Семантическая классификация русских предложных конструкций с использованием моделей Transformer

А. В. Белый, Д. В. Бойцова, Е. А. Ботвиньева,
В. В. Выборная, А. М. Гончарова, О. А. Митрофанова, А. А. Родина

***Аннотация*— В статье обсуждаются частотные характеристики соотношения предлогов и их значений в базе данных русских предложных конструкций и решается задача разработки эффективного семантического классификатора предложных конструкций. Представленный в статье ресурс был создан в рамках проекта кафедры математической лингвистики Санкт-Петербургского государственного университета «Квантитативная грамматика русских предложных конструкций». Дополнительным источником данных для исследования послужил корпус из 200 синтаксически неоднозначных предложений, заимствованных из диссертационного исследования Д.А. Черновой «Процесс обработки синтаксически неоднозначных предложений: психолингвистическое исследование». В данной работе предлагается принципиально новая древовидная архитектура классификатора, состоящая из главного многоклассового и вспомогательного бинарного классификаторов. Данное решение значительно улучшает качество классификации по сравнению с предыдущими исследованиями. В серии экспериментов лучшее решение основано на классификаторе SVM и модели DeepPavlov/rubert-base-cased, что обеспечивает значение F1-меры 0,76.**

***Ключевые слова*— предложные конструкции, синтаксемы, разрешение неоднозначности, классификация контекстов, языковые модели.**

I. Введение

В области автоматической обработки текста семантическая классификация предложных конструкций (ПК) представляет собой серьезную проблему, особенно в таких языках, как русский, где широко распространена полисемия.

Сложность автоматической работы с ПК и самими предлогами заключается в том, что они представляют собой весьма неоднородную группу с неясными и неструктурированными значениями. В результате они часто не получают должного внимания при автоматической/автоматизированной работе с текстом: во многих случаях они оказываются в составе «стоп-словаря» и пропускаются. Однако нельзя забывать о том, что предлоги «передают четкие семантико-синтаксические отношения между знаменательными словами» [1]. 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 предлогов, в текстах разных типов и стилей [5] [6].

Данная статья основана на результатах, полученных в ходе выполнения проекта «Квантитативная грамматика русских предложных конструкций» [4]. Данный ресурс создан на кафедре математической лингвистики Санкт-Петербургского государственного университета. Его основной целью была разработка комплексного квантитативного лексико-грамматического описания русских предлогов и предложных конструкций. В ходе реализации проекта была сформирована база данных ПК, насчитывающая 11122 контекста, которые послужили основным материалом для исследования [3]. Для каждой ПК указываются: используемый предлог, управляющее и зависимое слова, их леммы, части речи и другие морфологические характеристики. Более того, каждой предложной конструкции приписывается семантическая метка, определяющая значение, реализующееся в данном контексте [4]. Вслед за Г. А. Золотовой отдельные значения ПК рассматриваются как синтаксемы, представленные в «Синтаксическом словаре» [2]. Такое решение мотивировано тем, что значение предложной конструкции невозможно разложить на значение предлога и значение падежной формы. Предложная

Статья получена 25 сентября 2024 г.

Белый Андрей Владимирович, Санкт-Петербургский государственный университет, студент, email: st087202@student.spbu.ru

Бойцова Дарья Валерьевна, Санкт-Петербургский государственный университет, студент, email: st078167@student.spbu.ru

Ботвиньева Екатерина Александровна, Санкт-Петербургский государственный университет, студент, email: st085755@student.spbu.ru

Выборная Вероника Витальевна, Санкт-Петербургский государственный университет, студент, email: vvybornaa@gmail.com

Гончарова Алина Максимовна, Санкт-Петербургский государственный университет, студент, email: sssparzha@gmail.com

Митрофанова Ольга Александровна, Санкт-Петербургский государственный университет, канд. филол. наук, доцент, e-mail: o.mitrofanova@spbu.ru

Родина Анна Андреевна, Санкт-Петербургский государственный университет, студент, email: rodinany@gmail.com

Статья подготовлена по итогам выступления на Международной объединённой конференции «Интернет и современное общество» (IMS-2024).

конструкция выступает как единое целое, то есть как синтаксема, наделенная определенным значением. При этом одна синтаксема может быть представлена сразу несколькими парами «предлог – падежная форма» [4].

В рамках настоящей работы мы сосредоточились на двух задачах: (1) выявить закономерности распределения предлогов по синтаксемам и (2) разработать эффективный классификатор русских предложных конструкций, решив ранее обнаруженные В. В. Гудковым и др. [13] проблемы.

II. Теоретические предпосылки исследования

Классификация предлогов является предметом многих исследований, которые можно разделить на две категории: онтологическое описание предлогов и их значений и классификация предложных конструкций с использованием различных методов и моделей.

A. Онтологическое описание предлогов

Помимо вышеупомянутой фундаментальной работы Г. А. Золотовой по созданию «Синтаксического словаря», стоит упомянуть работу И. В. Азаровой и В. П. Захарова [23]. В ней описана базовая структура семантических категорий локативных и темпоральных синтаксем, полученная с использованием передового метода анализа частотности предложных конструкций. Они предлагают две рубрики: 1) локализация, включающая 4 синтаксемы, а именно: a) локатив, b) директив, c) сурсив и d) трансгрессив; и 2) темпоратив, представленный 3 синтаксемами: a) темпорал b) аспектив и c) таксис. Важно подчеркнуть, что в рамках данной онтологии существуют не только иерархические связи между синтаксемами и субсинтаксемами, но и горизонтальные связи между единицами, принадлежащими к разным семантическим категориям.

Другое онтологическое описание русских предлогов можно найти в статье [9], где особое внимание уделяется структуре синонимических и квазисинонимических семантических отношений между первообразными и производными предлогами в русском языке.

