DOI: 10.21603/2078-8975-2019-21-3-801-810

оригинальная статья
УДК 81-139

Предлоги и падежные формы русского языка как предмет идентификационной лингвистики

Николай Д. Голев [a]; Галина В. Напреенко [b, @, ID]

[a] Кемеровский государственный университет, 650000, Россия, г. Кемерово, ул. Красная, 6

[b] Кемеровский государственный медицинский университет, 650029, Россия, г. Кемерово, ул. Ворошилова, 22а

[@] vila1991@mail.ru

[ID] https://orcid.org/0000-0002-4404-0560

Поступила в редакцию 23.05.2019. Принята к печати 05.08.2019.

Аннотация: Статья посвящена изучению служебных слов русского языка в идентификационном аспекте. Тема включена в контекст более общей проблемы наличия и степени проявления идентификационной функции у различных единиц языка на его различных уровнях, в том числе морфологическом, в частности проблемы ее различий в словах, принадлежащих разным частям речи, и словоформах, относящихся к разным грамматическим категориям. Непосредственным предметом анализа в данном исследовании являются служебные части речи – предлоги, а также падежные формы и приписываемые им субъектами речи грамматические значения. Источник материала – интернет-переписка двух авторов. Статья включена в парадигму исследований, направленных на выявление и описание квантитативных закономерностей распределения единиц, свойств и отношений в текстах, прежде всего – закономерностей, связанных с установлением коэффициента устойчивости / вариативности таковых единиц, свойств и отношений. В решении этого вопроса авторы исходят из положения о том, что разные единицы имеют разный коэффициент: одни стремятся к полюсу устойчивости, другие меняют коэффициент в зависимости от разных характеристик текста. Для решения поставленных задач авторами применяется один из статистических методов исследования текста – статистический критерий К. Пирсона. Примененный метод, связанный с определением частотности рассматриваемых лексем во фрагментах текстов разных авторских профилей, позволил обнаружить идентификационный потенциал текстов, принадлежащих разным авторам.

Ключевые слова: идентификация текста, идентификатор, служебные части речи русского языка, лексико-квантитативный метод, критерий К. Пирсона

Для цитирования: Голев Н. Д., Напреенко Г. В. Предлоги и падежные формы русского языка как предмет идентификационной лингвистики // Вестник Кемеровского государственного университета. 2019. Т. 21. № 3. С. 801–810. DOI: https://doi.org/10.21603/2078-8975-2019-21-3-801-810

Введение

Настоящей статьей мы намерены поставить вопрос о наличии у элементов языковой системы идентификационной функции, проявляющейся в отношении к языковому материалу (прежде всего – тексту). Данная функция в самом широком плане отражает способность элементов нести информацию о системе в целом. В частности, элементы и фрагменты текста несут в себе отпечаток всего текста. Автор текста вольно или невольно следует инерционной энергетике текста, поэтому каждый последующий фрагмент воспроизводит его предшествующие и предстоящие особенности. Отсюда вытекает возможность автороведческого анализа и квалификации авторской принадлежности текста, что в силу практической надобности нередко становится предметом лингво-экспертных исследований, например [1–6]. Говоря об идентификационной функции, мы имеем в виду более широкий контекст. Целое как совокупность текстов определенного автора – частный случай целого. Таковым может быть и текст безотносительно к автору: целое как совокупность текстов определенного жанра, типа речи, дискурса (гендерного, возрастного и др.), эпохи и т. п. Столь же разнообразны и элементы, рассматриваемые как носители идентификационной функции, – от буквы до речевого жанра. В настоящее время под изоморфизмом части и целого подразумевается системно-структурный изоморфизм, причем чаще всего квантитативная структура текста, членами системных отношений (парадигм) предстают элементы в их количественном измерении и квантитативных оппозициях членов парадигм. Все сказанное обосновывает употребляемое нами далее выражение *идентификационная лингвистика,* которым мы хотели бы обозначить очерченное направление лингвистического

DOI: 10.21603/2078-8975-2019-21-3-801-810

анализа. В предлагаемой статье квантитативному анализу такого направления подвергаются единицы морфологического плана – части речи (в нашем случае служебные слова) и их функционально-семантические группировки (в нашем случае предлоги), реализуемые во фрагментах одного и того же текста и двух разных текстах.

Идентификационная функция речевых произведений и особенности предлогов и союзов по отношению к ней

Предварительно выскажем два общих соображения, значимых для дальнейшего анализа. Считаем важным, говоря об идентификационной функции элементов, иметь в виду некоторые тенденции квантитативных показателей, выявляемых в тексте. С одной стороны, они могут отражать устойчивые и постоянные соотношения количеств, характерные для всех русских текстов. Такие характеристики по сути представляют собой системные квантитативные характеристики данных элементов или их разрядов. Подобными устойчивыми характеристиками являются текстовые частности отдельных букв и их разрядов, например: гласных и согласных букв, йотированных и нейотированных гласных, твердых и мягких согласных, твердого и мягкого знака – в любом русском тексте такие соотношения в принципе тождественны или приближаются к таковым. Можно предположить и наличие тенденций к вариативности частотных характеристик букв и их типов в зависимости от функциональных типов текстов, текстов, принадлежащих разным эпохам или текстов разных поэтов. Можно предположить устойчивость соотношения слов, начинающихся на одну букву (об этом говорят толковые словари), одноморфемных и неодноморфемных слов и т. д.

