О роли статистики предлогов в определении стилистической принадлежности русскоязычных текстов

Ольга А. Митрофанова, Анна Д. Москвина

***Аннотация*— Данная работа посвящена роли статистики слов из служебных частей речи в автоматическом определении стилистических характеристик текста. В роли лингвистического параметра в нашем исследовании выступает соотношение семантически противопоставленных частотных предлогов русского языка. Рассмотрены семь пар частотных предлогов, обладающих пространственным значением и имеющих одно или несколько переносных употреблений: *под / над, в / из, к / от, за / перед, в / на, на / с*. Гипотеза исследования состоит в том, что коэффициенты соотношения частот предлогов в рассматриваемых парах могут указывать на стилевую принадлежность тестов. Материалом для наших экспериментов послужили следующие корпусы текстов разных функциональных стилей и разной тематики: общий, художественный, публицистический, нехудожественный, устный подкорпусы Национального корпуса русского языка (НКРЯ), корпусы русского языка из семейства сверхбольших корпусов Araneа, а именно, корпусы Araneum Russicum Russicum и Araneum Russicum Externum, а также корпус текстов из социальных сетей, состоящий из постов и комментариев с платформ Facebook и Twitter, и корпус художественных текстов, включающий тексты произведений с сайта Либрусек. Была проверена гипотеза о стилистическом сходстве устной речи и письменной речи людей в социальных сетях на основании статистического анализа многозначных предлогов. Эксперименты подтвердили значимость коэффициента *под / над* в диагностике стиля и типа текстов, а также выявили информативность коэффициентов соотношения частот предлогов *в / из* и *за / перед* в дифференциальной диагностике письменных и устных тестов. Были получены как данные о статистике употребления предлогов, так и информация о семантическом наполнении предложных конструкций, которая столь же важна для определения стилевых, жанровых характеристик текстов и их тематической принадлежности. На основе имеющихся корпусных данных были выявлены и проанализированы основные особенности функционирования многозначных предлогов.**

***Ключевые слова*— квантитативная лингвистика, корпусная лингвистика; статистика, стиль, тематика текста.**

I. Проблемы определения стилей и типов текстов

В связи с наблюдаемым многообразием текстов в различных письменных источниках возникает вопрос о критериях определения их типов. Решение этого вопроса важно для ряда прикладных задач, среди которых автоматическое формирование корпусов текстов, автоматическая классификация веб-документов [13] и т.д.

Если процесс формирования жанров ранее происходил медленно и контролируемо, то сегодня мы сталкиваемся с ситуацией, когда благодаря интернету постоянно происходит процесс создания и распространения письменных текстов и стихийного формирования их новых типов, что очевидно требует пересмотра и расширения существующей классификации текстовых источников.

Традиционное понятие жанра было скорректировано применительно к изучению интернет-текстов как в западных, так и в русскоязычных исследованиях. Задачи определения стиля и тематики текстов входят в более широкую задачу определения жанра (или типа) текста.

С нашей точки зрения, понятие типа текста опирается на следующие основания:

1) существование различных ситуаций речевого общения,

2) существование разных каналов связи с разной структурной организацией (форумы, посты в социальных сетях, наличие или отсутствие веток дискуссий),

3) различные коммуникативные цели сообщений,

4) выбор различных языковых средств.

Существуют надежные формальные критерии стилистической диагностики текстов: морфосинтаксические (распределение слов по частям речи,

Статья получена 20 октября 2020. Исследование поддержано грантом РФФИ № 17-29-09159 «Квантитативная грамматика русских предложных конструкций». Данная статья основана на материалах доклада авторов в рамках конференции «Интернет и современное общество 2020».

О.А.Митрофанова – Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия (e-mail: o.mitrofanova@spbu.ru)

А.Д.Москвина – Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ) (e-mail: admoskvina@hse.ru)

статистические свойства конструкций, оценки синтаксической сложности и т.п.) [1, 11, 12, 15], графические (формулы, иноязычные обозначения, аббревиатуры и т.д.) [8]. Использование семантических критериев затруднено тем, что они с трудом поддаются формализации. Одним из допустимых семантических критериев является распределение значений многозначных предлогов [14].

