Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2019»

Москва, 29 мая — 1 июня 2019 г.

АННОТИРОВАНИЕ ПРАГМАТИЧЕСКИХ МАРКЕРОВ В РУССКОМ РЕЧЕВОМ КОРПУСЕ: ПРОБЛЕМЫ, ПОИСКИ, РЕШЕНИЯ И РЕЗУЛЬТАТЫ[1]

Богданова-Бегларян Н. В. (n.bogdanova@spbu.ru)
Блинова О. В. (o.blinova@spbu.ru)
Мартыненко Г. Я. (g.martynenko@spbu.ru)
Шерстинова Т. Ю. (t.sherstinova@spbu.ru)
Зайдес К. Д. (kristina.zaides@student.spbu.ru)
Попова Т. И. (tipopova13@gmail.com)

Филологический факультет СПбГУ,
Санкт-Петербург, Россия

В статье описывается опыт аннотирования прагматических маркеров (ПМ) в двух русских речевых корпусах: «Один речевой день» (ОРД; диалоги) и «Сбалансированная аннотированная текстотека» (САТ; монологи). Для подготовки сплошной разметки ПМ было проведено 4 пилотных аннотирования на выборках из ОРД и САТ, что позволило сформировать итоговый список ПМ: 450 единиц, представляющих собой варианты 53 базовых структурных типов. В ходе обработки результатов пилотного аннотирования удалось получить предварительные данные о частоте встречаемости отдельных прагматических маркеров и их типов, а также о зависимости употребления ПМ от пола и уровня речевой компетенции говорящего. В результате обработки данных были получены частотные списки как самих ПМ, так и выполняемых ими функций.

Ключевые слова: русская повседневная речь, речевой корпус, прагматический маркер, корпусная разметка, монолог, диалог

[1] Исследование выполнено при поддержке гранта РНФ «Система прагматических маркеров русской повседневной речи» (проект № 18-18-00242).

PRAGMATIC MARKERS ANNOTATION IN RUSSIAN SPEECH CORPUS: RESEARCH PROBLEM, APPROACHES AND RESULTS

Bogdanova-Beglarian N. V. (n.bogdanova@spbu.ru)
Blinova O. V. (o.blinova@spbu.ru)
Martynenko G. Ya. (g.martynenko@spbu.ru)
Sherstinova T. Yu. (t.sherstinova@spbu.ru)
Zaides K. D. (kristina.zaides@student.spbu.ru)
Popova T. I. (tipopova13@gmail.com)

Philological Faculty of St. Petersburg State University,
St. Petersburg, Russia

The article describes the experience of pragmatic markers (PM) annotation in two Russian speech corpora: “One Speaker’s Day” (ORD; dialogues) and “Balanced Annotated Textotec” (SAT; monologues). To prepare an optimal PM annotation scheme, 4 pilot annotations were conducted on samples from ORD and SAT. It made it possible to form the final list of PM: 450 units, representing variants of 53 basic structural types. Processing the results of the pilot annotation allowed to obtain preliminary data on frequency of individual pragmatic markers and their types, as well as on the dependence of PM usage on sex and the level of speech competence of the speaker. As a result of statistical data processing, frequency lists of both PMs and their functions were obtained. The most commonly used in the dialogue are the PM *вот*, which is usually used as a «boundary marker» (G), and the PM *там*, which is usually used as a hesitative and/or rhythm-forming marker. In the monologue, the upper zone of the frequency list of the PMs is also full of boundary markers (G), marking the beginning/end of the monologue or serving as navigators in the text (*вот/ну вот, значит, так*). The most frequent types of PMs in dialogue are: X (hesitative markers), M (meta-communicative marker), GX (boundary/hesitative marker), K (xeno-indicator marker that introduces someone’s speech), RX (rhythm-forming/hesitative marker). In the list of the most frequent types of PMs in monologue speech, the markers of the type GX (boundary/hesitative marker) and X (hesitative marker) are in the lead. The analysis of the frequency lists of PMs showed that we can talk about statistically significant differences in the use of PMs in dialogue and monologue.

Keywords: Russian everyday speech, speech corpus, pragmatic marker, corpus annotation, monologue, dialogue

1. Введение

Обработка естественного языка, организованного в корпус, предполагает присвоение компонентам текстов знаков аннотации [Захаров 2005]; [Плунгян 2008]. Традиционные виды аннотации (морфологическая, синтаксическая и др.) реализуются в корпусах различного типа (см., например: [Gries, Berez 2017]), в том числе — с помощью различных программ (см., например: [Kuzmenko 2017]).