Морфологические элементы русского языка в данном аспекте – не исключение, о чем говорят уже имеющиеся работы, прежде всего в области частей речи [7–13]. Гораздо меньше исследованы категории и парадигмы в их пределах. Тенденцию к устойчивости здесь поддерживает то обстоятельство, что морфологические элементы в языках сильного синтетического типа (к которым относится и русский язык) детерминируются в первую очередь синтактикой, элементы прагматики и семантики здесь также несомненно проявляются (в категориях залога, наклонения, лица, возможно – времени, категории полноты-краткости; мы полагаем, например, что выбор причастий и деепричастий более свободный, здесь прагмастилистический фактор силен, в отличие от инфинитива, которым управляет синтаксис). Но в целом роль семантики и прагматики в выборе морфологических форм не столь значительна. Данное обстоятельство приводит к тому, что, подчиняясь законам синтагматики, говорящий не имеет большой свободы выбора и действует автоматически в рамках предписываемых синтагматикой алгоритмов. Такова грамматическая категория падежа и увязанная с ней система предложно-падежных форм. Мы выдвигаем гипотезу, согласно которой морфологические категории могут выступать в качестве языкового материала, обладающего идентификационным потенциалом. Нужно полагать, что выбор морфологических категорий как обязательных единиц текста осуществляется неосознанно. В данном исследовании идентификационно-статистическому анализу подвергается грамматическая категория падежа.

В последней мы находим своеобразный компромисс между синтагматикой и семантикой (номинативностью). В этом плане данная статья является продолжением ранее опубликованной нами (в соавторстве) статьи, посвященной идентификационному потенциалу категории падежа [14], в которой данная функция включена в парадигму противопоставления константных и неконстантных членов морфологических категорий, результаты показали, что статистический расчет падежных форм может служить материалом для идентификации и характеристики идиостиля автора.

В настоящей статье акцент сделан на идентификационных возможностях предложного компонента предложно-падежной системы русского языка. В его трактовке мы исходим из тезиса, согласно которому предлог берет на себя фиксацию номинативного компонента обозначения ситуации и одновременно функции управления определенной падежной формой управляемого существительного.

Следующие фрагменты призваны проиллюстрировать в конкретно-исследовательском плане высказанные теоретические положения.

Предлоги в аспекте оппозиции постоянных и вариативных квантитативных характеристик

Содержание в тексте определенного набора признаков, отражающих авторскую индивидуальность, позволяет говорить о наличии идентифицирующей функции у тех или иных языковых параметров. Мы полагаем, что предлоги как служебная часть речи, создавая особое строение предложений, выражая отношения между членами предложения, могут выражать неосознанное предпочтение автора, имеющее в квантитативном соотношении идентификационный потенциал данной части речи: могут выполнять идентифицирующую функцию текста и устанавливать принадлежность текста одному или разным авторам; идентифицировать фрагменты текстов (по принадлежности фрагментов текстов одному автору) или дифференцировать фрагменты текстов по данному признаку.

DOI: 10.21603/2078-8975-2019-21-3-801-810

Данные частотных словарей и количественные данные Национального корпуса русского языка (НКРЯ)[1] дают возможность ответа на вопрос о том, являются ли предлоги жестко распределёнными по текстам русского языка или их распределение вариативно. В указанном аспекте были выявлены данные по предлогам НКРЯ и Корпуса русского литературного языка (КРЛЯ)[2]. Данные занесены в таблицу 1, в которой представлены основной корпус и подкорпус устной речи НКРЯ и основной корпус и корпус «Драма» в КРЛЯ. Для сопоставления данных о частотности было использовано в качестве единицы измерения количество словоупотреблений на миллион – *imp* (*instances per million words*).

Таблица 1. Количественные данные по предлогам из НКРЯ и КРЛЯ, imp

Table 1. Quantitative data on the prepositions of the Russian National Corpus and Corpus of Standard Written Russian, imp

Предлог	НКРЯ		КРЛЯ	
	основной корпус	устный подкорпус	основной корпус	подкорпус «Драма»
на	15173	11169	15472	11407
возле	104	24	160	46
в (во)	31805	22705	27990	17638
у	4110	8592	5260	8735
перед	650	314	608	433
за	4116	3901	3919	4003
между	718	353	451	112
по	5980	4774	4425	2708
под (подо)	1264	539	1166	632
с (со)	12686	8907	11284	9076
над (надо)	603	190	641	423
около	275	133	285	51
от	3845	2044	3325	2162
из-за	246	242	336	388
из-под	89	24	95	41
к (ко)	6174	3920	5446	3983
о (об, обо)	4257	3426	3864	3600
через	781	520	727	500

Отчетливо видно, что предлоги являются высокоупотребимыми в речи, при этом данные по корпусам имеют небольшие различия. Эти различия не являются существенными, на их основании нельзя утверждать, что их употребление, фиксируемое в квантитативном индексе, является зависимым от подкорпуса (функциональной разновидности языкового материала), гораздо больше оснований утверждать об их тенденции к постоянству квантитативного индекса. Прежде всего, это касается наиболее частотных предлогов – *в, на, с, у, к, о*. Эту мысль доказывает сопоставление представленных данных с Частотным словарем русского языка (на материалах Национального корпуса русского языка) О. Н. Ляшевской, С. А. Шарова[3]. Частотный словарь отражает частоту словоупотребления в русском языке. Таблица в Частотном словаре представляет собой указание ранга (порядкового номера в частотном словаре), слова и частотность употребления слова. Порядок слова обратно пропорционален частоте словоупотребления: чем чаще употребляется слово, тем ниже его порядковый номер. Предлог *в* имеет ранг 2, *на* – 4, *с* – 8, *к* – 15, *у* – 21, *о* – 31.

