Interdisciplinary Approach to the Study of Pragmatic Markers in Everyday Spoken Discourse

Natalia Bogdanova-Beglarian, Olga Blinova, Tatiana Sherstinova

Saint Petersburg State University, Russia

Abstract

Pragmatic markers (PMs) are indispensable elements of spoken discourse in any language. They are speech elements, having major influence on a pragmatic aspect of spoken discourse and being practically devoid of their own referential meaning. In spite of PMs wide circulation, they are very poorly studied. The current research demonstrates an interdisciplinary approach to study of PMs based on two representative speech corpora – ORD corpus of Russian Everyday Speech known as "One Day of Speech"-corpus and the "Balanced Annotated Collection of Texts" (SAT corpus). The research involves methodologies of different linguistics branches (phonetics, discourse analysis, sociolinguistics, psycholinguistics, corpus linguistics, etc.), making it possible to built formal statistical schemes which may be used both for theoretical linguistic studies and the improvement of NLP tasks.

Pragmatic Markers (PMs)

PMs definition

Pragmatic markers (PMs) are discourse units (words and multiword expressions) with a weakened referential meaning, which perform a variety of pragmatic tasks. PMs form a mandatory component of oral communication in any language. It is largely PMs that are responsible for the effectiveness of communication. However, unlike other lexical units that are well-represented in numerous dictionaries, pragmatic markers for many languages are still very poorly studied. In earlier papers on spoken discourse, PMs were considered within a wider class of *discourse particles* or *discourse markers*.

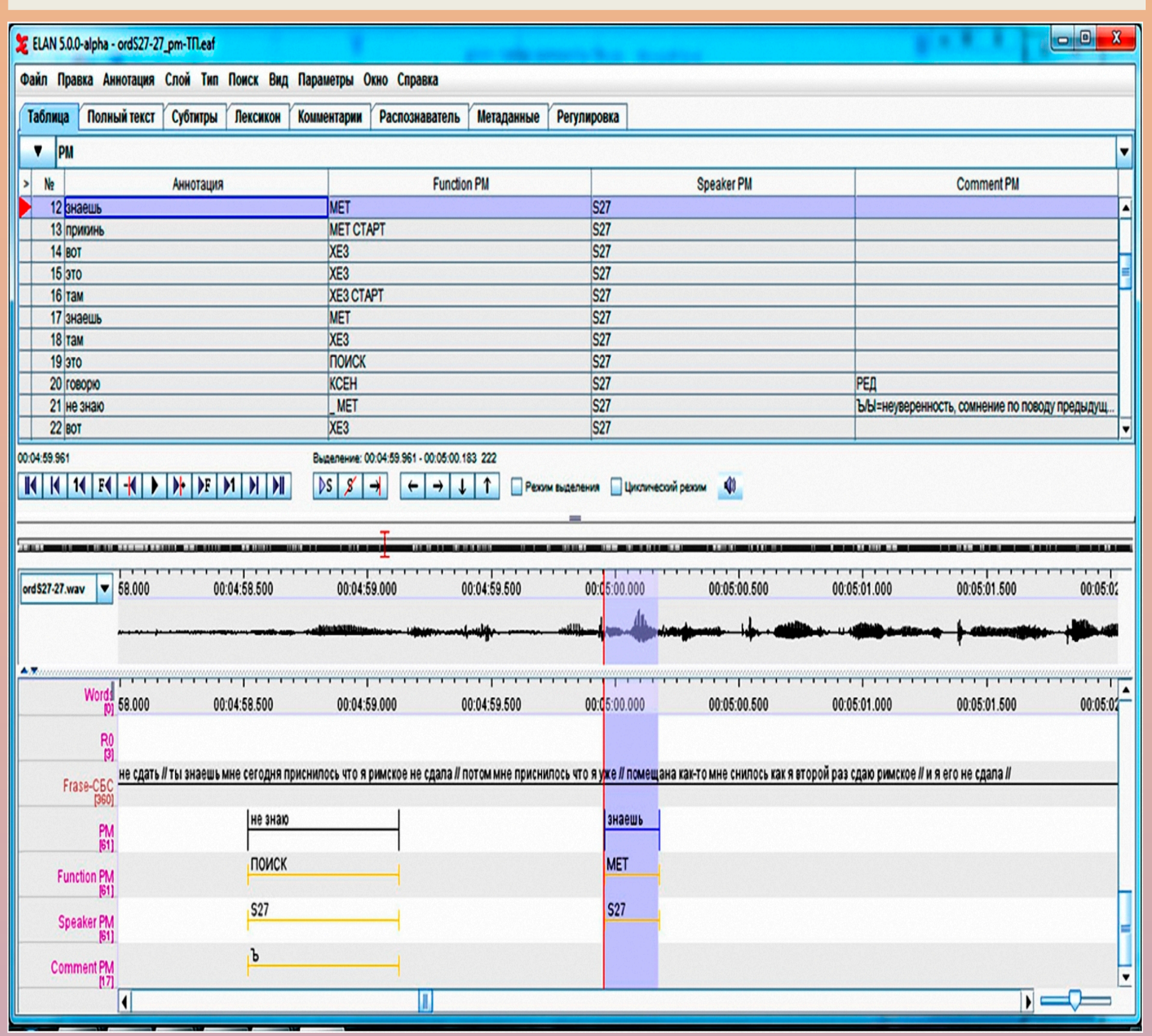
e.g.