II. Роль служебных частей речи при анализе текстов

Лексику любого языка можно разделить на два больших класса: содержательную (свободную) и служебную (функциональную, грамматикализованную). Содержательная лексика обладает большей информативностью, именно она формирует семантическое ядро, смысл любого текста. Объясним, почему служебные слова действительно представляют интерес в лингвистических исследованиях, в том числе при автоматическом анализе. В отличие от содержательных слов, они:

1) выполняют синтаксическую, структурообразующую функцию, поэтому их употребление независимо от тематики и, до некоторой степени, жанра текста;

2) в силу грамматикализованности, их употребление обязательно и зачастую не является результатом сознательного выбора автора;

3) служебные единицы обладают высокой частотностью, и как следствие, предоставляют обширный материал для наблюдений.

Несмотря на интуитивно ожидаемую независимость от тематики и жанра, распределения служебных (функциональных) единиц в разных типах текстов демонстрируют значимые закономерности [6]. Уже довольно традиционно (начиная с работы [7]) распределения функциональных слов используются как диагностический параметр в задачах атрибуции, например, [2]. Кроме того, хорошим примером таких закономерностей стала Дельта Бёрроуза [5] — простой и эффективный метод атрибуции текстов, основанный на сравнении распределений самых частотных слов для оценки близости текстов, причем такими словами ожидаемо оказываются служебные части речи — предлоги, артикли, частицы.

Предлоги, в свою очередь, представляют собой важнейший подкласс функциональных слов в русском языке и самую частотную служебную часть речи (по НКРЯ). Заслуживает внимания и отдельное изучение их взаимосвязи с различными типами текстов, в отрыве от других служебных частей речи. Так, например, была показана связь употребления конкретных предлогов с гендерной принадлежностью автора, что позволяет использовать такие предлоги в качестве параметров при классификации текстов по гендерному признаку [9] в задачах профилирования.

III. Роль статистики употребления предлогов в определении стилей и типов текстов

В своем исследовании Д. В. Сичинава подтверждает гипотезу о том, что распределение значений многозначных предлогов связано с типом текстов и их тематикой. В статье [14] рассматривается применение такого критерия, как соотношение частот предлогов под и над в качестве лингвистического параметра для наивно-«жанровой» классификации корпуса текстов (включает такие «жанры», как поэзия, художественная проза, детективы и др.), а также приводится семантический анализ данных предлогов. Автор обосновывает такую идею, во-первых, доказанной эффективностью специфики употребления служебных слов как признака авторского стиля, понимая тип текста как расширение данного понятия, и, во-вторых, тем, что пространственные значения выбранных предлогов обладают богатой семантикой и возможностью множественного метафорического употребления. Было показано, что по данному параметру (коэффициенту ***под/над***) тексты разбиваются на три группы: поэзия, художественная и нехудожественная проза.

IV. Эксперименты по применению статистики предлогов в стилистической диагностике

Опираясь результаты предыдущих исследований, в нашей работе мы анализируем соотношение частот употребления предложных значений и конструкций, а также некоторых других типов служебных слов в текстах разных функциональных стилей и тематики. Материалом исследования послужили подкорпусы НКРЯ (http://www.ruscorpora.ru/), данные Частотного словаря современного русского языка [10], корпусы кафедры математической лингвистики СПбГУ. Нашей задачей было выяснить, является ли диагностическим коэффициент соотношения пар семантически противоположных примитивных предлогов для определения стилистических (жанровых) характеристик текста. Продолжая работу [14], мы взяли семь пар частотных предлогов, обладающих пространственным значением и, очевидно, допускающим метафорическое употребление: ***под/над, в/из, к/от, за/перед, в/на, на/с***. Ниже приведены примеры употребления предлога ***под*** в пространственном (1) и объектном (2) значениях.

(1) *Когда Петя и Оля впервые встретились* ***под*** *столом в поисках упавшей ложки, между ними пробежала молния, которая однозначно указывает на внезапно вспыхнувшую любовь, полную искренности.* (Источник: НКРЯ. Форум: комментарии к фильму «Все будет хорошо» (2008-2011))

(2) *Во всех случаях носителем сообщения является сигнал — физический процесс, изменения в форме которого однозначно отображают сообщение,* ***под*** *которым понимаются определённым образом оформленные сведения, подлежащие обработке и доставке.* (Источник: НКРЯ. Интерпретации и смысл понятия «информация» // «Информационные технологии», 2004)

A. Эксперименты с данными НКРЯ

Первая серия экспериментов была проведена на подкорпусах НКРЯ (общий, художественный, публицистический, нехудожественный, устный). Были получены следующие данные: соотношения частот многозначных предлогов могут меняться в разных стилях (типах) текстов, при этом наиболее четко противопоставлены письменные тексты и устная речь (см. табл. 1).