Согласно Частотному словарю предлоги *в, на, с* входят в десять самых частотных слов русского языка. Сопоставление указанных данных с данными НКРЯ показывает, что эти предлоги являются частотными в разных корпусах текстов, их частотность характерна для языка в целом, а значит и в конкретном тексте их частотность гипотетически должна быть высокой. В связи с этим возникает проблема возможности выявления идентификационного потенциала у данных слов.

Проблема наличия или отсутствия грамматической (синтагматической) или семантической (номинативной) природы выбора пишущим предлога разрешается путем проведения эксперимента. В эксперименте приняли участие тринадцать испытуемых: пол – женский (одиннадцать человек), мужской (два человека); возраст – от 19 до 22 лет; социальный статус – студенты 3, 4 и 5 курсов, обучающиеся на гуманитарных программах.

Испытуемым было предложено описать четырнадцать изображений, взятых из учебника по немецкому языку[4] (рис.). В учебнике изображения приводились в качестве иллюстраций к упражнению на предлоги и слова, необходимые для пространственного ориентирования. Изображения были следующими: мультипликационный персонаж нарисован в том или ином пространственном отношении к достопримечательностям Берлина (под Бранденбургскими воротами или у Красной ратуши и т. д.). Инструкция к эксперименту: Опишите каждую из предложенных картинок одним предложением, отвечая на вопрос «Где находится медведь?» и сделайте запись под соответствующими номерами.

[1] Национальный корпус русского языка. Режим доступа: http://www.ruscorpora.ru/ (дата обращения: 20.10.2018).

[2] Корпус русского литературного языка. Режим доступа: http://narusco.ru/ (дата обращения: 25.10.2018).

[3] Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. 1090 с. Режим доступа: http://dict.ruslang.ru/freq.php (дата обращения: 20.10.2018).

[4] Themen Aktuell 1. 2010. С. 99–100.

DOI: 10.21603/2078-8975-2019-21-3-801-810

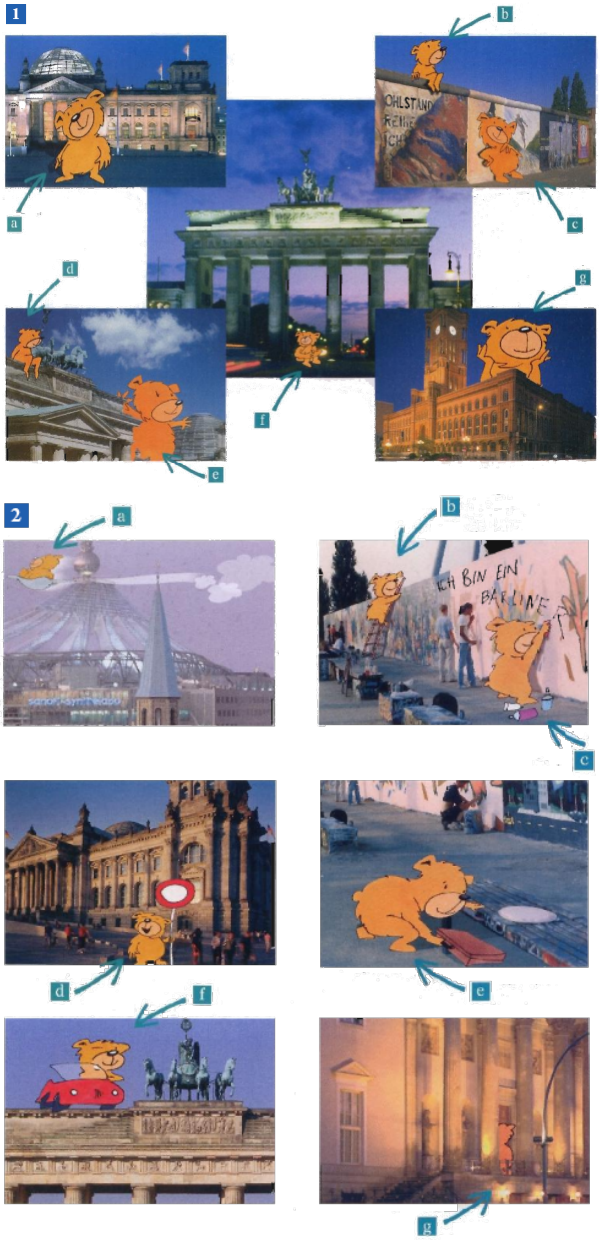

Рис. Иллюстрации к упражнению на предлоги и слова, необходимые для пространственного ориентирования
Fig. Illustrations for the exercise on prepositions and words necessary for spatial orientation

Результаты эксперимента показывают разнообразие языковых единиц, используемых информантами при описании иллюстраций. Например, иллюстрация, на которой изображен анимационный медведь между колоннами Французского собора, описывалась следующими способами: *прячется в проходе; стоит за колонной; между колонн, на крыльце здания; медведь возле библиотеки; мишка заходит в театр*; беспредложное описание: *мишка идет домой*. Иллюстрация, на которой изображен медведь под Бранденбургскими воротами, описана информантами следующим образом: *под аркой, возле арки, сидит на дороге, сидит у арки во Франции* и т. д. Все полученные варианты предлогов сгруппированы в таблицу 2, которая отражает распределение частоты употребления предлогов для описания каждой иллюстрации. В таблице иллюстрации обозначены буквами *A1, B1, C1, D1, E1, F1, G1, A2, B2, C2, D2, E2, F2, G2*.