English: well, you know, I think
French: comme ('like'), alors ('so'), bon ("well"), enfin ("well" / "I mean"), C'est-à-dire (que) ("in other words")
Italian: guarda ('look'), prego ('please'), dai ('come on')
Russian: вот ('well'), короче ('in short'), типа того ('sort of'), знаешь ('you know')

PMs Functions

A — **marker-approximator** ("tipa", "kak by", etc.);
G — **boundary marker**, including *starting*, *final*, and *navigational* markers ("vot", "koroche", etc.);
D — **deictic marker** ("vot etot vot", "vot takoj vot", etc.);
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");
K — **"xeno" marker** that introduces someone's speech ("tipa", "govorit", etc.);
M — **meta-communication marker** that refers to "communication about communication" ("znaesh", "vidish");
F — **"reflexive" marker** that expresses reflection on what is said ("tak skazat'");
R — **rhythm-forming marker** ("vot", "tam", etc.);
C — **marker of self-correction** ("v smysle", "vernej", etc.);
H — **hesitation markers** ("eto", "vot", "tam", etc.).

Annotation


ELAN 5.0.0-alpha - ordS27-27_pm-TП.eaf
Файл Правка Аннотация Слой Тип Поиск Вид Параметры Окно Справка
Таблица Полный текст Субтитры Лексикон Комментарии Распознаватель Метаданные Регулировка
PM
00:04:59.961
Выделение: 00:04:59.961 - 00:05:00.183 222
Режим выделения
Циклический режим
ordS27-27.wav
Words

№	Аннотация	Function PM	Speaker PM	Comment PM
12	знаешь	МЕТ	S27	
13	прикинь	МЕТ СТАРТ	S27	
14	вот	ХЕЗ	S27	
15	это	ХЕЗ	S27	
16	там	ХЕЗ СТАРТ	S27	
17	знаешь	МЕТ	S27	
18	там	ХЕЗ	S27	
19	это	ПОИСК	S27	
20	говорю	КСЕН	S27	РЕД
21	не знаю	_МЕТ	S27	ЪЫ=неуверенность, сомнение по поводу предыдущ...
22	вот	ХЕЗ	S27	