М. В. Хохлова и В. И. Рубинер [14] анализируют частоты и длины предложных конструкций, удовлетворяющих модели «предлог + именная группа», в юридических текстах. Их анализ фокусируется на четырех типах предложных конструкций: «предлог + существительное», «предлог + местоимение / местоименное прилагательное + существительное», «предлог + местоимение / местоименное прилагательное + существительное» и «предлог + прилагательное + существительное».

B. Семантическая классификация предложных конструкций

Несмотря на обозначенные во Введении сложности работы с ПК при автоматическом анализе текста на различных уровнях, семантическая классификация предлогов и содержащих их контекстов привлекает внимание многих исследователей.

Один из подходов, показавших достойные результаты, представлен в статье S. Pawar, A. Mittal et. al. [20]. В ней предлагается использовать языковую модель BERT для получения векторных представлений предлогов с учетом их контекстов и многослойный персептрон (multilayer perceptron, MLP) для предсказания класса значения предлога. Для обучения использовался англоязычный набор данных из соревнования SemEval-2007 Task 6 [17], включающий 34 предлога и 332 класса значений. Точность (accuracy) работы предложенной системы составила 0,87.

В другом исследовании по разрешению семантической многозначности предлогов, использующем набор данных SemEval, H. Gong, J. Mu et al. [12] сравнили два метода: (a) выведение значений без учителя, проводимое с помощью алгоритма кластеризации k-средних, и (б) обучение с учителем, для которого они использовали стандартные реализации машины опорных векторов (support vector machines, SVM), многослойного персептрона (MLP) и k-ближайших соседей (k-nearest neighbors, k-NN). Векторные представления были получены с помощью модели Word2Vec CBOW [19].

Поскольку разрешение семантической неоднозначности русских предлогов – это малоизученная область исследований, необходимо отметить результаты, полученные В. В. Гудковым, В. П. Захаровым и др. [13], [24]. В данных статьях представлен корпус русских ПК, извлеченных из новостного подкорпуса корпуса Taiga. При создании корпуса ПК использовалась предварительно обученная языковая модель UDPipe [22] – для извлечения предлогов и их контекстов; и библиотека pymorphy2 для языка программирования Python – для проведения морфологической разметки в терминах OpenCorpora. На материале корпуса были обучены классификаторы, использующие полученные с помощью моделей FastText [10] векторные представления. Общее качество классификации достигло значения F1-меры 0,65. Однако эффективно распознавались только 4 синтаксемы: дестинатив, квантитатив, локатив и темпоратив, для которых значения F1-меры находятся в интервале от 0,70 до 0,81.

В настоящей статье мы ставим перед собой цель превзойти эти результаты благодаря использованию более современных моделей векторизации ПК и предлагаемой нами древовидной архитектуре классификатора.

III. Исследовательские наборы данных

Напомним, что база данных предложных конструкций содержит 11122 аннотированных контекста употребления первообразных и производных, простых и составных предлогов. Перечень предлогов, которые должны найти отражение в базе данных, был составлен заранее на основе авторитетных словарей и корпусов, указанных в статье [13]. Для удобства работы с предлогами, имеющими различные варианты реализации, была произведена их нормализация (например, варианты предлога *о*: *о*, *об*, *обо*, сводились к единому виду *о*). Как уже говорилось, каждой ПК была

вручную приписана одна из 15 синтаксем из полной онтологии Г. А. Золотовой:

1) локатив: *лежит между горами*;
2) тематив: *думал о сыне*;
3) темпоратив: *откроют к ноябрю*;
4) объект: *ударить об стол*;
5) директив: *плыть против течения*;
6) квалификатив: *слишком стар для службы*;
7) сурсив: *услышала от подруги*;
8) дестинатив: *комната для курения*;
9) квантитатив: *состоять из 3 частей*;
10) комитатив: *Татьяна с Ольгой*;
11) каузатив: *краснеть при мысли*;
12) потенсив: *застраховать от забот*;
13) ситуатив: *богослужение под дождём*;
14) трансгрессив: *превращение из куколки*;
15) инструментив: *протрите сквозь сито*.

Авторы указывают на проблему дисбаланса синтаксем в полученной базе данных: классы потенсива, ситуатива, трансгрессива и интструментива представлены значительно меньшим числом ПК, чем все остальные. Отмечается, что некоторые синтаксемы являются семантически близкими, а ПК, в которых они реализуются, могут с равной вероятностью относиться к каждой из синтаксем.

A. Описание частотных характеристик предлогов и их значений в базе данных ПК

В данном разделе мы приводим подробное описание частотных характеристик наиболее полно представленных в базе данных 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 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.

1) Первообразные предлоги и их значения

Характерной особенностью первообразных предлогов является их полисемичность. Способность первообразных предлогов реализовывать разные значения в разных контекстах определяет их частотность.

a) Предлог в

Самым частотным первообразным предлогом оказывается предлог *в*, который охватывает 13 синтаксем и в базе данных предложных конструкций представлен сразу 3146 контекстами (см. рис. 1). 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*).

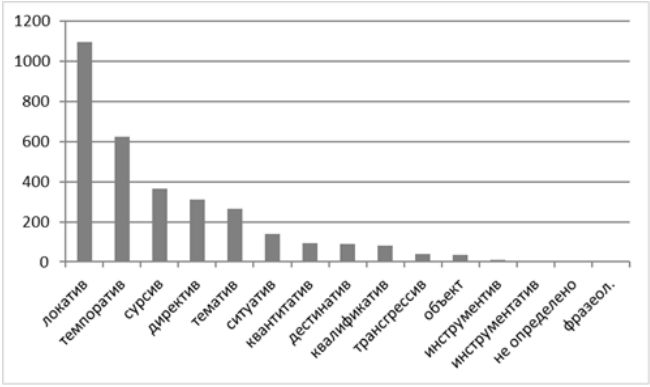

Рис. 1. Столбчатая диаграмма распределения синтаксем предлога *в*

b) Предлог на

Второй по частотности предлог, предлог *на*, представлен вдвое меньшим числом контекстов по сравнению с предлогом *в*. Тем не менее, предлог *на* охватывает 11 синтаксем (см. рис. 2). Чаще всего контексты с предлогом *на* реализуют синтаксему локатив, то есть значение местонахождения (*работал на киностудии*). Компонент, выражающий временные характеристики, представлен такими контекстами как *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*).