Результаты, отражающие общее количество частоты предлога, используемого для описания иллюстрации, при сопоставлении с данными НКРЯ (таблица 1) и данными Частотного словаря русского языка показывают, с одной стороны, наличие постоянной высокой частоты употребления предлогов (это касается непосредственно предлогов, входящих в 10 самых частотных слов русского языка – *в, на*), с другой стороны, показывают вариативность употребления предлогов, гипотетически несущих идентифицирующую функцию. Идентификационный потенциал может выражать, например, предлог *возле*, который в Частотном словаре имеет ранг 1088 (т. е. частота его употребления в языке невысока), однако по результатам эксперимента является вторым по частоте выбора информантами для описания картинки.

Таким образом, вариативность (в том числе лексическая), характерная для языка в целом, проецируется непосредственно на вариативность при индивидуальном выборе предлога в каждой отдельной ситуации речевого общения. Данный выбор осуществляется из ряда синонимичных единиц, представленных разнообразием языка.

Апробация гипотезы прошла на материале текстов интернет-переписки двух авторов. Восемь писем Автора 1 (пол: женский; социальный статус: студент; возраст (на момент написания): 18 лет) были рассмотрены в качестве единого текста и составили в общей сложности 9261 слово (Текст 1). Двенадцать писем Автора 2 (пол: женский; социальный статус: студент; возраст (на момент написания): 18 лет) также были проанализированы как одно речевое произведение, которое составило 8989 слов (Текст 2). Подсчет единиц был осуществлен с помощью компьютерной программы *Microsoft Word Starter* 14.0.5128.5000 (32-разрядная)[5]. Переписка осуществлялась посредством электронной почты в течение трех месяцев. Письма, отобранные для анализа, являются близкими по тематике, в них обсуждаются одни и те же события, предметы, лица, присутствуют вопросы одного автора и ответы на них другого.

Были исследованы некоторые простые первообразные предлоги: *в (во), до, без, вокруг, за, из, к (ко), между, на, о (об), около, от, по, под, при, про, с (со), у*. В таблице 3 представлены данные по выделенным ваше предлогам *в, на, с, у, к, о*.

[5] Расчеты произведены студентом О. Раевой.

DOI: 10.21603/2078-8975-2019-21-3-801-810

Таблица 2. Результаты эксперимента, отражающие предпочтения испытуемых в выборе предлога при описании иллюстрации
Table 2. Results of the test in preposition preferences during picture description

Предлог	Иллюстрации														Всего
	A1	B1	C1	D1	E1	F1	G1	A2	B2	C2	D2	E2	F2	G2	
на	7	12	4	12	2	2	5	9	9	4	3	6	9	1	85
возле	3	–	3	2	1	1	–	–	–	5	4	5	1	1	26
в	1	1	–	1	2	2	2	3	–	–	1	–	6	2	21
у	1	–	2	–	1	1	–	–	1	4	2	–	–	–	12
перед	3	–	2	–	1	1	–	–	–	–	3	–	–	–	10
за	–	–	–	–	–	–	4	–	1	–	–	–	–	4	9
между	–	–	–	–	4	1	–	–	–	–	–	–	–	4	9
по	–	–	1	–	–	2	–	1	1	–	–	–	3	–	8
под	–	–	–	–	–	5	–	–	–	–	–	3	–	–	8
с	–	–	–	1	–	–	–	1	–	–	2	–	–	–	4
над	–	–	–	–	–	–	–	3	–	–	–	–	–	–	3
около	1	–	1	–	–	–	–	–	–	–	1	–	–	–	3
от	1	–	–	–	–	–	–	–	1	1	–	–	–	–	3
из-за	–	–	–	–	–	–	1	–	–	–	–	–	–	–	1
из-под	–	–	–	–	–	–	–	–	–	–	–	1	–	–	1
к	–	–	–	–	–	–	–	–	1	–	–	–	–	–	1
о	–	–	–	–	–	–	1	–	–	–	–	–	–	–	1
через	–	–	–	–	–	–	–	–	1	–	–	–	–	–	1
Всего	17	13	13	16	11	15	13	17	15	14	16	15	19	12	**206**

Важным критерием является соотношение количества употребления предлогов на количество словоупотребления в тексте. В таблице 3 предлоги расположены согласно их рангу в Частотном словаре русского языка.

Описательный анализ показывает, что различие в количественном употреблении предлогов в текстах 1 и 2 не является существенным. Особенно это относится к частоте употребления самых частотных предлогов в русском языке (2 и 4 ранг): *в* и *на*. Гипотезу относительно отсутствия идентификационного потенциала у предлогов с наиболее высокой частотой употребления, характерной для языка, описательный анализ подтверждает. При этом предлоги, имеющие наименьшую частоту употребления в языке, могут обладать, по нашему мнению, наибольшим потенциалом (таблица 4).

Статистические способы обработки информации позволяют получить результаты с различной степенью достоверности [15]. В рамках данного исследования апробируется статистический критерий К. Пирсона (или χ^2). Отметим, что данный критерий не раз рассматривался в качестве одного из возможных методов идентификации текста [16–19]. Нулевая гипотеза данного исследования такова: между генеральными параметрами (т. е. общими признаками) сравниваемых групп (Текст 1 и Текст 2) разница равна нулю, а наблюдаемые различия носят не систематический, а исключительно случайный характер.

