Частотные характеристики предлогов и их значений в базе данных предложных конструкций

В. В. Выборная, А. М. Гончарова, А. А. Родина

Санкт-Петербургский государственный университет

vvybornaa@gmail.com, sssparzha@gmail.com, rodinany@gmail.com

Аннотация

В статье описываются частотные характеристики соотношения предлогов и их значений, а также исследуется синтаксическая неоднозначность предложных конструкций в русском языке. Материалом исследования послужила база данных предложных конструкций, созданная в ходе проекта «Квантитативная грамматика русских предложных конструкций», который разрабатывался на кафедре математической лингвистики Санкт-Петербургского государственного университета, а также корпус из 200 синтаксически неоднозначных предложений, заимствованных из диссертационного исследования Д. А. Черновой «Процесс обработки синтаксически неоднозначных предложений: психолингвистическое исследование». Анализ данных проведен с помощью инструментов модуля Pandas и других библиотек Python. Мы исходим из положения о том, что предлоги, в особенности первообразные, в разных контекстах реализуют разные значения. Мы рассматриваем соотношение предлогов и приписываемых им семантических меток в целях выявления закономерностей распределения предлогов по синтаксемам, а также определяем наиболее частотные синтаксемы и предлоги среди синтаксически неоднозначных предложений на основе многослойного перцептрона. Результаты данного исследования могут быть полезны при решении задач по снятию омонимии и вносят вклад в понимание структуры синтаксически неоднозначных предложений в русском языке, указывая на доминирующую роль синтаксемы «тематив» в их структуре.

Ключевые слова: русские предлоги, предложные конструкции, значение предлогов, синтаксемы, синтаксическая неоднозначность

Библиографическая ссылка: Выборная В. В., Гончарова А. М., Родина А. А., Частотные характеристики предлогов и их значений в базе данных предложных конструкций // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 61–69. DOI: 10.17586/2541-9781-2024-8-61-69.

1. Введение

Данная статья основана на результатах, полученных в ходе выполнения проекта РФФИ «Квантитативная грамматика русских предложных конструкций» [1, с. 17]. Данный проект разрабатывался на кафедре математической лингвистики Санкт-Петербургского государственного университета. Основной целью проекта стала разработка комплексного квантитативного лексико-грамматического описания русских предлогов и предложных конструкций. В результате работы над проектом была сформирована база данных предложных конструкций, насчитывающая 11122 контекста. Представленные в базе данных контексты послужили материалом для настоящего исследования. Контексты размечены

сразу по нескольким критериям. Для каждой предложной конструкции указывается предлог, управляющее слово, его лемма и часть речи, указывается также зависимое слово, его лемма, часть речи, падеж и число. Более того, каждой предложной конструкции приписывается семантическая метка, определяющее значение, реализующееся в данном контексте [1, с. 17].

Значения, приписываемые предложным конструкциям, представлены как семантические классы, выделенные на основе синтаксем «Синтаксического словаря» Г. А. Золотовой [2]. Такое решение мотивировано тем, что значение предложной конструкции невозможно разложить на значение предлога и значение падежной формы. Предложная конструкция рассматривается как единое целое, т. е. как синтаксема, наделенная определенным значением. При этом одна синтаксема может быть представлена сразу несколькими парами «предлог — падежная форма» [1, с. 20].

Сложность работы с предлогами заключается в том, что предлоги представляют собой весьма неоднородную группу с неясными и неструктурированными значениями. В результате предлоги зачастую остаются без внимания. Например, при автоматическом анализе текста предлоги, как правило, помечаются как «стоп-слова». Однако нельзя забывать о том, что предлоги «передают четкие семантико-синтаксические отношения между знаменательными словами» [3, с. 9]. При семантически ориентированном анализе семантико-синтаксические отношения, выражаемые предлогами, безусловно, оказываются важным аспектом. Более того, ряд исследований показывает, что существует определенная закономерность в распределении служебных единиц, в том числе и предлогов, в разных типах и стилях текстов [4; 5]. Еще одна область, в которой значение предлогов оказывается ключевым, – компьютерное зрение. Интеграция задач обработки естественного языка и компьютерного зрения позволяет найти новые перспективные подходы к описанию и поиску объектов в визуальном пространстве. Предлоги при этом рассматриваются как указатели на пространственные отношения.