Однако существует разметка, провести которую автоматически крайне сложно. Речь идет о таких элементах структуры устного дискурса, которые уместно назвать *прагматемами*, или *прагматическими маркерами* (ПМ) [Богданова-Бегларян 2014]. ПМ активно функционируют в нашей речи: говорящий использует их, вербализуя трудности речепорождения, подыскивая нужное слово или производя метаязыковое комментирование сказанного. Автоматическое аннотирование ПМ затруднено прежде всего тем, что внешне они ничем не отличаются от значимых единиц и лишь в контексте реализуют свой новый статус, появляющийся в результате процесса *прагматикализации* (см. подробнее, например: [Bogdanova-Beglarian, Filyasova 2018]). Разметка ПМ в речевом корпусе не является тривиальной и требует значительной ручной работы экспертов-филологов (о путях решения возникающих проблем см.: [Zaides et al. 2018]).

2. Материал и методика

Материалом для выявления инвентаря ПМ русской устной речи и построения такой их типологии, которая была бы пригодна для аннотирования больших массивов данных, стали два корпуса (см. о начале этой работы: [Bogdanova-Beglarian et al. 2018]).

1. Корпус повседневной русской речи «Один речевой день» (ОРД), один из наиболее представительных на сегодняшний день ресурсов для анализа русского устного дискурса ([Русский язык… 2016]; [Bogdanova-Beglarian et al. 2016a, b, 2017]; [Богданова-Бегларян и др. 2017а]).
2. Корпус «Сбалансированная аннотированная текстотека» (САТ), включающий записи монологической речи, полученные от разных профессиональных групп носителей языка. Все тексты в САТ построены в рамках 4-х коммуникативных сценариев (чтение[2], пересказ, описание изображения, рассказ (см.: [Звуковой корпус… 2013], [2014], [2015]; [Богданова-Бегларян и др. 2017б]).

За основу словника ПМ была взята типология прагматем Н. В. Богдановой-Бегларян [Богданова-Бегларян 2014]. Слегка переработанный, данный словник был расширен за счет всех возможных структурных вариантов (расширений

[2] О неподготовленном ч*тении* как разновидности спонтанного монолога см.: [Звуковой корпус… 2013].

базовой единицы), а также с учетом всех грамматических форм ПМ — для удобства их автоматического поиска по транскрипту и сведения в единую базу данных. В этот список попали как относительно частотные прагматические маркеры (*это самое, (ну) (я) не знаю, такой, короче, значит, (и) всё такое (прочее)* и др.), так и менее частотные, но регулярно употребляемые в устной речи (*типа того (что), боюсь (что), вроде, как бы, (и) всё такое (прочее)* и др.). Общий список изначально насчитывал 65 единиц. Тем самым был определен предварительный инвентарь единиц, подлежащих аннотированию в материалах монолога (САТ) и диалога (ОРД).

3. Подходы к аннотированию ПМ

Первичное аннотирование ПМ в корпусном материале проводилось непосредственно в среде ELAN, поддерживающей «привязку» разметки к определенному сегменту звукового сигнала. При этом было введено четырехуровневое аннотирование:

- уровень 1. PM — ПМ в той форме, как он представлен в транскрипте.
- уровень 2. Function PM — основные и дополнительные функции.
- уровень 3. Speaker PM — код говорящего.
- уровень 4. Comment PM — уровень комментариев.

Далее был пересмотрен и сокращен перечень функций ПМ; используемые для обозначения функций коды упрощены до однобуквенных, снято требование выделения основной функции:

- А — маркер-аппроксиматор,
- Г — разграничительный маркер (стартовый, финальный и навигационный),
- Д — дейктический маркер,
- З — все виды маркеров-заместителей (чужой речи, ряда перечисления или их частей),
- К — маркер-ксенопоказатель,
- М — маркер-метакоммуникатив,
- Ф — маркер-рефлексив,
- Р — ритмообразующий маркер,
- С — маркер самокоррекции,
- Х — хезитативный маркер.

Было введено также понятие базового варианта ПМ, который и указывался при аннотировании, что способствовало получению более однородной разметки.

Переработанная методика прошла успешную апробацию на материале второго этапа аннотирования [Bogdanova-Beglarian et al. 2018] и была адаптирована для разметки корпуса САТ.

4. Аннотирование ПМ в диалогической и монологической речи

Различия в подходах к аннотированию ПМ в речи разного типа вызваны тем, что в двух использованных корпусах данные представлены принципиально различно: в корпусе ОРД используется разметка, выполненная в среде ELAN (формат *.eaf), а корпус САТ не имеет многоуровневой разметки и представляет собой массив звуковых файлов и файлов расшифровки формата *.doc.