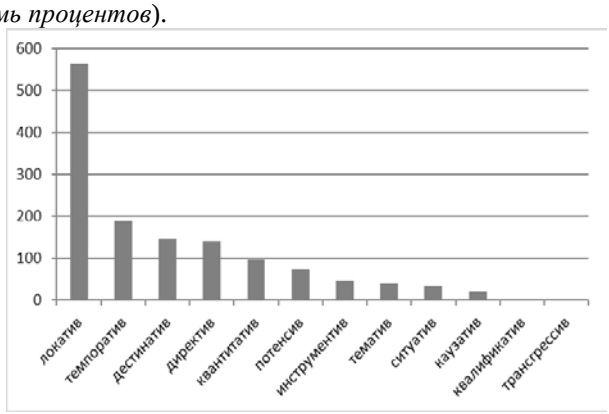

Рис. 2. Столбчатая диаграмма распределения синтаксем предлога *на*

c) Предлог о

Предлог *о* представлен лишь 856 контекстами и охватывает всего две синтаксемы (см. рис. 3). Среди самых частотных первообразных предлогов предлог *о*

оказывается семантически наиболее четко очерченным. При этом две синтаксемы: тематив и объект, делят весь объём контекстов с данным предлогом практически поровну.

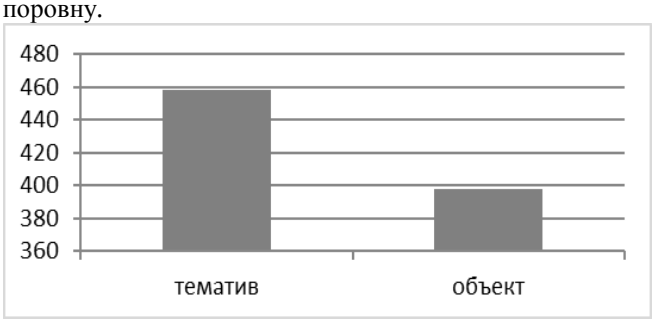

Рис. 3. Столбчатая диаграмма распределения синтаксем предлога *о*

d) Предлог по

В базе данных предложных конструкций 779 контекстов содержат предлог *по*, причем данный предлог охватывает 10 синтаксем (см. рис. 4). Наиболее характерным для него 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]. 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.

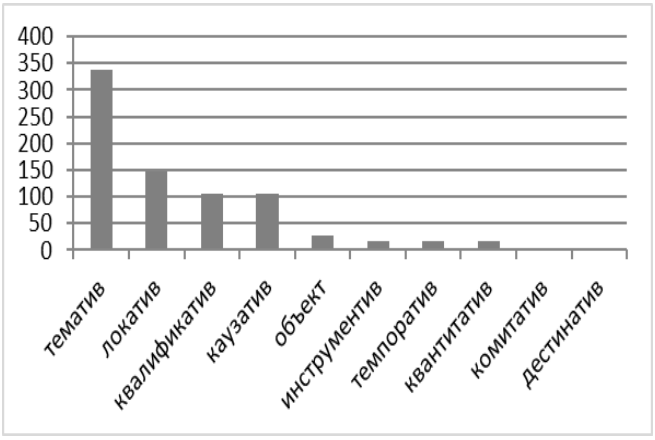

Рис. 4. Столбчатая диаграмма распределения синтаксем предлога *по*

e) Предлог с

Предлог *с* представлен в базе данных 768 контекстами и, как и первый по частотности предлог *в*, охватывает 13 синтаксем (см. 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]. Следующая по частотности синтаксема – объект (*гулять с друзьями*, *пакет с деньгами*).

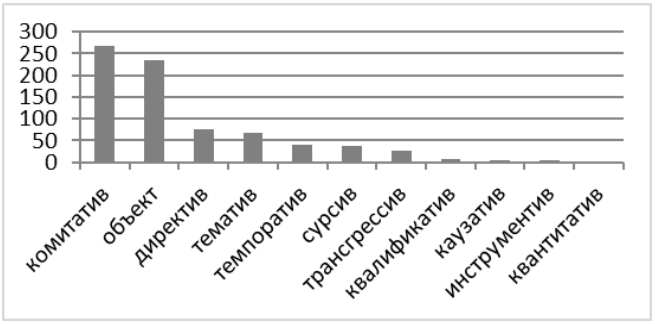

Рис. 5. Столбчатая диаграмма распределения синтаксем предлога *с*

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 (см. рис. 6).

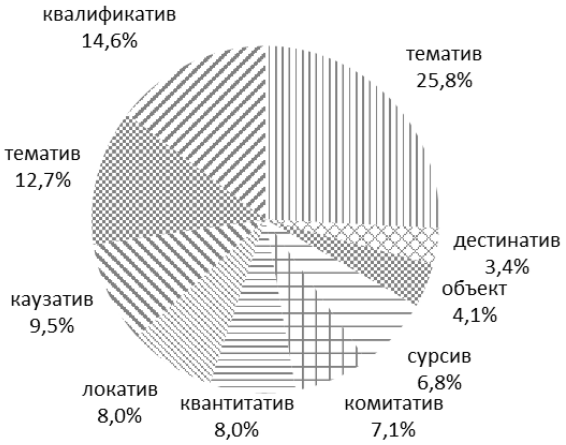

Рис. 6. Круговая диаграмма распределения синтаксем производных предлогов

a) 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 [1]. 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*).

b) 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*).

c) 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*).

d) 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.

e) 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.