Таблица 1. Соотношения частот по подкорпусам НКРЯ

	НКРЯ	**худож**	**публ**	**нехудож**	**устн**
под / над	1.99	1.83	1.99	2.01	2.54
в / из	7.04	6.29	7.04	7.17	9.17
к / от	1.55	1.57	1.55	1.49	1.82
за / перед	6.11	5.91	6.11	6.10	10.32
в / на	2.27	1.56	2.27	2.37	2.02
на / с	1.34	1.37	1.34	1.29	1.33

Так, наиболее информативными оказываются коэффициенты соотношения частот предлогов ***в/из*** и ***за/перед***, их значения в устных текстах достигают 9.17 и 10.32 соответственно, в то время как максимальные значения этих коэффициентов для письменных текстов равняются 7.17 и 6.10 (расчеты производились на подкорпусе нехудожественных текстов). С другой стороны, некоторые пары предлогов вели себя «равномерно» в разных подкорпусах, например, пара ***на/с***.

На следующих этапах работы мы решали вопросы о том,

1) насколько соотношение частот предлогов зависит от объема корпусов,

2) влияет ли тип источников (веб-страницы и их локализация) на предложные коэффициенты,

3) каков статус текстов социальных сетей с точки зрения статистики предлогов.

B. Эксперименты с данными корпусов Aranea

Вторая серия экспериментов была проведена на семействе сверхбольших корпусов русского языка Aranea [4] (http://unesco.uniba.sk/aranea_about/index.html; http://corpora.spbu.ru/aranea/). Результаты этих экспериментов подтвердили гипотезы об изменчивости коэффициентов пар ***под/над***, ***за/перед*** и о стабильности коэффициента ***на/с***. Таким образом, можно сделать вывод, что значения коэффициентов мало зависят от объема корпусов. При работе с корпусами Araneum Russicum Russicum и Araneum Russicum Externum были сделаны наблюдения о том, что возможно применение данных о предлогах, расширенных информацией о частотности предложных конструкций и их значений, для дифференциации региональных разновидностей русского языка по веб-данным.

C. Эксперименты с данными корпуса социальных медиа и художественного корпуса

В третьей серии экспериментов мы проверили гипотезу о характеристиках интернет-дискурса, развивая интуитивно очевидное предположение о том, что язык социальных сетей стилистически более близок к устной речи, чем к другим разновидностям письменных текстов.

В качестве источников данных мы взяли корпусы текстов с соревнования морфологических парсеров MorphoRuEval-2017, а именно корпус текстов из социальных сетей, состоящий из постов и комментариев с платформ Facebook и Twitter, а также корпус художественных текстов, представленный произведениями с сайта Либрусек (http://http.lib.rus.ec/).

Объем каждого из сформированных тестовых корпусов составлял 1 миллион словоупотреблений, тексты были выбраны из существующих коллекций случайным образом. Были получены данные по абсолютной частоте употребления всех рассматриваемых предлогов, а также данные по коэффициентам рассматриваемых пар. Эти данные мы сравнили с данными устного корпуса, описанного выше. В результате была выявлена наиболее значимая для наших целей пара: ***за/перед***. Коэффициент составил 12.34 для корпуса социальных сетей и 10.32 для устного, против 4.77 для художественных текстов (см. табл. 2).

Таким образом, первые данные, полученные в пристрелочных экспериментах, говорят в пользу гипотезы о стилистическом сходстве устной речи и письменной речи людей в социальных сетях на основании статистического анализа многозначных предлогов.

Таблица 2. Соотношения частот в корпусах Либрусек и социальных сетей

	Либрусек	**Twitter + Facebook**
под / над	1.29	2.92
в / из	6.42	8.23
к / от	1.51	0.60
за / перед	4.77	12.34
в / на	2.16	1.57
на / с	1.20	1.95

D. Наблюдения о семантике предложных конструкций

Важно исследовать не только статистику употребления предлогов, но и статистические данные о семантическом наполнении предложных конструкций [3], которые также могут стать диагностическим параметром для определения жанровых характеристик. Мы извлекли из тестового корпуса социальных сетей все предложные конструкции и, проанализировав их, выявили основные особенности использования многозначных предлогов.