Для аннотирования ПМ в обоих корпусах были подготовлены две выборки материала. Для анализа *диалогической речи* было отобрано 149 эпизодов из корпуса ОРД, записанных от 98 информантов (вместе с их коммуникантами) (всего 308 905 словоупотреблений). В подкорпус вошли эпизоды «речевых дней» информантов разных профессиональных групп, преимущественно ситуации неформального общения. В группе информантов — 45 женщин (46%) и 53 мужчины (54%). Для исследования монологической речи из корпуса САТ были отобраны тексты разного типа, записанные от 34 информантов (всего 50 128 словоупотреблений). В подкорпусе представлена речь информантов, принадлежащих к двум профессиональным группам, — юристов и медиков.

4.1. Процедура пилотного аннотирования

Всего было осуществлено 4 этапа пилотного аннотирования материала.

Первое пилотное аннотирование было выполнено на выборке из корпуса ОРД объемом в 16 000 словоупотреблений параллельно 4-мя экспертами, по правилам, разработанным на подготовительном этапе. В ходе разметки использовался расширенный список функций ПМ, при этом аннотаторы выделяли главную из них и помещали соответствующий тег на первое место в боксе уровня «Function PM». Дополнительные функции перечислялись далее в алфавитном порядке. На уровне «Comment PM» отмечались некоторые дополнительные особенности употребления маркеров: например, редукция формы ксенопоказателя *говорит* до *грит* или *гыт* или особое интонационное оформление ПМ.

Сами теги представляли собой обозначения соответствующей функции. Возможные новые ПМ, а также различные варианты уже имеющихся в списке отмечались с помощью специальной пометы на уровне «Comment PM».

Анализ результатов первого пилотного аннотирования показал, что инструкция по разметке требует доработки. В ходе подготовки инструкции для *второго пилотного аннотирования* было решено использовать более короткий список функций ПМ и перечислять основные и дополнительные функции в одном ряду по алфавиту, поскольку практически каждый ПМ в устной речи оказывался полифункциональным, а иерархия выполняемых им функций при разметке выстраивалась не всегда однозначно и единогласно. Анализ результатов позволил оптимизировать методику и выработать более эффективную инструкцию для разметчиков [Bogdanova-Beglarian et al. 2018]. Переработанная методика прошла успешную апробацию на втором этапе аннотирования и сохранялась без существенных изменений на третьем и четвертом этапах.

Третье пилотное аннотирование было проведено на подкорпусе САТ (15 000 словоупотреблений). Оно выполнялось для предварительной оценки особенностей употребления ПМ в монологической речи и позволило сопоставить частоту употребления ПМ в зависимости от УРК[3] говорящего.

Четвертое пилотное аннотирование было проведено на подкорпусе ОРД (60 000 словоупотреблений). Выполнялось оно для предварительной оценки особенностей употребления ПМ в диалогической речи и позволило сделать некоторые выводы об особенностях использования ПМ в мужской и женской речи.

В конце каждого этапа пилотного аннотирования осуществлялась экспертная корректура прагматической разметки, пересматривался и дополнялся перечень выделяемых ПМ. На данный момент рабочий список вариантов ПМ насчитывает 450 единиц, представляющих собой варианты 53 базовых структурных типов.

5. Некоторые количественные характеристики ПМ в диалогической и монологической речи (сравнение ОРД и САТ)

Обработка результатов аннотирования ПМ в корпусном материале позволила получить данные о частоте встречаемости отдельных прагматических маркеров, а также о зависимости употребления ПМ от характеристик говорящего. Статистическая обработка результатов третьего и четвертого этапа пилотного аннотирования позволила получить выводы относительно наиболее употребительных ПМ и их функциях. Приведем некоторые из полученных данных.

5.1. Прагматические маркеры в ОРД и САТ

Частотные списки ПМ представлены в табл. 1, где приведены: ранги, частоты ПМ в абсолютных цифрах, доли конкретных ПМ от всех ПМ в выборке (в %), доли конкретных ПМ от всех слов выборки (в %) и ipm.

Размеры выборки ОРД (диалогическая речь) — 60 000 словоупотреблений. Размеры выборки САТ (монологическая речь) — 15 000 словоупотреблений.