B. 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» [7]. Целью эксперимента являлось определение наиболее проблематичных для анализа синтаксем и предлогов

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%.

C. Описание набора обучающих данных

Для обучения семантического классификатора ПК из базы данных нами было выделено подмножество. А именно: мы исключили ПК, содержащие производные предлоги а также 5 наименее представленных синтаксем (то есть каузатив, потенсив, ситуатив, трансгрессив и инструментив). Это решение было обусловлено желанием сосредоточить внимание на разграничении диффузных классов и дистанцироваться от проблемы недостатка данных. Фразеологизированные употребления предлогов были отсечены.

После такой фильтрации объем набора данных составил 8240 пар ПК-синтаксема. На Рис. 7 показано частотное распределение синтаксем в нем.

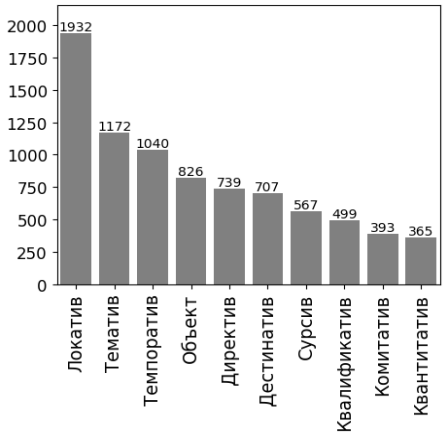

Рис. 7. Столбчатая диаграмма распределения синтаксем в наборе данных.

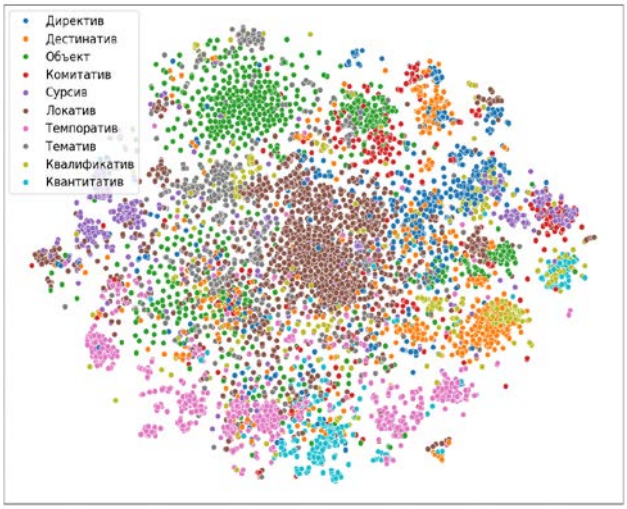

Рис. 8. Визуализация векторного пространства ПК (DeepPavlov/rubert-base-cased).

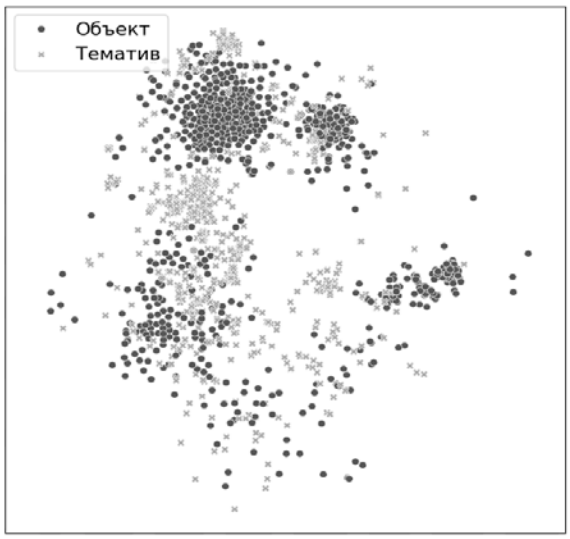

Рис. 9. Визуализация векторных представлений ПК со значениями объекта и тематива (DeepPavlov/rubert-base-cased).

На Рис. 8 показано двумерное отображение векторного пространства ПК, созданное при помощи алгоритма понижения размерности t-SNE, реализованного в библиотеке scikit-learn. Данная иллюстрация демонстрирует диффузную структуру классов. Рис. 9 демонстрирует расположение двух наименее четко противопоставленных классов – синтаксем объект и тематив.

IV. Классификация предложных конструкций

Одной из задач нашей работы было создание эффективного классификатора русских ПК.

Векторные представления ПК в базе данных были получены с использованием различных предобученных языковых моделей архитектуры Transformers:

- sentence-transformers/LaBSE [11];
- cointegrated/LaBSE_en_ru – усеченная версия модели LaBSE, словарь которой включает векторные представления только англо- и русскоязычных токенов [16];
- ai-forever/ruRoberta-large [25];
- RussianNLP/ruRoBERTa-large-rucola – модель RoBERTa-large, дообученная на корпусе RuCoLA (Russian Corpus of Linguistic Acceptability) [18];
- DeepPavlov/rubert-base-cased – дообученная на русскоязычных данных модель BERT, созданная на основе BERT-base-cased [15];
- ai-forever/sbert_large_mt_nlu – дообученная на русскоязычных данных модель BERT, созданная специально для получения векторных представлений предложений целиком [8].

Модель DeepPavlov/rubert-base-cased была преобразована в формат SBERT с помощью библиотеки sentence-transformers [21], которая позволяет создавать векторные представления предложений на основе моделей BERT с дополнительным слоем mean-pooling; остальные модели уже имели необходимый формат.

