ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 46 за 2023 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Э.С. Клышинский, В.А. Ганеева,
Е.А. Клыкова, О.И. Васина,
А.Е. Богданова, О.В. Карпик

Автоматическое извлечение корпуса управления глаголов русского языка

Рекомендуемая форма библиографической ссылки: Автоматическое извлечение корпуса управления глаголов русского языка / Э.С. Клышинский [и др.] // Препринты ИПМ им. М.В.Келдыша. 2023. № 46. 15 с. https://doi.org/10.20948/prepr-2023-46
https://library.keldysh.ru/preprint.asp?id=2023-46

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Э.С. Клышинский, В.А. Ганеева, Е.А. Клыкова,
О.И. Васина, А.Е. Богданова, О.В. Карпик

Автоматическое извлечение корпуса управления глаголов русского языка

Москва — 2023

Клышинский Э.С., Ганеева В.А., Клыкова Е.А., Васина О.И., Богданова А.Е., Карпик О.В.

Аннотация

Рассмотрено дальнейшее развитие корпуса управления глаголов русского языка. Первая версия корпуса была улучшена с помощью следующих методов: фильтрация несущественной информации с помощью частотности, распознавание вводных слов, семантический подход для различения аргументов и дополнений, а также группировка глаголов на основе их семантических фреймов.

Ключевые слова: управление слов, корпус управления слов, русский язык, автоматическая обработка текстов

Klyshinsky E.S., Ganeeva V.A., Klykova E.A., Vasina O.I., Bogdanove A.E., Karpik O.V.

Abstract

This paper presents further development of the Russian Verb Co-occurrences Corpus. The first version of the corpus was improved using a variety of methods, including the application of frequency thresholds to filter out irrelevant information, identification of parenthetical expressions, adoption of a semantics-based approach to differentiate between arguments and adjuncts, and the clustering of verbs based on their semantic frames.

Keywords: verbal government, word co-occurrences corpus, the Russian language, natural language processing

Оглавление

1. Введение

Модель управления, также известная как аргументная структура глагола, описывает грамматические и семантические особенности его зависимых (Клышинский, Кочеткова, 2012). Другими словами, она содержит информацию о семантических и синтаксических актантах глагола (в произвольном случае — некоторой лексемы), а также о способах их грамматического выражения. Разработка словаря управления глаголов является многообещающим направлением в области автоматической обработки текстов, поскольку это поможет выявить модели отношений между глаголами и связанными с ними прямыми, косвенными и предложными дополнениями и подлежащими (в произвольном случае — систему синтаксических связей произвольного слова с другими словами вне зависимости от части речи). Эта информация может быть использована для улучшения языковых моделей и приложений для интеллектуального анализа текста, для преподавания и изучения русского языка как иностранного и других областей лингвистики.

Основной целью данной работы является расширение корпуса управления русских глаголов, созданного Богдановой и Ладонцевым (2022), за счет улучшения метода поиска и фильтрации сочетаний, а также расширения синтаксически размеченной коллекции текстов. Корпус состоит из комбинаций «глагол + существительное» и «глагол + предлог + существительное». Наша работа направлена на фильтрацию различных ошибок, внедрение семантического подхода в классификации управления и создание базы данных с понятным веб-интерфейсом для пользователей.

Мы представляем сочетание методов, разработанных для улучшения корпуса, включая фильтрацию вводных слов, классификацию аргументов-дополнений и анализ фреймов слов. Код проекта и дополнительные материалы можно найти на GitHub. (https://github.com/eaklykova/verb_governance_dictionary_2023)

2. Обзор существующих решений

Ранее уже предпринимались попытки создать словарь управления слов. Однако, они были успешными лишь частично, так как для ручной сборки подобного словаря необходимо длительное время, что приводило к незаконченности и скромным объемам существующих вариантов. Один из таких словарей можно найти у Прокоповича и др. (1981), он включает список из 1219 словосочетаний наряду с теоретическими вопросами именного и глагольного управления. Денисов и Морковкин (1983) дают точные описания комбинаторных особенностей 2500 наиболее распространенных русских слов, включая 727 глаголов. Мельчук и Жолковский (2016) пытались формализовать лексические и синтаксические представления языка, предлагая стандартизировать информацию о структуре управления словами. Однако их база данных содержит только 283 лексические записи.