Таблица 3. Ранг и частотность употребления некоторых предлогов в исследуемых текстах
Table 3. Rank and frequency of the use of some prepositions in the texts under study

Ранг	Предлог	Частотность употребления	
		Текст 1	Текст 2
2 (76)	в (во)	233	198
4	на	145	111
8 (77)	с (со)	91	143
15 (444)	к (ко)	57	56
21	у	41	84

Таблица 4. Количественные данные по предлогам, имеющим наименьшую частоту употребления в языке
Table 4. Quantitative data on the prepositions with the lowest frequency of use in the language

Ранг	Предлог	Частотность употребления	
		Текст 1	Текст 2
51	до	31	19
63	при	10	2
755	вокруг	0	2

DOI: 10.21603/2078-8975-2019-21-3-801-810

В исследованиях обычно принимается 5 %-й уровень значимости [20], которому соответствует 95 %-й уровень надежности. Мы также используем данную величину, и это значит, что если фактически полученная величина χ^2 ($\chi^2_ф$) окажется меньше его критического значения (χ^2_{st}), которое представлено в специальной таблице, то нулевая гипотеза будет подтверждена. И наоборот, если значение $\chi^2_ф$ окажется больше значения χ^2_{st}, тогда величина коэффициента Тамбовцева (ТМВ) ($\chi^2_ф/\chi^2_{st}$)>1, то нулевая гипотеза будет опровергнута.

Итак, необходимо вычислить сумму значений χ^2 для каждого из анализируемых предлогов. Предлог *без* в Тексте 1 встретился 4 раза на 8989 слов, а в Тексте 2 – 9 раз на такое же количество слов. Необходимо произвести следующие вычислительные шаги для определения значения критерия К. Пирсона: 9+4=13; 9–4=5; 5^2=25; χ^2=25/13=1,92. Проделаем те же шаги для остальных исследуемых предлогов. Результаты представлены в таблице 5.

Таблица 5. Результаты вычисления значений χ^2 относительно употребления исследуемых предлогов

Table 5. Values of χ^2 relative to the use of the prepositions under study

Предлог	Частота встречаемости		χ^2
	Текст 1	Текст 2	
в (во)	233	198	2,84
до	31	19	2,88
без	4	9	1,92
вокруг	0	2	2
за	27	24	0,18
из	26	27	0,02
к (ко)	57	56	0,01
между	3	1	1
на	145	111	4,51
о (об)	41	50	0,89
около	3	3	0
от	23	25	0,08
по	44	42	0,05
под	6	2	2
при	10	2	5,33
про	21	18	0,23
с (со)	91	143	11,55
у	41	84	14,79
$\Sigma\chi^2$	–	–	50,28

Необходимо определить соотношение вычисленного и критического значения χ^2 (т. е. величину ТМВ). Для этого необходимо вычислить число степеней свободы по выше формуле $k=(c-1)(r-1)$, где c – число строк в таблице, а r – число граф. Из этого следует, что k=(18–1)(2–1)=17*1=17. Нами также выбран 5 %-й уровень значимости. Обращаясь к таблице критических значений χ^2-критерия Пирсона [20, с. 271], выясняем, что критическое значение χ^2, соответствующее уровню значимости 5 % и числу степеней свободы 17, равно 27,59.

Вычислим коэффициент ТМВ по формуле $\chi^2_ф/\chi^2_{st}$=50,28/27,59=1,82. Значение ТМВ>1 (или $\chi^2_ф>\chi^2_{st}$), а это значит, что предположение о том, что между общими признаками сравниваемых текстов разница равна нулю, а наблюдаемые различия носят не систематический, а исключительно случайный характер, должно быть отвергнуто в соответствии со статистическим законом.

Эмпирические данные подтвердили тот факт, что Текст 1 и Текст 2 написаны разными авторами. Такой результат говорит об эффективности применения χ^2-критерия К. Пирсона как статистического метода.

Номинативный аспект семантики падежных форм

Категория падежа – это словоизменительная категория имени, которая реализуется в системе ряда форм, противопоставленных друг другу. Значение данной категории заключено в выражении отношения имени к другому слову в составе словосочетания или предложения. Падеж способен реализовывать свои как присловные, так и неприсловные позиции и связи. Существует комплекс основных, самых общих значений, присущих падежу и совпадающих и в присловных, и в неприсловных его позициях [21]: объектное (отношение предмета к действию, которое направлено на этот предмет), субъектное (отношение предмета к действию, которое совершается самим этим предметом) и определительное (отношение предмета к другому предмету, действию, состоянию либо к целой ситуации, которая этим отношением характеризуется). Семантический компонент падежных форм не обладает признаками обязательности, регулярности и стандартности, он сильно вариативен, и тенденция к константности проявляется в них не всегда достаточно четко.

Так, для именительного падежа характерны значения субъектное и определительное, для родительного падежа – субъектное, определительное (исключая обстоятельственное) и предметное, дательному падежу присущи субъектное и объектное значения, винительному падежу – объектное, творительный падеж выражает определительное и объектное значения, а значения предложного падежа во многом совпадают со значениями беспредложных падежных форм.

Очевидно, что выполняющие различные синтаксические функции падежные формы характеризуются целым спектром значений. Эти значения обязательны для выражения в русском языке, т. к. без них невозможно будет составить ни словосочетание, ни простейшее предложение. Но сам процесс выражения этих значений настолько бессознателен, что он «ускользает от подделок» [22].

DOI: 10.21603/2078-8975-2019-21-3-801-810

В связи с этим предположением мы решили обратиться к падежным формам анализируемых в данной работе текстов и проверить, насколько однородными являются Текст 1 и Текст 2 с этой точки зрения. Однако, на наш взгляд, необходимо обратиться к частотным и высокочастотным единицам рассматриваемых речевых произведений. Дело в том, что в силу ограниченного количества имеющихся в русском языке падежей, каждый из них может встретиться в том или ином языковом материале. Для нас же будет показательным то, какие падежные формы и грамматические значения приписывают субъекты речи частотным, а вместе с этим и значимым для них единицам текста.