После получения векторных представлений ПК нами были проведены предварительные эксперименты с помощью нескольких многоклассовых классификаторов, реализованных в библиотеке scikit-learn, а именно:

- машины опорных векторов (SVM);
- машины опорных векторов, обученной с использованием метода стохастического градиентного спуска (stochastic gradient descent, SGD) (SVM-SGD);
- логистической регрессии (logistic regression, LR);
- логистической регрессии, обученной с использованием метода стохастического градиентного спуска (LR-SGD);
- многослойного перцептрона (MPL);
- многослойного перцептрона, обученного использованием метода стохастического градиентного спуска (MLP-SGD).

Базовая реализация представляла собой простую архитектуру с одним классификатором, но при тестировании было обнаружено, что она была неспособна решить проблему разделения диффузных классов.

Как и в работе В. В. Гудкова и др. [13], наиболее заметными примерами неоднозначности являются синтаксемы объект и тематив, получившие значения F1-меры 0,39 и 0,51 соответственно. Заметим, однако, что 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]. Кроме того, эти классы представлены в базе данных значительным числом контекстов, и потому мы считаем, что полученные для них результаты имеют высокую степень достоверности, а вклад в общее качество классификации значителен.

Чтобы повысить качество работы с этими классами, мы предлагаем древовидную архитектуру классификатора, состоящую из двух отдельных последовательных классификаторов.

Таким образом, наша система состоит из основного многоклассового классификатора, присваивающего начальные метки, и вспомогательного бинарного классификатора, специально обученного на примерах из диффузных классов. В случае если первый классификатор относит ПК к одному из диффузных классов, данная ПК отправляется специализированному классификатору, решение которого считается окончательным. Теоретически, такая древовидная архитектура классификатора может быть расширена не одним, а несколькими бинарными классификаторами, однако это остается за рамками описанного здесь эксперимента. В данной работе мы используем идентичные алгоритмы классификации и для базового, и для вспомогательного классификаторов. На Рис. 10 представлена описанная архитектура, а на Рис. 11 – алгоритм ее работы.

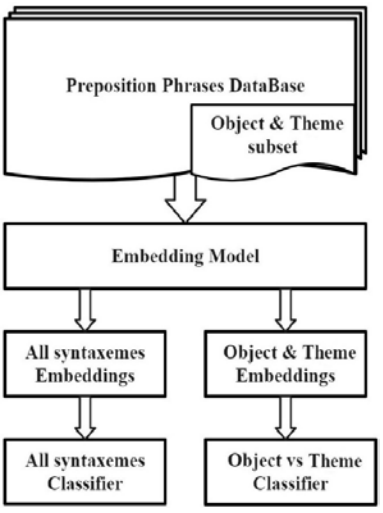

Рис. 10. Архитектура древовидного классификатора.

Результаты классификации оценивались с помощью стандартных метрик. Наиболее важным показателем мы считали значение F1-меры. Чтобы учесть

несбалансированность классов, использовалась ее взвешенная версия, которая вычисляет среднее значение по всем классам пропорционально количеству элементов в каждом классе.

Для обеспечения достоверности оценок мы использовали метод кросс-валидации StratifiedKFold с 10 фолдами (fold, подмножество обучающих данных). Стандартное отклонение значений F1-меры, полученными для каждого фолда, не превышает 0,002.

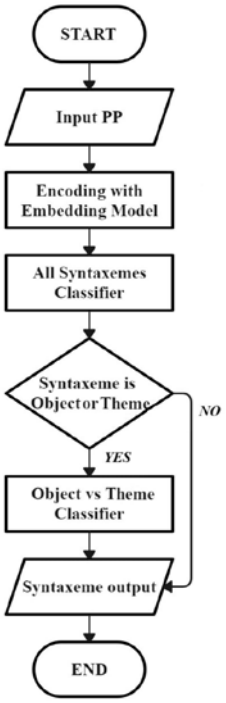

Рис. 11. Блок-схема работы классификатора древовидной архитектуры.

V. Результаты экспериментов по классификации предложных конструкций

В таблице сравниваются результаты одинарного (baseline) и древовидного (final) классификаторов. Можно заметить, что и выбор классификатора, и выбор языковой модели влияют на качество.

В целом, все классификаторы с обучением методом SGD получили несколько более низкие оценки, чем базовые реализации тех же классификаторов. Это различие становится более значительным для модели LaBSE, где оценка F1 снижается с 0,75 для простого SVM до 0,53 для SVM-SGD.

Таблица. Качество классификации (F1-мера)

Model		**LaBSE**	**LaBSE_ en_ru**	**ruRoberta-large**
SVM	baseline	0,71	0,71	0,72
	final	0,75	0,75	0,73
MLP	baseline	0,65	0,68	0,7
	final	0,68	0,71	0,75
LR	baseline	0,70	0,68	0,71
	final	0,72	0,72	0,75
SVM-SGD	baseline	0,50	0,65	0,70
	final	0,53	0,68	0,74
MLP-SGD	baseline	0,47	0,66	0,69
	final	0,48	0,68	0,72
LR-SGD	baseline	0,53	0,66	0,72
	final	0,57	0,70	0,75

Model		**ruRoBERTa-large-rucola**	**rubert-base-cased**	**sbert_large_ mt_nlu_ru**
SVM	baseline	0,71	0,75	0,65
	final	0,72	0,76	0,68
MLP	baseline	0,66	0,67	0,64
	final	0,71	0,71	0,68
LR	baseline	0,70	0,70	0,66
	final	0,70	0,70	0,70
SVM -SGD	baseline	0,70	0,70	0,65
	final	0,71	0,70	0,68
MLP-SGD	baseline	0,63	0,62	0,65
	final	0,64	0,64	0,68
LR-SGD	baseline	0,69	0,69	0,65
	final	0,70	0,70	0,68

Из трех типов классификаторов, использованных для тестирования, классификатор SVM в большинстве случаев оказывается лучшим. Классификатор MLP демонстрирует худшие показатели, что можно объяснить его склонностью фокусироваться на несущественных особенностях набора данных и проблемах с коллинеарностью.