Research Data and Interdisciplinary Approach

ORD corpus

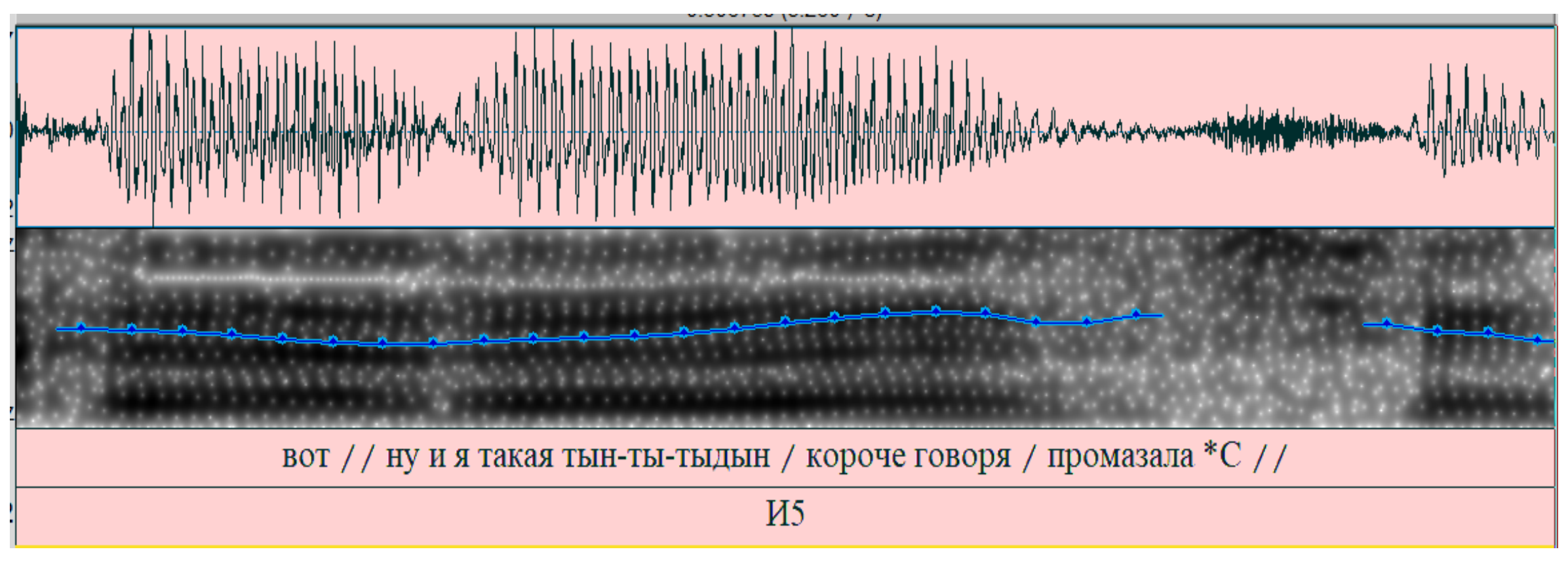
"ONE DAY OF SPEECH"

or ONE DAY WITH A VOICE RECORDER AROUND YOUR NECK

1400 hours of recordings
128 (+1000) participants
2850 macroepisodes
1 mln tokens in transcripts

Interdisciplinary Approach

The research involves methodologies of different linguistics branches: phonetic studies, discourse analysis, sociolinguistics, psycholinguistics, corpus linguistics, lexical studies, morphological and syntactic studies, communication studies, computational linguistics, etc.).



SAT corpus

SAT includes monologue speech recordings received from 5 professional groups: 1) doctors; 2) lawyers; 3) Russian teachers; 4) IT-specialists; 5) students, etc. of native Russian speakers. Texts were obtained in 4 experiments – reading, retelling, image description, storytelling.

Some Statistics

PMs in ORD corpus of everyday Russian: their share can reach **up to 6% of the total number of words** in speech of individual speakers; in some speech fragments PMs may even exceed the share of «standard words».

Russian Top List of PMs

1	*vot*	2483
2	*tam*	1950
3	*da*	1367
4	*govorit*	1167
5	*kak by*	1000
6	*eto*	733
7	*eto samoe*	717
8	*znaesh'*	683
9	*koroche*	633
10	*tak*	600

"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
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

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

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

Applications

The results of the project will find their practical application: **1) in the field of the applied linguistics, informational and speech technologies – to support the systems of automatic speech monitoring, voice search, speech synthesis and recognition systems, artificial intelligence, voice dialog systems when communicating with a computer or robot, 2) for teaching Russian as a foreign language, and 3) for conducting linguistic and forensic expertise based on audio records of speech communication.**

Acknowledgements

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