В ряд частотных словоформ были включены имена существительные и личные местоимения (ед. ч. и мн. ч.), встречающиеся в тексте более десяти раз. Выявление частотности было произведено с помощью компьютерной программы *Simple Word Sorter Vertion 1.0*[6]. Далее было определено количество тех или иных падежных форм с учетом омонимичности (например, *его* – Р. п. или В. п. в соответствие с контекстом и т. д.).

Сопоставляя полученные данные, мы увидели, что количество словоформ в форме И. п. приблизительно одинаково для Текста 1 и Текста 2. Это объяснимо с точки зрения общих закономерностей построения речевого высказывания. Форма И. п. в роли первого актанта, т. е. подлежащего, в предложении является регулярной и продуктивной моделью построения коммуникативной единицей. Так принято обозначать субъект действия, что является непосредственной репрезентацией действительности, в которой тот или иной объект вещественного мира наделен в сознании человека определенного рода активностью. Это и отражается структурной стороной языка.

В случаях с остальными падежными формами при поверхностном анализе нельзя однозначно сказать о достоверном различии или очевидном сходстве Текста 1 и Текста 2. Необходимо обратиться к более надежным способам обработки информации – например, к χ^2-критерию К. Пирсона.

Нулевая гипотеза в данном случае будет совпадать с предыдущей: разница между признаками сравниваемых групп равна нулю и различия, наблюдаемые между ними, носят исключительно случайный характер. Только под признаками в данном случае подразумевается распределение по падежным формам.

Перед непосредственным проведением вычислительных операций необходимо отметить, что в крайних классах недопустимо количество вариантов менее 5. В таком случае их необходимо объединить с частотами соседних классов. Число степеней свободы (k) определяется уже по вторичному числу классов [20]. Данное замечание является актуальным для наших вычислений, в связи с чем данные были модифицированы (таблица 6). После произведенных модификаций данных таблицы были вычислены значения χ^2 для каждой группы и определена их сумма.

Таблица 6. Результаты вычисления значений χ^2 относительно употребления падежных форм
Table 6. Values of χ^2 relative to the use of case forms

Падеж	Частота словоупотреблений		χ^2
	Текст 1	Текст 2	
И. п.	521	524	0,01
Р. п.	63	82	2,49
Д. п.	76	120	9,88
В. п.+Т. п.+П. п.	76+0+1=77	90+0+0=90	1,01
$\Sigma\chi^2$	–	–	13,39

Следующим шагом стало сопоставление полученного значения χ^2 (χ^2=13,39) и его критического значения. Уровень значимости был принят такой же, как и в предыдущем вычислении (a) – 5 %. Для определения табличного значения χ^2 необходимым осталось вычислить число степеней свободы: $k=(c-1)(r-1)=(4-1)(2-1)=3$. При таких значения a и k критическое значение χ^2-критерия (χ^2_{st}) соответствует числу 7,81. Вычисленный коэффициент ТМВ=$\chi^2_ф$/χ^2_{st}=13,39/7,81=1,71, это позволяет нам говорить о том, что различия, наблюдаемые между Текстом 1 и Текстом 2, носят систематический характер и не являются случайными. Нулевая гипотеза опровергается. Так, употребление той или иной падежной формы является индивидуальным для разных субъектов речи и может обладать идентификационным потенциалом.

Заключение

Несмотря на высокочастотность употребления ряда предлогов в русском языке, некоторые из них могут проявлять идентифицирующую функцию, характеризуя тем самым авторские предпочтения (выраженные неосознанно). Нами была проведена проверка эффективности одного из статистических методов χ^2-критерия К. Пирсона на материале распределения предлогов в тексте. Проведенный методический эксперимент позволил убедиться в том, что предлоги действительно являются единицами, которые бессознательно используются человеком для выражения отношений, существующих в действительности. Разнообразие синонимичных рядов в языке предоставляет выбор, индивидуальный для каждого. Применив χ^2-критерий и опровергнув

[6] Расчеты произведены студентом О. Раевой.

DOI: 10.21603/2078-8975-2019-21-3-801-810

гипотезу о случайности наблюдаемых в тексте различий, мы пришли к выводу об эффективности данного статистического метода. Использование критерия К. Пирсона относительно падежных форм, реализованных в тексте, выявляет вариативность способов их выражения, продиктованную индивидуальным характером субъекта речи. Можно говорить об универсальности критерия К. Пирсона и возможности его эффективного применения. Полагаем, что характер распределения в текстах временных форм будущего простого и будущего сложного (*я спою песню* и *я буду петь песню* и т. д.), употребление сочинительных и подчинительных союзов и особенности структуры предложений также могут быть проанализированы с помощью критерия К. Пирсона. Данное предположение необходимо апробировать в последующих исследованиях.