В рамках настоящего исследования мы сосредоточились на выявлении закономерностей распределения предлогов по синтаксемам. Это позволит нам получить более точное представление обо всём объёме значений, которые реализуются отдельным предлогом в составе различных предложных конструкций. При этом результаты нашего исследования могут быть полезны при решении целого ряда задач: задач по снятию омонимии, задач атрибуции текстов, задач по определению стилей и типов текстов, задач визуального описания и т. д.

2. Описание частотных характеристик предлогов и их значений

На основании контекстов, представленных в базе данных, можно выделить 5 наиболее частотных синтаксем: локатив (2053 контекстов), темпоратив (1463 контекста), тематив (1334 контекста), объект (943 контекста) и директив (890 контекстов). При этом каждая синтаксема представлена некоторым набором предлогов, которые в определенных контекстах реализуют значение, соответствующее семантической метке. Так, синтаксема локатив чаще всего реализуется при помощи следующих предлогов: в, на, по, за, у. Для синтаксемы темпоратив наиболее частотными являются предлоги в, на, до, за, после. Синтаксема директив чаще всего реализуется при помощи предлогов в, из, на, с, к.

Нетрудно заметить, что, по большому счету, наборы предлогов для каждой синтаксемы несильно отличаются друг от друга. Например, предлог «в» входит в пятерку самых частотных предлогов каждой группы. Следовательно, большую ценность представляет то, в каких именно контекстах тот или иной предлог реализует определенное значение и как на основе этого мы можем описать семантику того или иного предлога.

2.1. Первообразные предлоги и их значения

2.1.1. Предлог «в»

Характерной особенностью первообразных предлогов является их полисемичность. Способность первообразных предлогов реализовывать разные значения в разных контекстах определяет их частотность. Самым частотным предлогом оказывается предлог «в», который охватывает 13 синтаксем, и в базе данных предложных конструкций представлен сразу 3 146 контекстами (рис. 1).

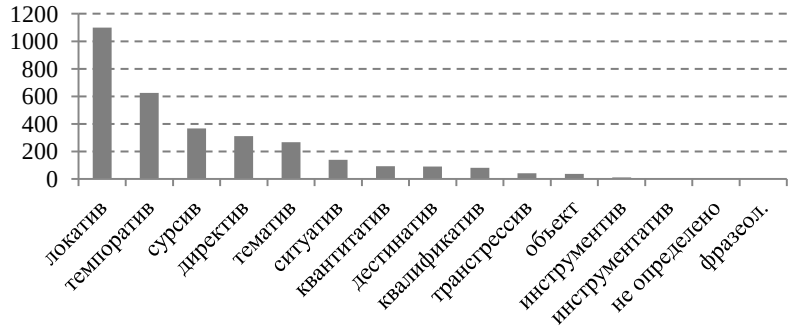

Рис. 1. Столбчатая диаграмма распределения синтаксем предлога «в»

Значение локатива («*происходит в городе*») реализуется почти в два раза чаще, чем значение темпоратива («*в ближайшее время*»). Примечательно то, что следующей по частотности синтаксемой для данного предлога оказывается сурсив, синтаксема со значением «источник информации» («*говорится в письме*»).

2.1.2. Предлог «на»

Второй по частотности предлог, предлог «на», представлен вдвое меньшим числом контекстов по сравнению с предлогом «в». Несмотря на значительную разницу в частотности, предлог «на» охватывает 11 синтаксем (рис. 2).

