".

Table 2. The most frequent PMs in Russian dialogue speech

Rank	PM	Abs. number	The share (%) of PM among the other PMs	IPM (to the total number of tokens)
1	vot	149	14.06	2483
2	tam	117	11.04	1950
3	da	82	7.74	1367
4	govorit	70	6.60	1167
5	kak by	60	5.66	1000
6	eto	44	4.15	733
7	eto samoe	43	4.06	717
8	znaesh'	41	3.87	683
9	koroche	38	3.58	633
10	tak	36	3.40	600
11	ne znayu	25	2.36	417
12	slushaj	23	2.17	383
13	znachit	21	1.98	350
14	nu vot	21	1.98	350
15	tipa	21	1.98	350
16	ponimaesh'	19	1.79	317
17	takoj	17	1.60	283
18	vidish'	11	1.04	183
19	takie	11	1.04	183
20	na samom dele	10	0.94	167
21	etot	9	0.85	150
22	vot takoj vot	7	0.66	117
23	govoryu	7	0.66	117
24	vot eti vot	6	0.57	100
25	dumaju	5	0.49	83

Table 3. The most frequent PMs in Russian monologue speech

Rank	PM	Abs. number	The share of PM am. the other PMs (%)	IPM (to the total number of tokens)
1	vot	139	51.48	9232
2	znachit	15	5.56	996
3	tak	15	5.56	996
4	tam	13	4.81	863
5	kak by	12	4.44	797

Table 3. (*continued*)

Rank	PM	Abs. number	The share of PM am. the other PMs (%)	IPM (to the total number of tokens)
6	*nu vot*	12	4.44	797
7	*vsyo*	4	1.48	266
8	*i tak dalee*	4	1.48	266
9	*vot tak vot*	3	1.11	199
10	*nu tak*	3	1.11	199
11	*takaya*	3	1.11	199
12	*takie*	3	1.11	199
13	*vot eta vot*	2	0.74	133
14	*da*	2	0.74	133
15	*kak eto nazyvaetsya*	2	0.74	133
16	*ya dumayu chto*	2	0.74	133
17	*v eti*	1	0.37	66
18	*vot sejchas by vot*	1	0.37	66
19	*vot takaya vot*	1	0.37	66
20	*vot takoe vot*	1	0.37	66
21	*vot eti vot*	1	0.37	66
22	*vot eto vot*	1	0.37	66
23	*vot etot vot*	1	0.37	66
24	*vrode*	1	0.37	66
25	*vrode by*	1	0.37	66

In the top zone of the frequency list for dialogue speech we can also see meta-communications (M) —"da", "znaesh", "ponimaesh", "vidish", hesitation markers (H) "eto", "eto samoje", "koroche tak", the xeno-indicator marker (K) "govorit" (which is often reduced), and many others. In the monologue speech subsample, the upper zone of PM frequency list contains mainly different types of boundary markers (G), which mark the beginning or end of the monologue or are serving as discourse navigators—"vot"/"nu vot", "znachit tak", "vsyo". We should mention the high frequency of the deictic marker (D) "vot tak vot" as well.

Table 4 presents the lists of most typical PMs for Russian dialogue and monologue speech types measured in difference of correspondent IPM values. It could be seen from this table, that the maximum difference is observed for PM "vot", which occurs much more often in monologues then in dialogue.

Table 4. The most obvious differences in PMs usage between dialogue and monologue speech

"Dialogue" PMs			"Monologue" PMs		
Rank	PM	IPM difference	Rank	PM	IPM difference
1	*da*	−951	1	*vot*	7262
2	*govorit*	−926	2	*znachit*	718
3	*tam*	−684	3	*tak*	520

(*continued*)

Table 4. (*continued*)

"Dialogue" PMs			"Monologue" PMs		
Rank	PM	IPM difference	Rank	PM	IPM difference
4	*eto samoe*	−569	4	*nu vot*	519
5	*znaesh'*	−542	5	*i tak dalee*	266
6	*eto*	−516	6	*nu tak*	186
7	*koroche*	−503	7	*takaya*	159
8	*slushaj*	−304	8	*vot tak vot*	146
9	*tipa*	−278	9	*kak eto nazyvaetsya*	120
10	*ne znayu*	−265	10	*ya dumayu chto*	120

The research presented in [15] has shown that the differences in the use of PMs between dialogue and monologue types of speech according to the Mann-Whitney test can be considered statistically significant.

4   The Functions of Pragmatic Markers in Monologue and Dialogue Speech

Table 5 presents the top frequency lists of PMs functions in dialogue and monologue Russian speech (the meanings of the codes were listed above in Sect. 2.3), and in Table 6 one may see the comparison of PMs functions frequency in these two types of spoken Russian.