После усреднения оценок по всем классификаторам лучшей векторной моделью оказалась ai-forever/ruRoberta-large, где спад для SGD версий классификаторов менее всего выражен. Однако самый лучший результат был получен на другой модели: классификаторе SVM с моделью DeepPavlov/rubert-base-cased, получившем значение F1-меры равное 0,76. Оно превосходит указанное в [13] значение F1-меры в 0,65 по всем синтаксемам.

Чтобы оценить влияние предложенной нами архитектуры, результаты были сопоставлены с базовым решением – одинарным классификатором того же типа, обученным на тех же векторных представлениях. Несмотря на среднюю производительность вспомогательных классификаторов (F1-мера ≈ 0,6), их добавление стабильно улучшает общую производительность в подавляющем большинстве случаев. Изменения могут показаться незначительными, но стоит отметить, что что улучшение по всем классам достигается за счет улучшения по наиболее диффузным синтаксемам (тематив и объект). Значение F1-меры для этих двух классов растет с 0,4 до 0,6 в лучшем случае.

Достигнутые результаты позволяют рассчитывать на применение предложенной архитектуры к большему числу классов для дальнейшего повышения производительности, но эти эксперименты остаются за рамками текущей работы.

VI. Заключение

В статье рассмотрены различные подходы к описанию системы русских предлогов, проведен анализ базы данных русских предложных конструкций, предложена, реализована и протестирована новая древовидная архитектура классификатора.

Семантическая структура некоторых предлогов и синтаксем оказывается недостаточно четко очерченной, однако это не помешало нам установить, что частота встречаемости синтаксемы прямо коррелирует с количеством ее значений, и превзойти ранее полученные на данном материале показатели, продвинувшись в решении проблемы разделения диффузных классов. Наилучший результат (средневзвешенная F1-мера равная 0,76) получен для классификаторов SVM и модели DeepPavlov/rubert-base-cased.

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

Библиография

[1] И.В. Азарова, В.П. Захаров, А.Д. Москвина, “Семантическая структура русских предложно-падежных конструкций,” *Компьютерная лингвистика и вычислительные онтологии – Интернет и современное общество*: труды XXI Международной объединенной конференции (Санкт- Петербург, 30 мая – 2 июня 2018 г., выпуск 2). СПб. № 2. С. 9–16. 2018.

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

[3] Квантитативная грамматика русских предложных конструкций / В.П. Захаров [и др.]. URL: https://vintagentleman.github.io/qt_prep_gram/ (дата обращения: 24.11.2024).

[4] Квантитативная онтология и база данных русских предлогов / В.П. Захаров, [и др.], *Вестник РФФИ. Гуманитарные и общественные науки*. № 109. С. 17–26. 2022.

[5] О.А. Митрофанова, А.Д. Москвина, “О роли статистики предлогов в определении стилистической принадлежности русскоязычных текстов,” *International Journal of Open Information Technologies*. Vol. 8. № 11. P. 91–96. 2020.

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

[7] Д.А. Чернова, “Процесс обработки синтаксически неоднозначных предложений: психолингвистическое исследование,” Автореф. на соиск. ученой степ. канд. филолог. наук: 10.02.19 – теория языка. СПб., 2016.

[8] ai-forever/sbert_large_mt_nlu_ru | HuggingFace, URL: https://huggingface.co/ai-forever/sbert_large_mt_nlu_ru, (дата обращения: 24.11.2024).

[9] I. Azarova, M. Khokhlova, V. Zakharov & V. Petkevič, “Ontological Description of Russian Prepositions,” *Proceedings of the III International Conference on Language Engineering and Applied Linguistics*, Saint Petersburg, Russia, November 27, 2019. P. 245–257. 2019.

[10] P. Bojanowski, E. Grave, A. Joulin & T. Mikolov, “Enriching Word Vectors with Subword Information,” *Transactions of the Association for Computational Linguistics*. Vol. 5. P. 135–146. 2016.

[11] F. Feng, Y. Yang, D.M. Cer, N. Arivazhagan & W. Wang, “Language-agnostic BERT Sentence Embedding,” *Annual Meeting of the Association for Computational Linguistics*. 2020.

[12] H. Gong, J. Mu, S. Bhat & P. Viswanath, “Preposition sense disambiguation and representation,” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP, Brussels, Belgium. P. 1510–1521. 2018.

[13] V. Gudkov, A. Golovina, O. Mitrofanova & V. Zakharov, “Russian Prepositional Phrase Semantic Labelling with Word Embedding-based Classifier,” A. Ronzhin, T. Noskova, A. Karpov (eds.) *R. Piotrowski's Readings in Language Engineering and Applied Linguistics*. PRLEAL-2019. CEUR Workshop Proceedings. Vol. 2552. P. 272–284. 2019.

[14] M. V. Khokhlova & V. I. Rubiner, “On quantitative analysis of Russian prepositional constructions based on legislative texts,” *Proceedings of the International Conference Corpus Linguistics-2019*. Saint Petersburg, Russia. P. 149–154. 2019.

[15] Y. Kuratov & M. Arkhipov, “Adaptation of Deep Bidirectional Multilingual Trans formers for Russian Language,” *ArXiv preprint*. URL: https://arxiv.org/abs/1905.07213. (дата обращения 24.11.2024). 2019.

[16] LaBSE-en-ru | HuggingFace, URL: https://huggingface.co/cointegrated/LaBSE-en-ru, (дата обращения: 24.11.2024).