Самыми употребительными в диалоге ПМ оказались: *вот*, чаще всего выступающий как дискурсивный маркер Г, и *там*, выступающий, как правило, в роли хезитативного и/или ритмообразующего маркера. Входят в эту зону также метакоммуникативы (М) *да* и *знаешь*, хезитативы (Х) *как бы*, *это*, *это*

[3] *Уровень речевой компетенции* определяется как степень свободы говорящего в выборе речевых средств, уровень его владения языковыми возможностями, его способность решать те или иные коммуникативные задачи. УРК коррелирует с двумя социальными характеристиками говорящего: высшее образование + профессиональное отношение к речи (преподаватели, актеры, лекторы, дикторы, политики…) → высокий УРК; высшее образование + непрофессиональное отношение к речи → средний УРК; отсутствие высшего образования + непрофессиональное отношение к речи → низкий УРК. Как убедительно показал анализ материала, реальные лингвистические корреляты имеют только полярные типы — высокий и низкий УРК [Звуковой корпус… 2013].

самое, *короче* и *так* и маркер-ксенопоказатель (К) *говорит* (чаще редуцированный). Частотность маркеров типа М в ОРД не случайна: в диалоге говорящие действительно постоянно вынуждены обращаться к собеседнику, так или иначе привлекая, а затем и удерживая его внимание, или передавать чужую речь.

Таблица 1 Наиболее частотные ПМ в ОРД и САТ

ранг	ПМ	f (ПМ)	доля от ПМ (%)	доля по выборке (%)	ipm
ОРД					
1	*вот*	149	14,06	0,25	2483
2	*там*	117	11,04	0,20	1950
3	*да*	82	7,74	0,14	1367
4	*говорит*	70	6,60	0,12	1167
5	*как бы*	60	5,66	0,10	1000
6	*это*	44	4,15	0,07	733
7	*это самое*	43	4,06	0,07	717
8	*знаешь*	41	3,87	0,07	683
9	*короче*	38	3,58	0,06	633
10	*так*	36	3,40	0,06	600
САТ					
1	*вот*	139	51,48	0,92	9232
2	*значит*	15	5,56	0,10	996
3	*так*	15	5,56	0,10	996
4	*там*	13	4,81	0,09	863
5	*как бы*	12	4,44	0,08	797
6	*ну вот*	12	4,44	0,08	797
7	*всё*	4	1,48	0,03	266
8	*и так далее*	4	1,48	0,03	266
9	*вот так вот*	3	1,11	0,02	199
10	*ну так*	3	1,11	0,02	199

В корпусе САТ верхняя зона частотного списка ПМ полна маркеров типа Г, маркирующих начало/конец монолога или служащих навигаторами по тексту: *вот/ну вот, значит, так*. Присутствуют в этой зоне и дейктические маркеры (Д) — *вот так вот*.

Для оценки различий между данными, полученными после составления частотных списков ПМ в ОРД и САТ, был использован тест Манна-Уитни; применялась программная среда R [R Core Team 2019]; сравнению подвергались значения ipm одних и тех же ПМ, см. **табл. 2**:

Таблица 2 Таблица для оценки различий между ОРД и САТ (фрагмент)

ПМ	ipm САТ	ipm ОРД
вот	9231,587	2483,333
значит	996,2144	350
так	996,2144	600
там	863,3858	1950
как бы	796,9715	1000
ну вот	796,9715	350

В результате получены значения W = 2082, p-value = 1,077e-06, то есть $p < 0{,}001$. Таким образом, различия в употреблении ПМ между ОРД и САТ (т. е. в диалогической и монологической речи) можно признать статистически значимыми.

В ходе аннотирования помечались все встретившиеся в материале типы ПМ, как «чистые», так и «смешанные» (полифункциональные употребления) (АГ, АГХ, ГРХ, АФ и т. п.), отражающие общую полифункциональность ПМ, что весьма свойственно устной речи. Оказалось, что монофункциональных употреблений ПМ в ОРД существенно больше (68,7%), чем в САТ (37,4%). В ОРД чаще всего монофункциональными выступают такие ПМ, как Ф (рефлексив) (100,0%), М (метакоммуникатив) (93,0%) и З (маркер-заместитель) (91,7%). В САТ — К (ксенопоказатель) и Д (дейктический маркер) (по 100,0%). Среди полифункциональных преобладают ПМ, выполняющие, среди прочего, хезитационную функцию (АХ, АГХ, АКХ, ГХ, РХ и под.).

Наиболее употребительными в ОРД оказались прагматические маркеры Х (5283 ipm), М (3317 ipm) и ГХ (2417 ipm). В САТ — ГХ (6110 ipm), Х (4250 ipm) и АХ (2125 ipm). Видно, что среди маркеров, наиболее распространенных в САТ, преобладают ПМ хезитативного типа. В ОРД отчетливо преобладают метакоммуникативы (М): в диалоге говорящие часто вынуждены обращаться к собеседнику, так или иначе привлекая, а затем и удерживая его внимание, или передавать (пересказывать) чужую речь.