Table 5. The top frequency lists of PMs functions in dialogue and monologue Russian speech

"Dialogue" PMs			"Monologue" PMs		
Rank	PM function	IPM	Rank	PM function	IPM
1	H	5283	1	GH	6110
2	M	3317	2	H	4251
3	GH	2417	3	AH	2125
4	K	1717	4	RH	1528
5	RH	1167	5	G	1129
6	AH	867	6	D	332
7	G	550	7	DH	332
8	A	500	8	Z	332
9	N/A	383	9	N/A	266
10	D	333	10	GM	199

Table 6. The difference in distribution of PMs functions in dialogue and monologue speech

"Dialogue" PMs			"Monologue" PMs		
Rank	PM function	IPM difference	Rank	PM function	IPM difference
1	M	3184	1	GH	−3693
2	K	1650	2	AH	−1259
3	H	1033	3	G	−579
4	A	367	4	RH	−361
5	N/A	118	5	DH	−282
6	ARH	50	6	FH	−199
7	GR	50	7	Z	−149
8	GRH	50	8	AF	−133
9	AR	34	9	GDM	−133
10	KR	33	10	R	−99

It is seen from Table 5, that in both speech types, PMs of hesitation (H), boundary markers (G), and deictic markers (D), as well as bifunctional GH, AH, and RH are rather frequent. In general, it turned out that monofunctional use of PMs in dialogue speech is significantly higher (68.7%) than in monologue speech (37.4%).

According to Table 6, the most frequent "dialogue" functions are the following: (1) meta-communicative marker (M), (2) "xeno" marker (K), (3) hesitation marker (H), and (4) approximator (A). Among the polyfunctional PMs the rhythm-forming function is a typical component. It should be mentioned that in dialogue speech there occurred more cases when the experts could not attribute the PM function (N/A).

As for monologue speech, there are comparatively more boundary markers (G), replacement (Z), and rhythm-forming markers (R) here. Among the polyfunctional PMs the hesitation function prevails, probably because of the difference in communication scenarios of the SAT corpus [13].

5 Top Frequency Lists of Russian PMs for Speakers of Different Gender and Age Groups

This section reflects peculiarities of PMs usage in Russian everyday dialogue speech by different groups of speakers. Thus, Table 7 presents the top lists of PMs in male and female speech. In this table, as well as in the following one (Table 8), the numbers refer to the share of PMs in the correspondent social group.

Table 7. The top frequency lists of PMs in female and male Russian speech

Female speech			Male speech		
Rank	PM	%	Rank	PM	%
1	*vot*	15.92	1	*tam*	19.24
2	*govorit*	8.80	2	*vot*	10.20
3	*da*	7.40	3	*da*	8.45

(*continued*)

Table 7. (*continued*)

Female speech			Male speech		
Rank	PM	%	Rank	PM	%
4	*tam*	7.12	4	*koroche*	6.12
5	*kak by*	6.28	5	*tak*	4.37
6	*eto*	5.03	6	*kak by*	4.37
7	*eto samoe*	4.75	7	*tipa*	3.21
8	*znaesh'*	4.47	8	*eto samoe*	2.62
9	*tak*	2.93	9	*znaesh'*	2.62
10	*koroche*	2.37	10	*takie*	2.33

Table 8. The top frequency lists of PMs in Russian speech of different age groups

	Youth group		Middle-age group		Senior group	
Rank	PM	%	PM	%	PM	%
1	*vot*	12.06	*tam*	15.90	*vot*	21.77
2	*tam*	11.36	*vot*	10.88	*eto samoe*	8.87
3	*da*	8.22	*govorit*	10.04	*govorit*	6.85
4	*kak by*	6.99	*da*	7.53	*da*	6.85
5	*koroche*	5.59	*kak by*	6.69	*tam*	5.65
6	*govorit*	5.07	*znachit*	5.02	*eto*	5.65
7	*znaesh'*	4.02	*eto samoe*	5.02	*znaesh'*	5.24
8	*eto*	3.67	*tak*	4.18	*tak*	4.84
9	*ne znayu*	3.32	*eto*	3.35	*ponimaesh'*	2.82
10	*tipa*	3.15	*koroche*	2.51	*nu vot*	2.42

The limited volume of this article does not give us the opportunity to consider in detail all the differences observed in speech of different social groups, even for the top zone of frequency lists. However, we should mention high frequency of "tam" and "koroche" in male speech, whereas "govorit" and "kak by" are more peculiar to female speech. These results largely coincide with those obtained earlier [16–18]. As for age variation, "eto samoe" and "ponimaesh'" are mainly used by seniors, whereas "koroche" and "tipa" are more frequent in speech of youth group.