Чтобы увеличить объем и качество таких ресурсов, последние годы велась работа над Активным словарем русского языка (Апресян и др., 2014-2017). Цель этого словаря – не только представить значения слов, но и дать информацию, которая помогает в речеобразовании, такую как комбинаторные особенности слова в различных значениях и практические условия словоупотребления. Главным недостатком является невероятно трудоемкий процесс сбора информации из словарей и их недоступность в машиночитаемом формате.

Чтобы преодолеть описанные выше трудности, были разработаны автоматические методы создания подобных словарей. Большинство из них опираются на Национальный корпус русского языка (НКРЯ). Словарь глагольной сочетаемости непредметных имен русского языка (Бирюк и др., 2008) и FrameBank (Ляшевская и Кашкин, 2015) используют лингвистические аннотации для создания базы управления слов. Аналогичным образом онлайн-проекты Collocations, Colligations, and Corpora (CoCoCo) (Копотев и др., 2015; Кормачева и др., 2014) или Корпус синтаксических комбинаций (КоСиКо) (Клышинский и др., 2018) использовали НКРЯ для извлечения сочетаний слов и создания баз данных.

Однако следует отметить, что среди этих крупных баз данных недостаточно представлены корпуса глагольного управления, а специализированные печатные и онлайн-словари имеют ограничения по размеру и не содержат конкретной информации о грамматических особенностях аргументов.

Эти вопросы были рассмотрены в исследовательской работе Богдановой и Ладонцева (2022). Они разработали пайплайн, который использует модель DeepPavlov (https://docs.deeppavlov.ai/en/master/) для извлечения моделей глагольного управления. Несмотря на впечатляющий размер и характеристики корпуса, у него есть некоторые недостатки, среди прочего – наличие вводных слов в аргументной структуре. Вводными словами являются такие фразы, как «например» и «по-моему», которые не являются обязательными в предложении и не меняют его первоначальный смысл, поэтому они не являются частью глагольной аргументной структуры. Другим недостатком корпуса, разработанного Богдановой и Ладонцевым (2022), является отсутствие разделения глагольных зависимых на обязательные (аргументы) и необязательные (дополнения).

Работа над корпусом управления русских глаголов все еще продолжается. Например, новый метод Клышинского и Богдановой (2023) предполагает анализ групп слов и их взаимосвязей для улучшения качества корпуса. Однако будущая работа авторов не предполагает фильтров для вводных выражений и более глубокого анализа фреймов. Опираясь на предыдущие исследования как основу, мы стремимся решить эти проблемы.

3. Набор данных

Для достоверности и качества корпуса управления глаголов потребовался значительный объем данных. Тексты из различных источников были собраны с

помощью поисковых систем, дампов Википедии и т.д. Наш корпус включает в себя в общей сложности 3,8 миллиарда слов в шести различных жанрах, как указано в таблице 1.

Таблица 1.

Жанровая структура и размеры корпуса

Жанр	Количество предложений	Количество слов	Предложений %	Слов %
Законодательные акты	12 082 916	354 259 132	5.69%	9.22%
Тематические новости	20 679 206	413 693 321	9.75%	10.77%
Википедия	25 932 220	544 161 541	12.22%	14.16%
Наука	26 764 667	560 907 459	12.61%	14.60%
Беллетристика	66 566 588	963 210 918	31.37%	25.07%
Новости	601 57 847	1 005 384 776	28.35%	26.17%
Всего	**212 183 444**	**3 841 617 147**	**100.00%**	**100.00%**

Для синтаксического анализа мы использовали модель deeppavlov ‘ru_syntagrus_joint_parsing’ (https://github.com/deeppavlov/DeepPavlov), поскольку Богданова и Ладонцев (2022) обнаружили, что эта модель является наиболее подходящим вариантом для создания корпуса управления глаголов, достигнув показателя точности грамматического анализа более 0,95.