5.2. Функции прагматических маркеров в ОРД

Верхняя зона частотного списка типов ПМ в подкорпусе ОРД включает в себя следующие разновидности (см. **табл. 3**, в 23 случаях аннотаторы не смогли приписать ПМ функцию, такие случаи обозначены как NA).

Количество ПМ в речи отдельных информантов колеблется от 1 (0,09% от всего объема речевого материала по данному информанту) до 70 (И118; 6,6% от объема речевого материала информанта в выборке).

Таблица 3 Наиболее частотные типы прагматических маркеров диалогической речи

ранг	функция	кол-во	ipm	ранг	функция	кол-во	ipm
1	Х	317	5283	11	З	11	183
2	М	199	3317	12	Ф	9	150
3	ГХ	145	2417	13	ГМ	7	117
4	К	103	1717	14	МХ	7	117
5	РХ	70	1167	15	Р	6	100
6	АХ	52	867	16	АР	6	100
7	Г	33	550	17	ДХ	3	50
8	А	30	500	18	АРХ	3	50
9	NA	23	383	19	ГР	3	50
10	Д	20	333	20	ГРХ	3	50

В **табл. 4** представлены данные, отражающие количество самых частотных функциональных типов ПМ в речи женщин и количество ПМ с теми же функциями в речи мужчин.

Таблица 4. Наиболее частотные типы ПМ в диалогической речи мужчин и женщин (фрагмент)

функция	ЖЕН	МУЖ
Х	216	101
М	143	56
ГХ	91	54
К	84	19
АХ	37	15
РХ	28	42
Г	24	9
А	17	13
Д	17	3
NA	11	12
З	8	3

Оценка статистической значимости различий в употребительности ПМ с разными функциями в речи мужчин и женщин выполнена с помощью критерия «хи-квадрат». Такая оценка показала, что различия являются статистически значимыми (X-squared = 273,18; $p < 0{,}001$), но аппроксимация может быть неправильной, то есть гипотеза о наличии значимых различий нуждается в дальнейшей проверке с привлечением большего объема размеченных данных. В речи женщин употребляются ПМ с 25 разными тегами функций, в речи мужчин — с 20 разными тегами, при этом 8 тегов в речи мужчин не встречается вовсе (АР, ГРХ, ДХ, КР, АГ, АГХ, ЗХ, МС).

5.3. Функции прагматических маркеров в САТ (зависимость от УРК говорящего)

Во всех частотных списках типов ПМ (общем и по трем УРК) первые две позиции уверенно занимают маркеры типа ГХ (пограничный маркер/хезитатив) и Х (хезитатив), ср.:

- общий список: 34,07 и 23,7 % соответственно;
- высокий УРК: 23,6 и 22,47 %;
- средний УРК: 38,89 и 19,44 %;
- низкий УРК: 39,45 и 27,52 %.

Видно, что поиск нужного слова или продолжения монолога, а также стремление просто выстроить связный текст — это главное в механизме спонтанного порождения устного текста, что вынуждает говорящего обращаться к специальным единицам — прагматическим маркерам соответствующих типов. Видно также, что доля таких ПМ возрастает по мере снижения УРК говорящего.

На третьем месте в трех списках из четырех (исключение — средний УРК) — маркер типа АХ, аппроксиматор/хезитатив, с помощью которого говорящий выражает и речевое колебание (чаще всего — поиск), и свою неуверенность в том, что подобрал нужное слово или верно выражает мысль. Количество таких ПМ в монологах достаточно велико:

- общий список: 11,85 %;
- высокий УРК: 15,73 %;
- низкий УРК: 11,93 %.

В речи информантов со средним УРК маркер АХ отошел на четвертую позицию, уступив место типу Г — маркерам начала/конца монолога или навигаторам по тексту.

Очевидно, что говорящие с любым УРК в равной степени испытывают трудности при спонтанном речепорождении и преодолевают эти трудности с помощью более или менее единого набора ПМ.

Анализ частотных списков ПМ в речи информантов САТ с разным УРК позволил сделать ряд наблюдений.

Так, в целом ПМ составили 1,8 % от общего массива слов в монологах-рассказах женщин-медиков (270 употреблений). Больше всего ПМ пришлось на группу информантов со средним УРК (группа Б, 1,39 %; 72 употребления), минимум — на группу с низким УРК (группа В, 0,92 %; 109 употреблений). Доля ПМ в группе информантов с высоким УРК (группа А) — 1,12 % (89 употреблений).