Так, для предлога ***за*** основным значением является объект действия, морфологическим маркером которого является винительный падеж (***за*** *Федеральные*

Собрания, респект ***за*** *его ответы, пенсионеров* ***за*** *справедливость*). Для его парного предлога ***перед*** — значение темпоратива (*праймериз* ***перед*** *выборами, прям перед выходом, перед Днем Победы*). Следует отметить, что с точки зрения тематики этот корпус тяготеет к дискуссиям о политике, что, возможно, послужило одной из причин значительного преобладания предлога ***за***, характерного в конструкциях типа ***за*** *ЕР,* ***за*** *ЛДПР*, над предлогом ***перед***, темпоральное значение которого кажется тематически независимым.

Среди наиболее частотных значений предлога ***под*** можно назвать коррелятив (*косит* ***под*** *припадочного, рядится* ***под*** *героем*) и квалификатор действия (***под*** *угрозой увольнения*).

Для предлога ***над*** основным значением стало значение объект-делибератив (*победа* ***над*** *нацизмом, издеваются* ***над*** *учителями*).

V. Итоги

В результате проделанной работы были сделаны наблюдения об изменчивости коэффициентов соотношения частот некоторых предлогов, определена информативность различных пар предлогов с точки зрения стилистической диагностики, была доказана стабильность коэффициентов при увеличении объема корпусов.

В ходе экспериментов была подтверждена гипотеза о сопоставимости устных текстов и письменных текстов социальных сетей.

Мы планируем расширить рамки исследования, анализируя не только частотность, но и семантическое наполнение предложных конструкций, что также может стать важным инструментом для определения стилевой принадлежности текстов.

Литература

[1] Andreev V.S., Beliaeva L.N. Internal Dynamics of Text: Parts of Speech Distribution in Verse // PRLEAL-2019: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), CEUR Workshop Proceedings, Vol-2552, pp. 151-160.

[2] Argamon S., Levitan S. Measuring the usefulness of function words for authorship attribution // Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, 2005.

[3] Azarova I., Khokhlova M., Zakharov V., Petkevič V. Ontological description of Russian prepositions // PRLEAL-2019: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), CEUR Workshop Proceedings, Vol-2552, pp. 245-257.

[4] Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora // Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014, pp. 257-264.

[5] Burrows J. ‘Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship” // Literary and Linguistic Computing, Vol. 17, No. 3, 2002, pp. 267-287.

[6] Kestemont M. Function Words in Authorship Attribution. From Black Magic to Theory // Proceedings of the 3rd Workshop on Computational Linguistics for Literature, Gothenburg, 2014.

[7] Mosteller F., Wallace D. Inference in an Authorship Problem // Journal of the American Statistical Association, 58(302), 1963, pp. 275-309.

[8] Воронов С.О. Фильтрация и тематическое моделирование коллекции научных документов. Долгопрудный, 2014.

[9] Литвинова Т.А. Стилеметрическое исследование текстов участников экстремистского форума: гендерный аспект // Известия Воронежского государственного педагогического университета. Серия «Филологические науки», 2019, № 4 (285), с. 227-236.

[10] Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М., 2009.

[11] Мартыненко Г.Я. Основы стилеметрии. Л., 1988.

[12] Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л., 1990.

[13] Рубинер В.И. Классификация интернет-страниц: алгоритмы // Структурная и прикладная лингвистика. Вып. 10. СПб., 2014.

[14] Сичинава Д.В. Об одном лингвистическом параметре типологии текстов: коэффициент «под/над» // Научно-техническая информация, Серия 2, № 10, 2003, с. 27-35.

[15] Тулдава Ю. Проблемы и методы квантитативно-системного исследования лексики. Тарту, 1987.