4. Метод фильтрации ошибок синтаксического анализа

Ручной анализ случайной выборки из корпуса выявил несколько классов ошибок. В широком смысле их можно разделить на две группы: орфографические ошибки (в т.ч. опечатки) и ошибки синтаксического анализа. Дальнейшее исследование показало, что ошибки синтаксического анализа часто связаны с неправильным определением вводных выражений в качестве аргументов или дополнений к глаголу или неправильным определением грамматических признаков существительного. Например, в следующем предложении *семью* также может пониматься как числительное, означающее количество глаз:

(1) *Он видел их семью своими глазами.*

Разные типы ошибок устранялись с использованием разных подходов. Мы использовали два фильтра, разработанные ранее. Предложный фильтр сопоставляет каждый предлог с набором падежей, с которыми он может встречаться, на основе SynTagRus, размеченного вручную подмножества НКРЯ.

Затем невозможные комбинации предлога и падежа были исключены из нашего корпуса. Помимо этого, мы проверили полученные глаголы по словарю OpenCorpora, используя библиотеку PyMorphy.

Затем мы применили частотный фильтр для уменьшения шума, включив в корпус только глаголы с частотой встречаемости не менее четырех. Кроме того, мы разработали три метода, направленные на минимизацию более специфических типов ошибок, которые описаны в следующих разделах.

4.1. Вводные выражения

Вводные выражения, содержащие существительные, распознаются синтаксическим анализатором deeppavlov как связанные с глаголом, что приводит к неверной информации о структуре актантов глагола. Чтобы улучшить качество словаря, эти конструкции необходимо отфильтровать.

Для составления списка таких выражений был использован следующий простой метод. В русском языке вводные слова и конструкции, находящиеся в начале предложения, обычно обособляются от остальной части предложения запятыми. Большинство русских вводных выражений содержат от 1 до 3 слов. Учитывая это, мы сначала выбрали все выражения из 1-3 слов, обособленные запятыми или расположенные в начале предложения и выделенные запятой. Затем n-граммы с частотой встречаемости в нашей размеченной коллекции менее 3 были отфильтрованы. 500 наиболее частотных конструкций были проанализированы вручную, и выяснилось, что примерно 150 из них действительно являются общепринятыми вводными выражениями. Менее частотные из них содержали значительное количество шума, поэтому они были исключены из дальнейшего рассмотрения.

Фильтрация словаря управления глаголов на основе списка вводных выражений потребовала дополнительного шага в методе составления корпуса глагольного управления: идентификации выражений, содержащих по крайней мере одно существительное. Это было необходимо, поскольку при разработке словаря в качестве возможных актантов глагола рассматривались только существительные. В результате в общей сложности остались 53 конструкции, приведенные в приложении А. Чтобы убедиться, что существительные, содержащиеся в этих конструкциях, не были засчитаны как зависимые от глагола, мы проверили, содержит ли каждое предложение конкретное вводное выражение. Если в предложении встречалось такое выражение, проставленная синтаксическим анализатором связь считалась неверной.

Побочным продуктом такой фильтрации является список наиболее распространенных русских вводных слов и конструкций, упорядоченный по их абсолютной частоте в нашей размеченной коллекции. Этот список может быть полезен для различных задач обработки естественного языка (таких как синтаксический анализ, извлечение информации и т.д.), а также для изучающих русский язык как иностранный. Важно отметить, что мы не фокусировались на взаимном расположении этих конструкций; мы использовали только

подмножество наших данных для их извлечения и не утверждаем, что рейтинг точно отражает весь русский язык. Информация доступна на GitHub (https://github.com/eaklykova/verb_governance_dictionary_2023).

Хотя список не является исчерпывающим, нашей основной целью было избежать ошибочного отбрасывания каких-либо конструкций, а не отфильтровывать каждое отдельное вводное выражение. Более того, в соответствии с законом Ципфа (Zipf, 1939), идентифицируя относительно небольшое число наиболее часто встречающихся вводных выражений, мы можем эффективно отфильтровать большинство таких выражений из окончательного словаря.

4.2. Актанты и сирконстанты

С точки зрения теоретической лингвистики, слова в предложении могут быть разделены на два типа: актанты (необходимые для грамматической корректности предложения или заполняемые значением по умолчанию, выводимым из контекста) и сирконстанты (структурно необязательные) (см., например, (Dowty, 2003) или (Теньер, 1988)). Например, в предложении:

(2) *Он заплатил за работу в евро*.