6   Conclusion

This study showed that PMs are indeed the integral elements of Russian spoken discourse. In speech of individual speakers, their share can reach up to 6.6% of the total number of words, and in individual speech fragments it can even exceed the share of significant units. The most common PM in almost all frequency lists is "vot", and in monologue speech, the high frequency of occurrence has shown PM "znachit".

PMs frequency lists analysis showed that we can confidently talk about statistically significant differences in use of PMs in dialogue and monologue. The most frequent PMs functions in speech of all groups of informants are metacommunicative, boundary-marking, and xeno-indicator. Pragmatic markers of these classes are often multifunctional and implement a number of additional functions.

Finally, the pilot annotation of the corpus data showed the qualitative heterogeneity of PMs, which manifests itself both in terms of the diversity of functions performed by them, and in terms of uniqueness of their identification and classification. The presented quantitative data may be used for improvement of NPL and discourse modeling systems.

References

1. Fraser, B.: Pragmatic markers. Pragmatics **6**(2), 167–190 (1996)
2. Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G., Baeva, E.: Towards a description of pragmatic markers in Russian everyday speech. In: Karpov, A., Jokisch, O., Potapova, R. (eds.) SPECOM 2018. LNCS (LNAI), vol. 11096, pp. 42–48. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99579-3_5
3. Schiffrin, D.: Discourse Markers. Cambridge University Press, Cambridge (1987)
4. Fraser, B.: An approach to discourse markers. J. Pragmat. **14**, 383–395 (1990)
5. Fischer, K., Drescher, M.: Methods for the description of discourse particles: contrastive analysis. Lang. Sci. **18**(3–4), 853–861 (1996)
6. Scheler, G., Fischer, K.: The many functions of discourse particles: a computational model of pragmatic interpretation. In: Proceedings of CogSci 1997 (1997)
7. Shourup, L.: Discourse markers. Lingua **107**, 227–265 (1999)
8. Bogdanova-Beglarian, N.V.: Pragmatems in Spoken Everyday Speech: Definition and General Typology [Pragmatemy v ustnoj povsednevnoj rechi: opredelenie pon'atia i obshchaja tipologia]. In: Perm University Herald. Russian and Foreign Philology [Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia], issue no. 3(27), pp. 7–20 (2014)
9. 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., et al. (eds.) Proceedings of the FRUCT'23, Bologna, Italy, 13–16 November 2018, pp. 69–77. FRUCT Oy, Finland (2018b). https://doi.org/10.23919/FRUCT.2018.8588101
10. 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. LNCS (LNAI), vol. 5729, pp. 250–257. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04208-9_36
11. Bogdanova-Beglarian, N., et al.: Sociolinguistic extension of the ORD corpus of Russian Everyday speech. In: Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS (LNAI), vol. 9811, pp. 659–666. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43958-7_80

12. Sherstinova, T.: Macro episodes of Russian everyday oral communication: towards pragmatic annotation of the ORD speech corpus. In: Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM 2015. LNCS (LNAI), vol. 9319, pp. 268–276. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23132-7_33

13. Bogdanova-Beglarian, N., Sherstinova, T., Zajdes, K.: Corpus "Balanced Annotated Text Library": methodology of multi-level analysis of the Russian monological speech [Korpus "Sbalansirovannaja Annotirovannaja Tekstoteka": metodika mnogourovnevogo analiza russkoj monologicheskoj rechi]. In: Analysis of Spoken Russian (AR3-2017). Proceedings of the 17th International Seminar, St. Petersburg, pp. 8–13 (2017)

14. Web: ELAN (Version 5.2) [Computer software]. Max Planck Institute for Psycholinguistics, Nijmegen. https://tla.mpi.nl/tools/tla-tools/elan/. Accessed 04 Apr 2019

15. Bogdanova-Beglarian, N., Blinova, O., Sherstinova, T., Martynenko, G.: Pragmatic markers annotation in Russian speech corpus: research problem, approaches and results. In: Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, 2019 International Conference on Computational Linguistics and Intellectual Technologies, Dialogue 2019, Moscow, Russian Federation, vol. 2019-May, issue no. 18, pp. 72–85 (2019)

16. Sherstinova, T.: The most frequent words in everyday spoken Russian (in the Gender Dimension and Depending on Communication Settings). In: Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, 2016 International Conference on Computational Linguistics and Intellectual Technologies, Dialogue 2016, Moscow, Russian Federation, 30 May 2018–2 June 2018, Moscow, RGGU, vol. 2016, issue no. 15(22), pp. 616–631 (2016)

17. Potapova, R.K., Potapov, V.V.: Language, Speech, Personality [Yazyk, rech', lichnost']. Yazyki slavyanskoj kul'tury, Moscow (2006)

18. Romaine S.: Corpus Linguistics and Sociolinguistics, Corpus Linguistics: An International Handbook, Mouton de Gruyter, Berlin, vol. 1, pp. 96–111 (2008)