On the Role of Prepositional Statistics for Genre Identification of Russian texts

Olga A. Mitrofanova, Anna D. Moskvina

Abstract— **In this work we investigate the role of statistical data on function words for automatic identification of genre and topical characteristics of Russian texts. We use the ratio of semantically related prepositions as the principal linguistic parameter. We consider seven frequent prepositions which have spatial meaning and also reveal one or more figurative meanings: *под (under) / над (over), в (in) / из (from), к (to) / от (from), за (behind) / перед (in front of), в (in) / на (at), на (at) / с (from)*. Our research hypothesis claims that coefficients of preposition frequency ratios in the above mentioned pairs may indicate stylistic properties of the texts. We based our research on several corpora representing different genres and topics: general, literary, publicistic, non-literary, oral subcorpora of the Russian National Corpus (RNC), Russian corpora from the Aranea superlarge corpora family, namely, Araneum Russicum Russicum and Araneum Russicum Externum corpora, as well as social media corpus including posts and comments from Facebook and Twitter networks, and literary corpus including texts from Librusec digital library. We verified the hypothesis on the stylistic homogeneity of oral and written speech of social media users, our verification was based on statistical analysis of polysemous prepositions. Experiments proved the significance of *под (under) / над (over)* coefficient in style and text type detection, and revealed the role of *в (in) / из (from)* and *за (behind) / перед (in front of)* in differentiation of written and oral texts. We obtained evidence on the statistics of preposition occurrence, as well as the information on the semantic content of prepositional phrases, which is of great significance for text style, genre and topic detection. We found out and analyzed the main properties of the use of polysemous prepositions.**

***Keywords*— quantitative linguistics, corpus linguistics, genre, topic, text statistics**

Olga Alexandrovna Mitrofanova
PhD, Associate Professor, Department of Mathematical Linguistics, Faculty of Philology, Saint-Petersburg State University, Russia (https://spbu.ru/)
e-mail: o.mitrofanova@spbu.ru
elibrary: authorid=4169-6068
scopus.com: authorId=36932407200
ORCID: orcidid=0000-0002-3008-5514

Anna Denisovna Moskvina
Lecturer, Saint-Petersburg School of Arts and Humanities, Department of Philology, Higher School of Economics National Research University, Russia (https://www.hse.ru/)
e-mail: admoskvina@hse.ru
elibrary: 6920-6233
ORCID: 0000-0001-7400-8097

References

[1] Andreev V.S., Beliaeva L.N. Internal Dynamics of Text: Parts of Speech Distribution in Verse // PRLEAL-2019: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), CEUR Workshop Proceedings, Vol-2552, pp. 151-160.

[2] Argamon S., Levitan S. Measuring the usefulness of function words for authorship attribution // Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, 2005.

[3] Azarova I., Khokhlova M., Zakharov V., Petkevič V. Ontological description of Russian prepositions // PRLEAL-2019: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), CEUR Workshop Proceedings, Vol-2552, pp. 245-257.

[4] Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora // Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014, pp. 257-264.

[5] Burrows J. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship // Literary and Linguistic Computing, Vol. 17, No. 3, 2002, pp. 267-287.

[6] Kestemont M. Function Words in Authorship Attribution. From Black Magic to Theory // Proceedings of the 3rd Workshop on Computational Linguistics for Literature, Gothenburg, 2014.

[7] Mosteller F., Wallace D. Inference in an Authorship Problem // Journal of the American Statistical Association, 58(302), 1963, pp. 275-309.

[8] Voronov S.O. Fil`traciya i tematicheskoe modelirovanie kollekcii nauchny`x dokumentov. Dolgoprudny`j, 2014.

[9] Litvinova T.A. Stilemetricheskoe issledovanie tekstov uchastnikov e`kstremistskogo foruma: genderny`j aspekt // Izvestiya Voronezhskogo gosudarstvennogo

pedagogicheskogo universiteta. Seriya «Filologicheskie nauki», 2019, № 4 (285), pp. 227-236.
[10] Lyashevskaya O.N., Sharov S.A. Chastotny`j slovar` sovremennogo russkogo yazy`ka (na materialax Nacional`nogo korpusa russkogo yazy`ka). M., 2009.
[11] Marty`nenko G.Ya. Osnovy` stilemetrii. L., 1988.
[12] Marusenko M.A. Atribuciya anonimny`x i psevdonimny`x literaturny`x proizvedenij metodami raspoznavaniya obrazov. L., 1990.
[13] Rubiner V.I. Klassifikaciya internet-stranicz: algoritmy` // Strukturnaya i prikladnaya lingvistika. Vy`p. 10. SPb., 2014.
[14] Sichinava D.V. Ob odnom lingvisticheskom parametre tipologii tekstov: koe`fficient «pod/nad» // Nauchno-texnicheskaya informaciya, Seriya 2, № 10, 2003, pp. 27-35.
[15] Tuldava Yu. Problemy` i metody` kvantitativno-sistemnogo issledovaniya leksiki. Tartu, 1987.