он и *за работу* являются актантами, в то время как *в евро* является сирконстантом. Учитывая, что актанты, в отличие от сирконстант, необходимы для грамматической правильности предложения, крайне важно выделить их в корпусе глагольного управления.

Для автоматического различения аргументов и дополнений использовались данные из FrameBank (https://github.com/olesar/framebank/tree/master). Данные FrameBank содержат значительное количество тегов для каждой конструкции, не все из которых были необходимы для нашей задачи. Мы сохранили следующие признаки: сказуемое, зависимая фраза, положение зависимого по отношению к сказуемому, грамматическая форма зависимого и зависимый тип. Поскольку FrameBank изначально не предназначался для различения актантов и сирконстант, пока не было возможности реализовать фильтр, способный распознавать такие различия в более крупных группах тегов, таких как "Периферия", которые охватывали конструкции типа "предлог + существительное". В результате наш метод классификации был сосредоточен на выявлении подтипов аргументов, в частности типов "субъект" и "объект", которые уже были помечены. Остальным группам был присвоен тег "CONSTR" для облегчения классификации.

Поскольку целью нашего анализа было противопоставление именных субъектов, объектов и других участников фрейма, мы исключили все глаголы из списка зависимых. Был сохранен только первый глагол из каждой видовой пары глаголов в списке сказуемых, поскольку это не оказывает существенного влияния на сочетаемость глаголов. Хотя такое сокращение параметров может привести к уменьшению разнообразия анализируемых данных, это важно для правильного функционирования классификационной модели.

Впоследствии целевые фразы и связанные с ними ключевые глаголы были векторизованы с использованием предварительно обученной модели FastText (https://fasttext.cc/docs/en/crawl-vectors.html) на основе данных общего сканирования и Википедии. Использование предварительно обученной модели имело крайне важное значение, поскольку включение семантики является фундаментальным аспектом этой части проекта.

На заключительном этапе было использовано несколько классификаторов, таких как логистическая регрессия, Random Forest и Cat boost, чтобы определить наилучший алгоритм для нашего массива данных. Результаты оценивались на основе меткости, точности, полноты и f1-мары, которая будет обсуждаться в разделе 5.

4.3. Словарные фреймы

Чтобы еще улучшить качество корпуса, мы провели анализ распространенных комбинаций партиципантов, чтобы отфильтровать аномальные случаи. Давайте рассмотрим следующее предложение в качестве примера:

(3) *Папа читает книгу дочке.*

В этом предложении глагол *читать* имеет три зависимые: *Папа* тип NO_Nom, *книгу* типа NO_Acc, и *дочке* типа NO_Dat (префикс ‘NO_’ означает, что существительному не предшествует предлог).

Собрав все возможные комбинации, мы подсчитали количество глаголов, которые могли бы управлять каждым из зависимых. Затем мы проанализировали сочетания, которые были связаны только с одним глаголом, чтобы выявить потенциальные ошибки и редкие конструкции. Например, структура аргумента *в_Acc сквозь_Acc* была возможна только для одного глагола, а именно *зарониться*. В НКРЯ этот конкретный глагол, управляющий такими зависимыми, встречается только один раз в тексте 19-го века, указывая на то, что эта модель нетипична для современного русского языка.

5. Результаты

5.1. Фильтрация и оценка

Фильтр OpenCorpora отбросил 0,52% всех глаголов в корпусе. Дополнительно 1,17% случаев были отфильтрованы путем POS-маркировки каждого глагола с библиотеки PyMorphy и исключения тех, что были проанализированы как другая часть речи. У нас осталось в общей сложности 33 170 уникальных глаголов, количество которых было сокращено до 25 839 после применения частотного фильтра с пороговым значением 4.

Что касается комбинаций "глагол + существительное" и "глагол + предлог + существительное", предложный фильтр, основанный на SynTagRus, удалил 0,17% всех (неуникальных) комбинаций, в то время как исключение зависимых

от глагола вводных выражений помогло отбросить 2 878 923 вхождения. Окончательная версия корпуса содержит 216 202 518 комбинаций.