Во всех частотных списках (общем и по трем УРК) первое место уверенно занимает маркер *вот*, доля которого во всех случаях близка к 50 %: общий список — 51,48 %, высокий УРК — 48,31 %, средний УРК — 58,33 %, низкий УРК — 49,54 %.

Второе место в общем частотном списке ПМ занимает маркер *значит* (5,56 %), чаще свидетельствующий о низком УРК. Наши данные полностью подтвердили это предположение: употреблений *значит* в роли ПМ совсем не обнаружилось в речи информантов группы А, в группах же Б и В он занимает также второе место в соответствующих частотных списках (6,94 и 8,26 %), что

и обеспечило ему общее второе место по корпусу. Видно также, что по мере снижения УРК доля *значит* заметно возрастает.

Во всех четырех списках присутствуют и маркеры *как бы, ну вот, там, так* — по всей видимости, они более всего нужны любому говорящему для построения спонтанного монолога и менее всего при этом способны диагностировать УРК человека. *Так* и *там* в роли ПМ более всего представлены в речи информантов из группы В (низкий УРК — 7,34 и 6,42% соответственно; данные по всем монологам — 5,56 и 4,81%), Употреблений *как бы* и *ну вот* в роли ПМ больше всего в речи информантов из группы А (6,74 и 5,62%; общие данные — 4,44 и 4,44%). Средний УРК в рассматриваемом отношении ничем не примечателен.

Обращает на себя внимание также дейктический маркер *вот (…) вот*. В варианте *вот так вот* он присутствует в верхней зоне трех частотных списков ПМ: общем (1,11%), высокого УРК (2,25%) и низкого УРК (0,92%). В речи информантов группы Б (средний УРК) его в этой зоне не обнаружилось, в речи же информантов из группы В (низкий УРК) он представлен еще двумя структурными вариантами: *вот сейчас бы вот* и *вот эта вот* (по 0,92%).

6. Заключение

Проведенное исследование показало, что ПМ действительно представляют собой неотъемлемые элементы русского устного дискурса. В речи отдельных говорящих их доля может доходить до 6,6% от общего количества словоупотреблений, а в отдельных речевых фрагментах даже превышать долю значимых единиц.

Анализ частотных списков ПМ (общих по обоим корпусам и отдельных для разных групп говорящих) показал, что можно уверенно говорить о статистически значимых различиях в употреблении ПМ в диалоге и монологе.

Наиболее частотными функциями ПМ в речи всех групп информантов являются метакоммуникативная, разграничительная (дискурсивная), хезитативно-поисковая, и функция ксенопоказателя. Прагматические маркеры этих классов часто оказываются полифункциональными и реализуют ряд дополнительных функций.

Наиболее распространенным ПМ во всех частотных списках оказался *вот* (чаще — в разграничительной функции). В монологической речи высокую частоту встречаемости проявил маркер *значит* (как правило, в разграничительной или хезитативной функциях).

На основании полученных данных можно предварительно предположить, что частота использования ПМ в речи коррелирует с УРК говорящего: в среднем, чем он выше, тем меньше используется ПМ определенных типов, свидетельствующих о больших затруднениях говорящего в построении дискурса.

Все ПМ являются неизбежными элементами устной спонтанной речи, однако одни («хорошие» ПМ) не снижают качества речи, свидетельствуют об умении говорящего преодолевать естественные речевые сбои и не мешают восприятию и пониманию (высокий УРК), другие («плохие» ПМ) настолько ломают

структуру устного текста, что затрудняют понимание и свидетельствуют о низком качестве речи и неумении говорящего выстраивать связный устный текст (низкий УРК).

Наконец, пилотное аннотирование корпусного материала показало качественную неоднородность ПМ, проявляющуюся как в плане разнообразия выполняемых ими функций, так и в плане однозначности их выделения и отнесения этих единиц к прагматическим элементам устного дискурса, поэтому одной из перспективных задач предложенного направления исследования русской устной речи представляется выявление качественной дифференциации прагматических маркеров.