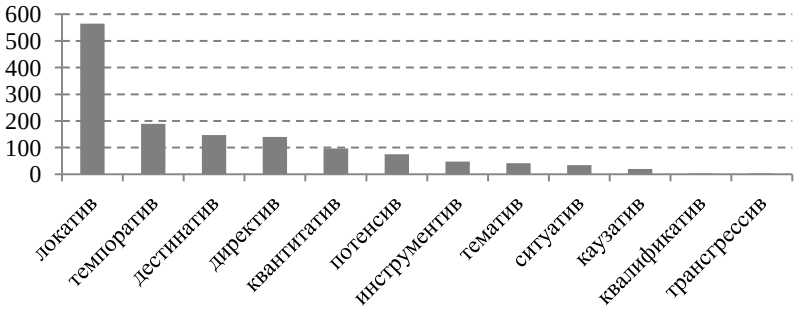

Рис. 2. Столбчатая диаграмма распределения синтаксем предлога «на»

Чаще всего контексты с предлогом «на» реализуют синтаксему локатив, т. е. значение местонахождения («*работал на киностудии*»). Компонент, выражающий временные характеристики, представлен среди прочего следующими контекстами: «*на прошлой неделе*», «*выступит на церемонии*». Чуть меньше 300 контекстов приходится в совокупности на синтаксемы дестинатив и директив, выражающие значения назначения предмета («*цена на газ*») и направления действия или движения («*подняться на второй этаж*») соответственно. В 96 конструкциях реализуется компонент, содержащий

количественные характеристики, синтаксема квантитатив («*сократился на семь процентов*»).

2.1.3. Предлог «о»

Предлог «о» представлен лишь 856 контекстами и охватывает всего две синтаксемы (рис. 3).

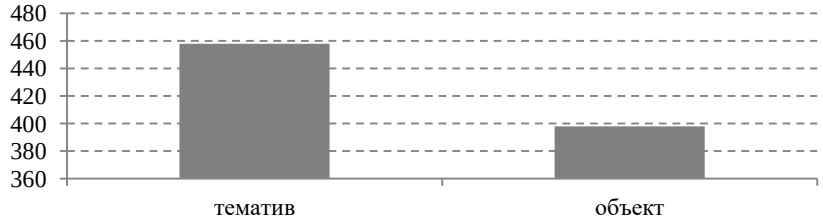

Рис. 3. Столбчатая диаграмма распределения синтаксем предлога «о»

Среди самых частотных первообразных предлогов предлог «о» оказывается семантически наиболее четко очерченным. При этом две синтаксемы: тематив и объект, делят практически поровну весь объём контекстов с данным предлогом. В «Синтаксическом словаре» Г. А. Золотовой тематив и объект определены достаточно четко: тематив — тема оцениваемой ситуации, объект — компонент с предметно-вещественным значением, подвергающийся воздействию [2, с. 431]. Однако при составлении базы данных предложных конструкций, по-видимому, не всегда оказывалось возможным однозначно определить, какая именно синтаксема реализуется в том или ином контексте. Это подтверждается конструкциями, получившими сразу две семантических метки: «*фильмы о Гарри Поттере*», «*уведомить о результатах*» и т. д. Совершенно ясно, что подобные контексты реализуют одновременно две синтаксемы. Отсюда возникают трудности с разграничением конструкций с предлогом «о» на практике.

2.1.4. Предлог «по»

779 контекстов в базе данных предложных конструкций содержат предлог «по», причем данный предлог охватывает 10 синтаксем (рис. 4).

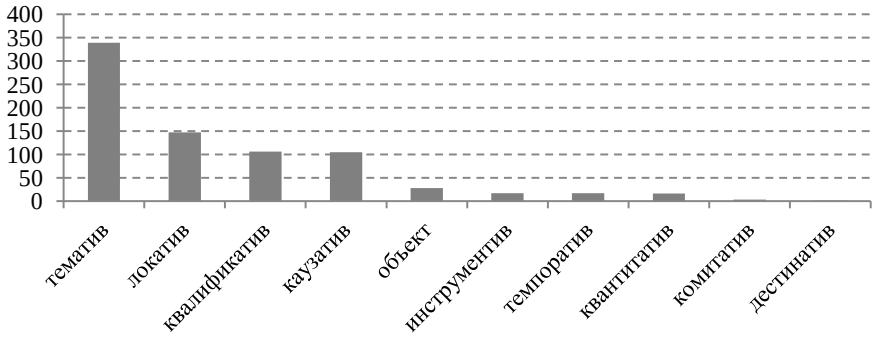

Рис. 4. Столбчатая диаграмма распределения синтаксем предлога «по»