Одной из самых больших трудностей при разработке корпуса глагольного управления была оценка его точности и полноты. Из-за новизны автоматически извлекаемого корпуса такого рода золотого стандарта не существует. Сравнение нашего корпуса с существующими ресурсами, такими как Apresyan et al. (2014-2017), казалось наиболее разумным, но у него был ряд недостатков. Во-первых, Активный словарь содержит только лексемы, начинающиеся с букв А-З. Во-вторых, он включает только наиболее активно используемые лексические единицы (отсюда и название), в то время как корпус управления глаголов может содержать глаголы, отсутствующие во всех словарях, но используемые в современном русском языке. Кроме того, Активный словарь недоступен в машиночитаемом формате, что еще больше усложняет его использование в целях оценки.

Настройка для оценки была следующей. Мы разделили глаголы в словаре на подмножества в зависимости от их частотности: топ-100 наиболее частых глаголов, следующие 100 глаголов и так далее. Из каждого подмножества случайным образом были выбраны пять глаголов, начинающихся с букв А-З, которые сравнили с Активным словарем вручную. Мы обнаружили, что наш словарь охватывает от 78 до 90% сочетаний глагол и зависимое слово (типа "глагол + предлог (необязательно) + падеж"), перечисленных в Активном словаре, в зависимости от частоты. В то же время Активный словарь содержит только 55% из тех, которые встречаются в корпусе глагольного управления. Это говорит о том, что наш корпус обладает более высокой полнотой, что неудивительно, учитывая автоматизацию его составления.

К результатам такой оценки следует отнестись со всей серьезностью, поскольку количество глаголов, которые могут быть оценены вручную, ограничено. Чтобы решить эту проблему, была предпринята попытка оцифровать Активный словарь. Мы использовали программное обеспечение ABBYY FineReader для преобразования PDF-файлов в машиночитаемый формат, а затем извлекли все словарные статьи, относящиеся к глаголам. Для конструктивного сравнения мы рассматривали только глаголы из корпуса управления, которые начинаются с букв А-З. Наш корпус содержит 9045 таких глаголов, в то время как Активный словарь описывает только 1073. Это приблизительное число, поскольку возможны ошибки синтаксического анализа из-за неточностей распознавания текста. Только один глагол из Активного словаря не был найден в нашем корпусе, а именно *высмаркиваться*. Его отсутствие оказалось вызвано одним из примененных нами фильтров: этого глагола нет в словаре OpenCorpora.

Что касается моделей управления, то в Активном словаре насчитывалось в общей сложности 2376 сочетаний глаголов и зависимых слов (около 2,28 на глагол). Важно отметить, что 858 комбинаций были отфильтрованы, поскольку они не соответствовали формату, требуемому для автоматической оценки. К

примеру, мы отбросили сочетания с *КУДА, ОТКУДА, С КАКОЙ ЦЕЛЬЮ* и т.д. Их необходимо было вручную преобразовать в формат "глагол + (необязательный) предлог + падеж" на основе примеров использования, приведенных в Активном словаре; однако это противоречило бы цели автоматизированного метода оценки.

В нашем корпусе управления глаголов насчитывается 174 705 сочетаний глагол + зависимое слово для глаголов, начинающихся с А-З (19,32 на глагол), что делает очевидным факт, что наш корпус обладает гораздо более высокой полнотой. Что касается точности, то в Активном словаре была 31 комбинация, которую не удалось найти в нашем корпусе. Все эти случаи были проверены вручную с помощью НКРЯ.

Отсутствие этих сочетаний объясняется несколькими причинами. Во-первых, некоторые из них неграмматичны. Например, *верить в + РОД, войти в + РОД, восстановить против + ВИН, встать на + РОД, въехать для + ПРЕДЛ, выйти за + РОД, выйти на + РОД, гореть от + ТВОР, договориться о + ТВОР, дуться из-за + ВИН, затрудняться с + ПРЕДЛ, злиться на + РОД* не встречаются в НКРЯ. У *выйти за + РОД, выйти на + РОД* действительно есть 5 и 3 вхождения соответственно, но все это ошибки в Активном словаре, который иногда приводит правильный пример, но помечает его неправильным регистром. Например, в (4) подходящим падежом для сочетания *гореть от* является *РОД* (родительный):

(4) **Гореть**: А2 • *от* ТВОР: *гореть от смущения* (Апресян и др., 2014-2017: том 2, 656).