Литература

1. *Богданова-Бегларян Н. В.* (2014) Прагматемы в устной повседневной речи: определение понятия и общая типология // Вестник Пермского университета. Российская и зарубежная филология. — Вып. 3 (27), 2014. — С. 7–20.
2. *Богданова-Бегларян Н. В., Шерстинова Т. Ю., Блинова О. В., Мартыненко Г. Я.* (2017а) Корпус «Один речевой день» в исследованиях социолингвистической вариативности русской разговорной речи // Анализ разговорной русской речи (АРЗ–2017): Труды седьмого междисциплинарного семинара / Науч. ред. Д. А. Кочаров, П. А. Скрелин. — СПб.: Политехника-принт, 2017. — С. 14–20.
3. *Богданова-Бегларян Н. В., Шерстинова Т. Ю., Зайдес К. Д.* (2017б) Корпус «Сбалансированная Аннотированная Текстотека»: методика многоуровневого анализа русской монологической речи // Анализ разговорной русской речи (АРЗ–2017): Труды седьмого междисциплинарного семинара / Науч. ред. Д. А. Кочаров, П. А. Скрелин. — СПб.: Политехника-принт, 2017. — С. 8–13.
4. *Захаров В. П.* (2005) Корпусная лингвистика: Учебно-методическое пособие. — СПб.: СПбГУ, 2005. — 48 с.
5. *Звуковой корпус как материал для анализа русской речи* (2013) Коллективная монография. Часть 1. Чтение. Пересказ. Описание / Отв. ред. Н. В. Богданова-Бегларян. — СПб.: Филологический ф-т СПбГУ, 2013. — 532 с.
6. *Звуковой корпус как материал для анализа русской речи* (2014) Коллективная монография Часть 2. Теоретические и практические аспекты анализа. Том 1. О некоторых особенностях устной спонтанной речи разного типа. Звуковой корпус как материал для преподавания русского языка в иностранной аудитории / Отв. ред. Н. В. Богданова-Бегларян. — СПб.: Филологический ф-т СПбГУ, 2014. — 396 с.
7. *Звуковой корпус как материал для анализа русской речи* (2015) Коллективная монография. Часть 2. Теоретические и практические аспекты анализа. Том 2. Звуковой корпус как материал для новых лексикографических проектов / Отв. ред. Н. В. Богданова-Бегларян. — СПб.: Филологический ф-т СПбГУ, 2015. — 364 с.
8. *Плунгян В. А.* (2008) Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. — № 16 (2), 2008. — С. 7–20.

9. *Русский язык повседневного общения: особенности функционирования в разных социальных группах* (2016) Коллективная монография / Отв. ред. Н. В. Богданова-Бегларян. — СПб.: ЛАЙКА, 2016. — 244 с.

References

1. *Bogdanova-Beglarian, N. V.* (2014), Pragmatems in Spoken Everyday Speech: Definition and General Typology [Pragmatemy v ustnoj povsednevnoj rechi: opredelenie pon'atia i obshchaja tipologia] // Perm University Herald. Russian and Foreign Philology [Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia]. Iss. 3 (27), pp. 7–20.
2. *Bogdanova-Beglarian, N., Baeva, E., Blinova, O., Martynenko, G., Sherstinova T.* (2018), Towards a Description of Pragmatic Markers in Russian Everyday Speech // Speech and Computer. SPECOM 2018. Lecture Notes in Computer Science, vol. 11096. Springer, Cham / Karpov, A., Jokisch, O., Potapova, R. (eds.), pp. 42–48.
3. *Bogdanova-Beglarian, N., Blinova, O., Martynenko, G., Sherstinova T., Zaides, K.* (2018), Pragmatic Markers in Russian Spoken Speech: an Experience of Systematization and Annotation for the Improvement of NLP Tasks // Proceedings of the FRUCT'23. Bologna, Italy, 13–16 November 2018 / S. Balandin, T. Salmon Cinotti, F. Viola, T. Tyutina (eds.). FRUCT Oy, Finland, pp. 69–77.
4. *Bogdanova-Beglarian, N. V., Blinova, O. V., Sherstinova, T. Iu., Martynenko, G. Ja.* (2017a), 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] // Analysis of Spoken Russian (AR3–2017). Proceedings of the seventh interdisciplinary seminar [Trudy sed'mogo mezhdisciplinarnogo seminara]. St. Petersburg, pp. 14–20.
5. *Bogdanova-Beglarian, N., Filyasova, Yu.* (2018), Active Processes in Modern Spoken Language (Evidence from Russian) // Digital Transformation and Global Society. Third International Conference, Conference proceedings DTGS 2018, St. Petersburg, Russia, May 30 — June 2, 2018, Revised Selected Papers, Part II. Communications in Computer and Information Science (CCIS). Vol. 859 / D. A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Yu. Kabanov, O. Koltsova (eds.). Springer, Cham, pp. 391–403.
6. *Bogdanova-Beglarian N., Sherstinova T., Blinova O., Baeva E., Martynenko G., Ryko A.* (2016b), Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech // SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, pp. 659–666.
7. *Bogdanova-Beglarian, N. V., Sherstinova, T. Iu., Zajdes, K. D.* (2017b), Corpus "Balanced Annotated Text Library": Methodology Multi-Level Analysis of the Russian Monological Speech [Korpus «Sbalansirovannaja Annotirovannaja Tekstoteka»: metodika mnogourovnevogo analiza russkoj monologicheskoj rechi] // Analysis of Spoken Russian (AR3–2017). Proceedings of the seventh interdisciplinary seminar. Trudy sed'mogo mezhdisciplinarnogo seminara. St. Petersburg, pp. 8–13.
8. *Everyday Russian Language: Functioning Features in Different Social Groups* (2016), [Russkij jazyk povsednevnogo obshchenia: osobennosti funkcionirovania