Наиболее характерным для предлога «по» оказывается синтаксема тематив («*комитет по обороне*»). Почти вдвое реже реализуется синтаксема локатив («*гулять по городу*»). Больший интерес вызывают следующие две синтаксемы: квалификатив и каузатив. Поскольку для других производных предлогов данные синтаксемы не так характерны, в совокупности с локативом и темативом они очерчивают семантику предлога «по».

Квалификатив определяется как компонент, обозначающий качество, свойство предмета («*фасад по чертежам*») [2, с. 431]. Каузатив выражает значение причины действия или проявления признака, свойства («*прибывшие по вызову*»). При этом как каузатив, так и квалификатив для предлога «по» обнаруживают одинаковую частотность: на каждую из синтаксем приходится примерно по 100 контекстов.

2.1.5. Предлог «с»

Предлог «с» представлен в базе данных 768 контекстами и, как и первый по частотности предлог «в», охватывает 13 синтаксем (рис. 5). Последний факт ещё раз указывает на разрозненность значений первообразных предлогов. Предлог «с» примечателен среди прочего тем, что наиболее частотная для него синтаксема, комитатив, оказывается лишь на десятом месте среди всех синтаксем первообразных предлогов. В «Синтаксическом словаре» комитатив определяется как «компонент, обозначающий сопровождающее действие, признак, сопутствующий предмет, соучаствующее лицо» («*дом с апартаментами*», «*два с половиной*») [2, с. 431]. Следующая по частотности синтаксема — объект («*гулять с друзьями*», «*пакет с деньгами*»).

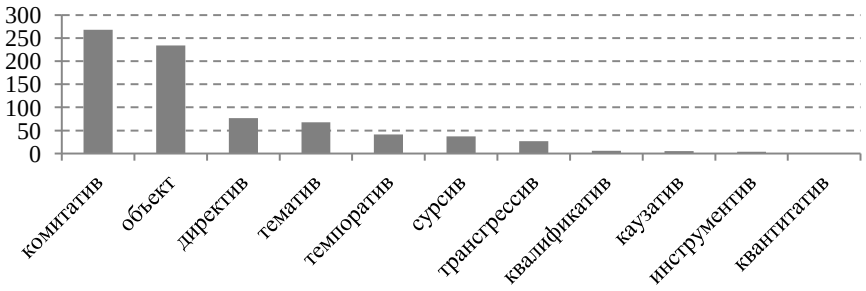

Рис. 5. Столбчатая диаграмма распределения синтаксем предлога «с»

2.2. Производные предлоги и их значения

Путем автоматического анализа свыше 1000 примеров использования было получено распределение синтаксем производных предлогов по частоте. Наиболее частотными оказались следующие: темпоратив, квалификатив, тематив, каузатив, локатив (рис. 6).

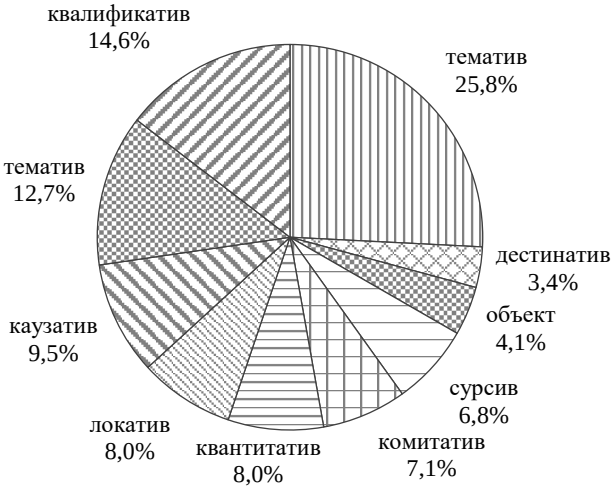

Рис. 6. Круговая диаграмма распределения синтаксем производных предлогов

2.2.1. Предлог «после»