[17] K.C. Litkowski & O. Hargraves, “SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions,” *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic. P. 24–29. 2007.

[18] V. Mikhailov, T. Shamardina, M. Ryabinin, A. Pestova, I. Smurov & E. Artemova, “RuCoLA: Russian Corpus of Linguistic Acceptability,” *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates. P. 5207–5227. 2023.

[19] T. Mikolov, K. Chen, G.S. Corrado & J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *International Conference on Learning Representations*. 2013.

[20] S. Pawar, S. Thombre, A. Mittal, G Ponkiya & P. Bhattacharyya, “Tapping BERT for Preposition Sense Disambiguation,” *ArXiv preprint*. URL: https://arxiv.org/pdf/2111.13972 (дата обращения 24.05.2024). 2021.

[21] SentenceTransformers, URL: https://www.sbert.net/docs/training/overview.html, (дата обращения 24.05.2024).

[22] M. Straka & J. Straková, “Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe,” *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada. P. 88–99. 2017.

[23] V. Zakharov & I. Azarova, “Grammatical Parallelism of Russian Prepositional Localization and Temporal Constructions,” *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020*, Brno, Czech Republic, September 8–11, 2020, Proceedings. P. 122–134. Springer-Verlag, 2020.

[24] V. Zakharov, A. Golovina, E. Alexeeva & V. Gudkov, “Russian Secondary Prepositions: Methodology of Analysis,” *CEUR Workshop Proceedings*. Vol. 2780. P. 187–201. 2020.

[25] D. Zmitrovich, A. Abramov, A. Kalmykov, M. Tikhonova, E. Taktasheva, D. Astafurov, M. Baushenko, A. Snegirev, T. Shavrina, S. Markov, V. Mikhailov & A. Fenogenova, “A Family of Pretrained Transformer Language Models for Russian,” *ArXiv preprint*. URL: https://arxiv.org/html/2309.10931v4. (дата обращения 24.05.2024). 2024.

Белый Андрей Владимирович, Санкт-Петербургский государственный университет, студент, email: st087202@student.spbu.ru, ORCID: orcidID= 0000-0002-1565-3536.

Бойцова Дарья Валерьевна, Санкт-Петербургский государственный университет, студент, email: st078167@student.spbu.ru

Ботвиньева Екатерина Александровна, Санкт-Петербургский государственный университет, студент, email: st085755@student.spbu.ru

Выборная Вероника Вительевна, Санкт-Петербургский государственный университет, студент, email: vvybornaa@gmail.com, ORCID: orcidID= 0009-0008-0041-9705

Гончарова Алина Максимовна, Санкт-Петербургский государственный университет, студент, email: sssparzha@gmail.com, ORCID: orcidID= 0009-0003-3542-2801

Митрофанова Ольга Александровна, Санкт-Петербургский государственный университет, канд. филол. наук, доцент, e-mail: o.mitrofanova@spbu.ru, ORCID 0000-0002-3008-5514
Родина Анна Андреевна, Санкт-Петербургский государственный университет, студент, email: rodinany@gmail.com, ORCID: orcidID= 0009-0002-4730-1946

Semantic Classification of Russian Prepositional Phrases with Transformer Embeddings

Andrei V. Belyi, Daria V. Boitsova, Ekaterina A. Botvineva,
Veronica V. Vybornaya, Alina M. Goncharova, Olga A. Mitrofanova, and Anna A. Rodina

***Abstract*— The article describes frequency characteristics of the preposition's ratio and their meanings in the database of Russian prepositions and considers the task of creating an effective semantic classifier of prepositional phrases trained and testes on the dataset. The database of Russian prepositions discussed in the article was created within the framework of the project 'Quantitative Grammar of Russian Prepositional Constructions' developed at the Department of Mathematical Linguistics of Saint Petersburg State University. The study was also based on a corpus of 200 syntactically ambiguous sentences described in D.A. Chernova's doctoral research "The Process of Processing Syntactically Ambiguous Sentences: A Psycholinguistic Study". In the present work a novel tree-based classifier architecture consisting of a main multiclass classifier and a supportive binary classifier is proposed. This architecture significantly improves performance compared to previous work, both in overall and on previously troublesome highly confused classes. Experiments were conducted with different types of classifiers and various embedding models for the Russian language used for encoding the dataset. The best solution provides F1-score of 0,76 leveraging SVM classifiers and a DeepPavlov/rubert-base-cased model.**

***Keywords*— prepositional phrases, syntaxemes, word sense disambiguation, context classification, phrase embeddings, transformers.**

REFERENCES

[1] I.V. Azarova, V.P. Zakharov & A.D. Moskvina, "Semantic Structure of Russian Prepositional Constructions," *Computernaja lingvistika i vychislitelnye ontologii – Internet i sovremennoe obshhestvo*: trudy XXI Mezhdunarodnoj obedinennoj konferencii (Sankt-Peterburg, 30 maja – 2 iunja 2018 g., vypusk 2). SPb. № 2. P. 9–16. 2018.

[2] G.A. Zolotova, "Syntactic dictionary: Repertory of elementary units of Russian Syntax,". Moscow: Nauka, 1988. (In Russian)

[3] Quantitative grammar of Russian prepositional constructions / V.P. Zakharov et al. URL: https://vintagentleman.github.io/qt_prep_gram/, (last accessed 24.11.2024).

[4] Quantitative Ontology and Russian Preposition Database / V.P. Zakharov et al., *Russian Foundation for Basic Research Journal. Humanities and social sciences*. № 109. P. 17–26. 2022.

[5] O.A. Mitrofanova & A.D. Moskvina, "On the Role of Prepositional Statistics for Genre Identification of Russian texts," *International Journal of Open Information Technologies*. Vol. 8. № 11. P. 91–96. 2020.