v raznykh social'nykh gruppakh]. Bogdanova-Beglarian, N. V. (ed.). Collective Monograph, St. Petersburg, 244 p.
9. *Gries, S. T., Berez A. L.* (2017), Linguistic Annotation in/for Corpus Linguistics / N. Ide and J. Pustejovsky (eds.). Handbook of Linguistic Annotation. Berlin & New York: Springer, pp. 379–409.
10. *Kuzmenko, E.* (2017) Morphological Analysis for Russian: Integration and Comparison of Taggers / Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol. 661. Springer, Cham, pp.162–171.
11. *Plungjan, V. A.* (2008), Corpus as a Tool and as an Ideology: on Some Lessons of Modern Corpus Linguistics [Korpus kak instrument i kak ideologia: o nekotorykh urokakh sovremennoj korpusnoj lingvistiki] // Russian Language in Scientific Description [Russkij jazyk v nauchnom osveshchenii]. 16 (2), pp. 7–20.
12. *R Core Team* (2019), R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
URL: https://www.R-project.org/.
13. *Speech Corpus as a Base for Analysis of Russian Speech* (2013) Collective Monograph. Part 1. Reading. Retelling. Description [Zvukovoj korpus kak material dl'a analiza russkoj rechi: kollektivnaja monografia. Chast' 1. Chtenie. Pereskaz. Opisanie] / N. V. Bogdanova-Beglarian (ed.). St. Petersburg, 532 p.
14. *Speech Corpus as a Base for Analysis of Russian Speech* (2014) Collective Monograph. Part 2. Theory and Practice of Speech Analysis. Vol. 1. Some Features of Oral Spontaneous Speech of Various Types. Speech Corpus as a Base for Material for the Teaching of Russian as a Foreign Language [Zvukovoj korpus kak material dl'a analiza russkoj rechi. Kollektivnaja monografia. Chast' 2. Teoreticheskie i prakticheskie aspekty analiza. Tom 1. O nekotorykh osobennost'akh ustnoj spontannoj rechi raznogo tipa. Zvukovoj korpus kak material dl'a prepodavania russkogo jazyka v inostrannoj auditorii] / N. V. Bogdanova-Beglarian (ed.). St. Petersburg, 396 p.
15. *Speech Corpus as a Base for Analysis of Russian Speech* (2015) Collective Monograph. Part 2. Theory and Practice of Speech Analysis. Vol. 2. Speech Corpus as a Base for New Lexicographical Projects [Zvukovoj korpus kak material dl'a analiza russkoj rechi. Kollektivnaja monografia. Chast' 2. Teoretichskie i prakticheskie aspekty analiza. Tom 2. Zvukovoj korpus kak material dl'a novykh leksikograficheskikh proektov] / N. V. Bogdanova-Beglarian (ed.). St. Petersburg, 364 p.
16. *Zaides, K., Popova, T., Bogdanova-Beglarian, N.* (2018), Pragmatic Markers in the Corpus "One Day of Speech": Approaches to the Annotation // Computational Models in Language and Speech. Proceedings of Computational Models in Language and Speech Workshop (CMLS 2018) co-located with the 15th TEL International Conference on Computational and Cognitive Linguistics (TEL 2018). Vol-2303. Kazan, Russia, November 1, 2018 / Ed. by A. Elizarov, N. Loukachevitch. Kazan (Volga Region) Federal University, N. I. Lobachevsky, Institute of Mathematics and Mechanics, Kazan, Russia; Lomonosov Moscow State University, Research Computing Center, Moscow, Russia, pp. 128–143.
17. *Zakharov, V. P.* (2005), Corpus Linguistics: Teaching Aid [Korpusnaja lingvistika: Uchebno-metodicheskoe posobie]. St. Petersburg. 48 p.