Наиболее частотным среди производных предлогов оказывается предлог «после». Однако в сравнении с первообразными предлогами мы замечаем, что частотность производных в десятки раз меньше (ср. более 3000 контекстов для предлога «в» и чуть больше 100 контекстов для предлога «после»). При этом число производных предлогов почти в 8 раз превосходит число первообразных. Нужно отметить тот факт, что производные предлоги обнаруживают более чёткую семантическую структуру, что объясняется мотивированностью их знаменательными частями речи [3, с. 10]. Так, предлог «после» охватывает лишь три синтаксемы (ср. диапазон значений предлога «в» насчитывает 13 синтаксем). Наиболее характерной для этого предлога оказывается синтаксема темпоратив («*после долгого перерыва*»). Вдвое реже встречаются контексты с данным предлогом, размеченные как каузатив («*задержанные после массовой драки*»).

2.2.2. Предлог «около»

Вторым по частотности среди производных предлогов оказывается предлог «около» (90 контекстов). В первую очередь мы замечаем синтаксемы квантитатив («*около двух метров*») и темпоратив («*около месяца назад*»), на каждую из которых приходится примерно по 40 контекстов. И всего лишь 11 конструкций с предлогом «около» размечены как локатив («*припаркованный около дома*»).

2.2.3. Предлог «между»

Предлог «между» представлен 48 контекстами. При этом большая часть контекстов реализуют значение синтаксемы объект («*разница между сборами и выплатами*»). Нельзя не заметить, что в случае с предлогом «между» все контексты размечаются однозначно, мы не наблюдаем двух семантических меток (тематив и объект) у конструкций с данным предлогом. Всего лишь 11 контекстов приходится на синтаксему локатив («*граница между Турцией и Сирией*»).

2.2.4. Предлог «в связи с»

39 контекстов в базе данных предложных конструкций содержат предлог «в связи с», при этом все контексты почти поровну делятся между синтаксемами каузатив («*отложить в связи с неявкой*») и тематив («*утверждают в связи со вступлением*»). Заметим также, что и в случае с предлогом «в связи с» имеет место неоднозначность. Равное распределение контекстов по двум синтаксемам явно указывает на то, что в действительности синтаксемы тематив и каузатив часто пересекаются.

2.2.5. Предлог «по мнению»

Следующий по частотности предлог — «по мнению» — представлен 37 контекстами, все из которых получили метку сурсив («*по мнению строителей*»). Значение источника информации оказывается единственным для данного предлога.

3. Анализ синтаксической неоднозначности: выявление наиболее частотных синтаксем и предлогов в корпусе неоднозначных предложений

В рамках исследования была изучена синтаксическая неоднозначность на материале корпуса из 200 предложений, заимствованных из диссертационного исследования Д. А. Черновой «Процесс обработки синтаксически неоднозначных предложений: психолингвистическое исследование» [6]. Целью эксперимента являлось определение наиболее проблематичных для анализа синтаксем и предлогов.

Для выявления наиболее частотных предлогов в синтаксически неоднозначных предложениях был проведен компьютерный анализ на языке Python, включая использование стандартных библиотек для обработки строк и модуля collections для количественного анализа элементов текста. Разбиение текста на предложения осуществлялось с помощью библиотеки nltk, а подсчет частотности каждого предлога проводился с использованием структуры данных Counter.

Результаты анализа указывают на значительную частотность употребления предлога «в», составившего 35,29 % от общего числа употреблений предлогов, «на» — 21,32 %, «о» — 13,24 %, «с» — 12,5 %, «за» — 5,88 %, «у» — 5,15 %, «под», «после», и «над» составили 2,21 %.