[6] D.V. Sichinava, "Ob odnom lingvisticheskom parametre tipologii tekstov: koe`fficient «pod/nad»," *Nauchno-texnicheskaya informaciya*. Serija 2. № 10. P. 27–35. 2003.

[7] D.A. Chernova, "Process obrabotki sintaksicheski neodnoznachnyh predlozhenij: psicholingvisticheskoe issledovanie," Avtoref. na soick. uchenoj step. kand. filolog. nauk: 10.02.19 – teorija jazyka. SPb., 2016.

[8] ai-forever/sbert_large_mt_nlu_ru | HuggingFace, URL: https://huggingface.co/ai-forever/sbert_large_mt_nlu_ru, (last accessed 24.11.2024).

[9] I. Azarova, M. Khokhlova, V. Zakharov & V. Petkevič, "Ontological Description of Russian Prepositions," *Proceedings of the III International Conference on Language Engineering and Applied Linguistics*, Saint Petersburg, Russia, November 27, 2019. P. 245–257. 2019.

[10] P. Bojanowski, E. Grave, A. Joulin & T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*. Vol. 5. P. 135–146. 2016.

[11] F. Feng, Y. Yang, D.M. Cer, N. Arivazhagan & W. Wang, "Language-agnostic BERT Sentence Embedding," *Annual Meeting of the Association for Computational Linguistics*. 2020.

[12] H. Gong, J. Mu, S. Bhat & P. Viswanath, "Preposition sense disambiguation and representation," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP, Brussels, Belgium. P. 1510–1521. 2018.

[13] V. Gudkov, A. Golovina, O. Mitrofanova & V. Zakharov, "Russian Prepositional Phrase Semantic Labelling with Word Embedding-based Classifier," A. Ronzhin, T. Noskova, A. Karpov (eds.) *R. Piotrowski's Readings in Language Engineering and Applied Linguistics*. PRLEAL-2019. CEUR Workshop Proceedings. Vol. 2552. P. 272–284. 2019.

[14] M. V. Khokhlova & V. I. Rubiner, "On quantitative analysis of Russian prepositional constructions based on legislative texts," *Proceedings of the International Conference Corpus Linguistics-2019*. Saint Petersburg, Russia. P. 149–154. 2019.

[15] Y. Kuratov & M. Arkhipov, "Adaptation of Deep Bidirectional Multilingual Trans formers for Russian Language," *ArXiv preprint*. URL: https://arxiv.org/abs/1905.07213. (last accessed 24.11.2024). 2019.

[16] LaBSE-en-ru | HuggingFace, URL: https://huggingface.co/cointegrated/LaBSE-en-ru, (last accessed 24.11.2024).

[17] K.C. Litkowski & O. Hargraves, "SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions," *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic. P. 24–29. 2007.

[18] V. Mikhailov, T. Shamardina, M. Ryabinin, A. Pestova, I. Smurov & E. Artemova, "RuCoLA: Russian Corpus of Linguistic Acceptability," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates. P. 5207–5227. 2023.

[19] T. Mikolov, K. Chen, G.S. Corrado & J. Dean, "Efficient Estimation of Word Representations in Vector Space," *International Conference on Learning Representations*. 2013.

[20] S. Pawar, S. Thombre, A. Mittal, G. Ponkiya & P. Bhattacharyya, "Tapping BERT for Preposition Sense Disambiguation," *ArXiv preprint*. URL: https://arxiv.org/pdf/2111.13972 (last accessed 24.11.2024). 2021.

[21] SentenceTransformers, URL: https://www.sbert.net/docs/training/overview.html, (last accessed 24.11.2024).

[22] M. Straka & J. Straková, "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe," *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada. P. 88–99. 2017.

[23] V. Zakharov & I. Azarova, "Grammatical Parallelism of Russian Prepositional Localization and Temporal Constructions," *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020*, Brno, Czech Republic, September 8–11, 2020, Proceedings. P. 122–134. Springer-Verlag, 2020.

[24] V. Zakharov, A. Golovina, E. Alexeeva & V. Gudkov, "Russian Secondary Prepositions: Methodology of Analysis," *CEUR Workshop Proceedings*. Vol. 2780. P. 187–201. 2020.

[25] D. Zmitrovich, A. Abramov, A. Kalmykov, M. Tikhonova, E. Taktasheva, D. Astafurov, M. Baushenko, A. Snegirev, T. Shavrina, S. Markov, V. Mikhailov & A. Fenogenova, "A Family of Pretrained Transformer Language Models for Russian," *ArXiv preprint*, URL: https://arxiv.org/html/2309.10931v4. (last accessed 24.11.2024). 2024.

Andrei V. Belyi, Saint Petersburg State University, student, email: st087202@student.spbu.ru, ORCID: orcidID= 0000-0002-1565-3536.
Daria V. Boitsova, Saint Petersburg State University, student, email: st078167@student.spbu.ru
Ekaterina A. Botvineva, Saint Petersburg State University, student, email: st085755@student.spbu.ru
Veronica V. Vybornaya, Saint Petersburg State University, student, email: vvybornaa@gmail.com, ORCID: orcidID= 0009-0008-0041-9705
Alina M. Goncharova, Saint Petersburg State University, student, email: sssparzha@gmail.com, ORCID: orcidID= 0009-0003-3542-2801
Olga A. Mitrofanova, St. Petersburg State University, Ph.D. in Philology, Associate Professor, e-mail: o.mitrofanova@spbu.ru, ORCID 0000-0002-3008-5514
Anna A. Rodina, Saint Petersburg State University, student, email: rodinany@gmail.com, ORCID: orcidID= 0009-0002-4730-1946