Литература

1. Распопова Т. А. Судебная автороведческая экспертиза: опыт исследования // Ежегодник НИИ фундаментальных и прикладных исследований. 2015. № 1. С. 133–144.
2. Абрамкина Е. Е. Автороведческая экспертиза протокола допроса: основные особенности и методика анализа // Вестник Томского государственного университета. 2017. № 415. С. 158–163. DOI: 10.17223/15617793/415/22
3. Морозов А. В. Автороведческая экспертиза текста договора // Юрислингвистика. 2004. № 5. С. 290–297.
4. Суркова А. С. Проблема идентификации автора текста в юрислингвистике // Будущее технической науки: мат-лы 3-й молодежной науч.-практ. конф. (Н. Новгород, 26–27 мая 2004 г.) Н. Новгород: НГТУ, 2004. С. 51–52.
5. Резанова З. И., Романов А. С., Мещеряков Р. В. О выборе признаков текста, релевантных в авторoведческой экспертной деятельности // Вестник Томского государственного университета. Филология. 2013. № 6. С. 38–52.
6. Тамбовцев Ю. А., Тамбовцева Л. А., Тамбовцева Ю. Ю. Тексты Бахтина, Волошина и Медведева: авторство или плагиат? // Проблемы и перспективы языкового образования в XXI веке: мат-лы Междунар. науч.-практ. конф. (Новокузнецк, 08 апреля 2011 г.) / отв. ред. А. В. Колмогорова. Новокузнецк: OBERON, 2011. С. 402–421.
7. Морозов Н. А. Лингвистические спектры // Изв. императорской АН, отд. яз. и словесности XX. 1915. Кн. 4. С. 93–134.
8. Чащин С. В. Применение методов машинного обучения «с учителем» для атрибуции текста: отдельные подходы и промежуточные результаты при идентификации авторов русскоязычных текстов // Вопросы криминологии, криминалистики и судебной экспертизы. 2018. № 1. С. 139–147
9. Тихомирова Е. А. Статистический анализ в задаче идентификации автора текста, написанного на естественном языке // Наука и образование: научное издание МГТУ им. Н. Э. Баумана. 2017. № 6. С. 131–146. DOI: 10.7463/0617.0001245
10. Муха А. В., Розалиев В. Л., Орлова Ю. А., Заболеева-Зотова А. В. Автоматизированный подход к определению авторства текста // Известия Волгоградского государственного технического университета. 2013. № 14. С. 51–54.
11. Марусенко М. А. Атрибуция анонимных и псевдонимных текстов методами прикладной лингвистики // Прикладное языкознание. СПб.: СПбГУ, 1996. С. 469–473.
12. Головина Т. А. Квантитативный анализ лингвоперсонологического функционирования частей речи (на материале художественных произведений: тип текста – описание) // Вестник Томского государственного университета. 2006. № 120. С. 94–105.
13. Головина Т. А. Части речи в лингвоперсонологическом пространстве // Университетская филология образованию: человек в мире коммуникаций: мат-лы Междунар. науч.-практ. конф. «Коммуникативистика в современном мире: человек в мире коммуникаций». Барнаул: Изд-во Алт. ун-та, 2005. С. 86–87.
14. Голев Н. Д., Горюнова М. Е. Текст как объект квантитативно-морфологического исследования // Культура и текст. 2018. № 2. С. 216–229.
15. Захаров М. П. Автоматизация авторoведческих исследований // Судебная экспертиза. 2007. № 4. С. 63–69.
16. Тамбовцев Ю. А., Тамбовцева А. Ю., Тамбовцева Л. А. Типология распределения некоторых лингвистических единиц в тексте как показатель авторства текста // Вестник Омского университета. 2008. № 2. С. 88–96.
17. Степаненко А. А. Гендерная атрибуция текстов компьютерной коммуникации: статистический анализ использования местоимений // Вестник Томского государственного университета. 2017. № 415. С. 17–25. DOI: 10.17223/15617793/415/3
18. Батура Т. В. Математическая лингвистика и автоматическая обработка текстов на естественном языке. Новосибирск: РИЦ НГУ, 2016. 166 с.
19. Зенков А. В. Новый методстилеметрии на основе статистики числительных // Компьютерные исследования и моделирование. 2017. Т. 9. № 5. С. 837–850. DOI: 10.20537/2076-7633-2017-9-5-837-850
20. Лакин Г. Ф. Биометрия. М.: Высш. шк., 1980. 293 с.

DOI: 10.21603/2078-8975-2019-21-3-801-810

21. Русская грамматика. Т. 1. Фонетика. Фонология. Ударение. Интонация. Словообразование. Морфология / гл. ред. Н. Ю. Шведова. М.: Наука, 1980. 788 с.
22. Тамбовцев Ю. А. Фонотипологическая близость лингвистических объектов по критерию «хи-квадрат» // Вестник НГУ. Серия: Информационные технологии. 2010. Т. 8. Вып. 3. С. 46–54.

Prepositions and Case Forms of the Russian Language as a Subject of Identification Linguistics

Nikolay D. Golev [a]; Galina V. Napreenko [b, @, ID]

[a] Kemerovo State University, 66 Krasnaya St., Kemerovo, Russia, 650000

[b] Kemerovo State Medical University, 22a, Voroshilov St., Kemerovo, Russia, 650029

[@] vila1991@mail.ru

[ID] https://orcid.org/0000-0002-4404-0560

Received 23.05.2019. Accepted 05.08.2019.

Abstract: The article features Russian vocabulary from the aspect of identification linguistics, i.e. identification function on the morphological level, e.g. in various parts of speech and word forms belonging to different grammatical categories. The analysis focuses on auxiliary parts of speech, namely prepositions, related case forms, and grammatical meanings. The research is based on Internet correspondence. The article is included in the paradigm of research aimed at identifying and describing quantitative patterns in the distribution of units, properties, and relationships in texts and patterns of the stability / variability coefficient of units, properties, and relationships. The authors assume that different units have a different coefficient: some tend to be stable while others change their coefficient depending on different characteristics of the text. The research employed the method of Pearson's statistical criterion. The applied method determines the frequency of lexemes in texts belonging to different author profiles and reveals their identification potential.