Для определения наиболее часто встречающейся синтаксемы был задействован классификатор на основе многослойного перцептрона (MLP). В качестве обучающих данных использовались предложения, размеченные по основным синтаксемам русского языка: локатив, темпоратив, тематив, объект и директив, которые были выявлены в рамках текущего исследования. Была применена модель SentenceTransformer, основанная на предварительно обученной модели «DeepPavlov/rubert-base-cased-sentence», для преобразования предложных конструкций в векторные представления.

Классификатор был обучен на основе базы данных предложных конструкций, созданной в рамках исследовательского проекта «Квантитативная грамматика русских предложных конструкций» [7]. Из общего числа 11 122 контекстов 80 % были выделены для обучающей выборки, а 20 % использовались для тестирования. Для обучения классификатора был использован оптимизационный алгоритм Adam с параметром регуляризации равным 0,001 и максимальным количеством итераций обучения, установленным на уровне 10.

Средняя точность классификатора, взвешенная по объему выборки каждого класса, составила 76 % (табл.). Применение классификатора к исходному набору данных выявило, что синтаксема «тематив» преобладает в 71 % случаев, что указывает на ее доминирующую роль в структуре синтаксически неоднозначных предложений.

Таблица. Оценка работы MLP классификатора

Синтаксемы	**Меры**			
	precision	**recall**	**f1-score**	**support**
директив	0,91	0,65	0,76	115
локатив	0,82	0,92	0,87	303
объект	0,54	0,52	0,53	139
тематив	0,60	0,63	0,61	198
темпоратив	0,92	0,88	0,90	176
accuracy			0,76	931
macro average	0,76	0,72	0,73	931
weighted average	0,76	0,76	0,75	931

Таким образом, результаты исследования указывают на значительную частотность употребления предлога «в» в синтаксически неоднозначных предложениях, а также на преобладание синтаксемы «тематив» в структуре этих предложений.

4. Заключение

Семантическая структура предлогов оказывается недостаточно четко очерченной, но это совсем не означает, что значения предлогов не поддаются структурному описанию. Даже на примере, хоть и небольшой, но семантически размытой, группы первообразных предлогов мы видим, что самые частотные предлоги обнаруживают отчетливые различия в наборе реализуемых значений. Более того, частота встречаемости синтаксемы прямо коррелирует с количеством ее значений. Самые частотные предлоги и синтаксемы оказываются самыми неоднозначными, а со снижением частотности предлога сужается и

диапазон его значений. Интересно то, что данная закономерность характерна именно для производных предлогов. На примере же первообразных предлогов мы увидели, что и не самые частотные предлоги могут охватывать свыше 10 синтаксем.

Полученные результаты могут быть полезны для дальнейших исследований в области синтаксического и семантического анализа текста и разработки методов автоматизированного извлечения синтаксических структур.

Литература

[1] Захаров В. П., Азарова И. В., Головина А. В., Гудков В. В., Москвина А. Д. Квантитативная онтология и база данных русских предлогов // Вестник РФФИ. Гуманитарные и общественные науки. 2022. № 109. С. 17–26.

[2] Золотова Г. А. Синтаксический словарь: репертуар элементарных единиц русского синтаксиса. 4-е изд. М.: Наука, 1988. 440 с.

[3] Азарова И. В., Захаров В. П., Москвина А. Д. Семантическая структура русских предложно-падежных конструкций // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXI Международной объединенной конференции «Интернет и современное общество», IMS-2019, Санкт-Петербург, 30 мая – 2 июня 2018 г. Сборник научных статей). — СПб.: Университет ИТМО, 2018. С. 9–16.

[4] Митрофанова О. А., Москвина А. Д. О роли статистики предлогов в определении стилистической принадлежности русскоязычных текстов // International Journal of Open Information Technologies. 2020. Т. 8, № 11. С. 91–96.

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

[6] Чернова Д. А. Процесс обработки синтаксически неоднозначных предложений: психолингвистическое исследование: автореф. на соиск. ученой степ. канд. филолог. наук: 10.02.19 — теория языка. СПб., 2016. 23 с.

[7] Квантитативная грамматика русских предложных конструкций / Захаров В. П. [и др.] // Github. URL: https://vintagentleman.github.io/qt_prep_gram/ (дата обращения: 31.03.2024).

Frequency Characteristics of Russian Prepositions and Their Meanings

V. V. Vybornaya, A. M. Goncharova, A. A. Rodina

Saint–Petersburg State University

The article describes the frequency characteristics of the preposition's ratio and their meanings, and also explores the syntactic ambiguity of prepositional constructions in the Russian language. The research material was a database of prepositional constructions created during the project «Quantitative Grammar of Russian Prepositional Constructions» developed at the Department of Mathematical Linguistics of Saint Petersburg State University, as well as a corpus of 200 syntactically ambiguous sentences borrowed from D. A. Chernova's doctoral research «The Process of Processing Syntactically Ambiguous Sentences: A Psycholinguistic Study». Data analysis was conducted using tools from the pandas module and other Python libraries. We start from the position that prepositions, especially non-derived ones, implement different meanings in different contexts. We consider the ratio of prepositions and the semantic labels attributed to them with the aim of identifying patterns in the distribution of prepositions across syntactic constructions, and also determine the most frequent syntactic constructions and prepositions among syntactically ambiguous sentences based on a multilayer perceptron. The results of this study can be useful in addressing tasks related to disambiguation and contribute to the

understanding of the structure of syntactically ambiguous sentences in the Russian language, indicating the prevailing role of the «topic» synatxeme in their structure.

Keywords: Russian prepositions, prepositional constructions, prepositional meanings, synatxemes, syntactic ambiguity

Reference for citation: Vybornaya V. V., Goncharova A. M., Rodina A. A. Frequency Characteristics of Russian Prepositions and Their Meanings // Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024, St. Petersburg, June 24–26, 2024). — St. Petersburg: ITMO University, 2024. P. 61–69. DOI: 10.17586/2541-9781-2024-8-61-69.

References

[1] Zaharov V. P., Azarova I. V., Golovina A. V., Gudkov V. V., Moskvina A. D. Kvantitativnaya ontologiya i baza dannyh russkih predlogov // Vestnik RFFI. Gumanitarnye i obshchestvennye nauki. 2022. № 109. P. 17–26. (In Russian)

[2] Zolotova G. A. Sintaksicheskij slovar': repertuar elementarnyh edinic russkogo sintaksisa. 4-e izd. M.: Nauka, 1988. 440 s. (In Russian)

[3] Azarova I. V., Zaharov V. P., Moskvina A. D. Semanticheskaya struktura russkih predlozhno-padezhnyh konstrukcij // Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vypusk 8 (Trudy XXI Mezhdunarodnoj ob"edinennoj konferencii «Internet i sovremennoe obshchestvo», IMS-2019, Sankt-Peterburg, 30 maya – 2 iyunya 2018 g. Sbornik nauchnyh statej). — SPb.: Universitet ITMO, 2018. S. 9–16. (In Russian)

[4] Mitrofanova O. A., Moskvina A. D. O roli statistiki predlogov v opredelenii stilisticheskoj prinadlezhnosti russkoyazychnyh tekstov // International Journal of Open Information Technologies. 2020. T. 8, № 11. S. 91–96. (In Russian)

[5] Sichinava D. V. Ob odnom lingvisticheskom parametre tipologii tekstov: koefficient «pod/nad» // Nauchno-tekhnicheskaya informaciya. Seriya 2. 2003. № 10. S. 27–35. (In Russian)

[6] Chernova D. A. Process obrabotki sintaksicheski neodnoznachnyh predlozhenij: psicholingvisticheskoe issledovanie: avtoref. na soick. uchenoj step. kand. filolog. nauk: 10.02.19 — teorija jazyka. SPb., 2016. 23 s. (In Russian)

[7] Kvantitativnaya grammatika russkih predlozhnyh konstrukcij / Zaharov V. P. [i dr.] // Github. URL: https://vintagentleman.github.io/qt_prep_gram/ (access date: 31.03.2024). (In Russian)