Keywords: text identification, identifier, auxiliary parts of speech of the Russian language, lexical-quantitative method, Pearson's criterion

For citation: Golev N. D., Napreenko G. V. Prepositions and Case Forms of the Russian Language as a Subject of Identification Linguistics. *Vestnik Kemerovskogo gosudarstvennogo universiteta*, 2019, 21(3): 801–810. (In Russ.) DOI: https://doi.org/10.21603/2078-8975-2019-21-3-801-810

References

1. Raspopova T. A. Judicial authors examination: experience of research. *Ezhegodnik NII fundamental'nykh i prikladnykh issledovanii*, 2015, (1): 133–144. (In Russ.)
2. Abramkina E. E. The interview memo authorship examination: basic special aspects and the method of the analysis. *Vestnik Tomskogo gosudarstvennogo universiteta*, 2017, (415): 158–163. (In Russ.) DOI: 10.17223/15617793/415/22
3. Morozov A. V. Authorship examination of contract texts. *Legal Linguistics,* 2004, (5): 290–297. (In Russ.)
4. Surkova A. S. Issues of authorship identification in legal science. *The Future of Technical Science*: Proc. 3rd youth Sci.-Prac. Conf., Nizhny Novgorod, May 26–27, 2004. Nizhny Novgorod: NGTU, 2004, 51–52. (In Russ.)
5. Rezanova Z. I., Romanov A. S., Meshcheryakov R. V. Selecting text features relevant for authorship attribution. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya,* 2013, (6): 38–52. (In Russ.)
6. Tambovtsev Iu. A., Tambovtseva L. A., Tambovtseva Iu. Iu. Texts by Bakhtin, Voloshina, and Medvedev: authorship or plagiarism? *Problems and Prospects of Language Education in the XXI Century:* Proc. Intern. Sci.-Prac. Conf., Novokuznetsk, April 08, 2011, ed. Kolmogorova A. V. Novokuznetsk: OBERON, 2011, 402–421. (In Russ.)
7. Morozov N. A. Linguistic Spectra. *Izvestiia imperatorskoi AN, otd. iaz. i slovesnosti XX*, 1915, book 4, 93–134. (In Russ.)

DOI: 10.21603/2078-8975-2019-21-3-801-810

8. Chaschin S. V. Application of machine learning supervised learning methods in text attribution: distinct approaches and intermediate results in identification of the authors of Russian texts. *Voprosy kriminologii, kriminalistiki i sudebnoi ekspertizy,* 2018, (1): 139–147. (In Russ.)
9. Tikhomirova E. A. The statistical analysis in the problem of the author identification of a natural language text. *Nauka i obrazovanie: nauchnoe izdanie MGTU im. N. E. Baumana*, 2017, (6): 131–146. (In Russ.) DOI: 10.7463/0617.0001245
10. Mukha A. V., Rozaliev V. L., Orlova Y. A., Zaboleeva-Zotova A. V. Automated approach to authorship atribution. *Izvestiia Volgogradskogo gosudarstvennogo tekhnicheskogo universiteta*, 2013, (14): 51–54. (In Russ.)
11. Marusenko M. A. Identifying the authorship of anonymous and pseudonymous texts by methods of applied linguistics. *Applied linguistics.* Saint-Petersburg: SPbGU, 1996, 469–473. (In Russ.)
12. Golovina T. A. Quantitative analysis of the lingua-personal functioning of parts of speech in descriptive fiction. *Vestnik Tomskogo gosudarstvennogo universiteta*, 2006, (120): 94–105. (In Russ.)
13. Golovina T. A. Parts of speech in lingua-personal space. *University Philology Education: People in the World of Communications*: Proc. Intern. Sci.-Prac. Conf. "Communication in the Modern World: Man in the World of Communications". Barnaul: Izd-vo Alt. un-ta, 2005: 86–87. (In Russ.)
14. Golev N. D., Goryunova M. E. Text as object of quantitative-morphological research. *Kultura i tekst,* 2018, (2): 216–229. (In Russ.)
15. Zakharov M. P. Automation of authorship examination. *Sudebnaia ekspertiza,* 2007, (4): 63–69. (In Russ.)
16. Tambovtsev Iu. A., Tambovtseva A. Yu., Tambovtseva L. A. Typology of distribution of several linguistic units in text as authorship index. *Vestnik Omskogo universiteta*, 2008, (2): 88–96. (In Russ.)
17. Stepanenko A. A. Gender attribution in social network communication: the statistical analysis of pronouns frequency. *Vestnik Tomskogo gosudarstvennogo universiteta*, 2017, (415): 17–25. (In Russ.) DOI: 10.17223/15617793/415/3
18. Batura T. V. *Mathematical linguistics and automatic processing of texts in natural language*. Novosibirsk: RITs NGU, 2016, 166. (In Russ.)
19. Zenkov A. V. A novel method of stylometry based on the statistic of numerals. *Computer research and modeling*, 2017, 9(5): 837–850. (In Russ.) DOI: 10.20537/2076-7633-2017-9-5-837-850
20. Lakin G. F. *Biometrics*. Moscow: Vyssh. shk., 1980, 293. (In Russ.)
21. *Russian grammar. Vol. 1. Phonetics. Phonology. Stress Intonation. Word formation. Morphology*, ed. Shvedova N. Yu. Moscow: Nauka, 1980, 788. (In Russ.)
22. Tambovtsev Iu. A. Phono-typological closeness of linguistic objects by the "Chi-square" criterion. *Vestnik NSU. Series: Information Technologies*, 2010, 8(3): 46–54. (In Russ.)