Другие примеры, такие как *анализировать с точки зрения + РОД, браковать в качестве + РОД, брюзжать по поводу + РОД, взыскать в пользу + РОД, возвысить в глазах + РОД, выбежать навстречу + ДАТ, выиграть в мнении + РОД, выиграть по сравнению с + ТВОР, вымогать под угрозой + РОД, выписать на адрес + РОД, гулять по случаю + РОД, дискредитировать в глазах + РОД, договориться по вопросу + РОД, допустить по результатам + РОД, запастись на случай + РОД, заявлять по поводу + РОД*, скорее всего, отсутствовали из-за распознавания DeepPavlov сложных предлогов как сочетания предлогов и существительных (что приводит к неверной трактовке существительного в качестве аргумента или дополнения глагола).

5.2. Автоматическая классификация актантов и сирконстант

На выделенных данных мы обучили несколько классификаторов. Результаты их работы для актантов и сирконстант представлены в Таблице 2.

Таблица 2.

Доля правильных ответов, точность, полнота и F1-мера трех алгоритмов классификации

Классификатор	Доля правильных ответов	Точность	Полнота	F1-мера
Логистическая регрессия	0.962	0.963	0.962	0.962
Классификатор Случайный лес	0.957	0.958	0.957	0.957
Классификатор CatBoost	**0.974**	**0.975**	**0.974**	**0.974**

Как мы можем видеть, результаты сравнительно высоки, а различия между итоговыми баллами находятся в пределах допустимой погрешности. Тем не менее самым успешным экспериментом стал классификатор Cat boost.

5.3. Словарные фреймы

Как упоминалось ранее, уникальные глагольные фреймы могут указывать на ошибки синтаксического анализа или неграмматичные высказывания. Однако они также могут указывать на редкие лингвистические явления или просто необычные конструкции, поэтому их отфильтровывание привело бы к потере важной информации. Вместо этого такие данные следует анализировать вручную.

Проблема в том, что из более чем 360 000 возможных фреймов, найденных в корпусе управления глаголами, примерно 216 000 были связаны только с одним глаголом, что делало невозможным их раздельный анализ. Частотное распределение глагольных фреймов показано на рисунке 1.

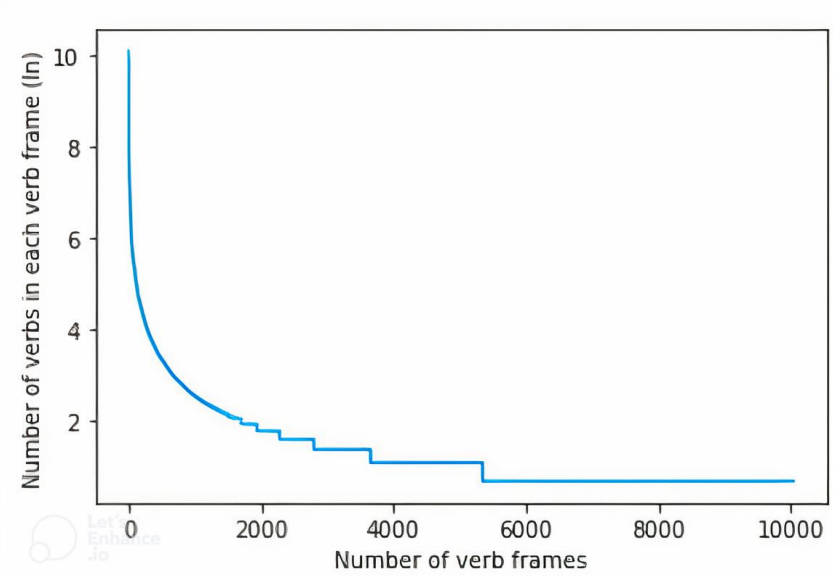

Рис. 1. Частотное распределение глагольных фреймов, представленное в виде натурального логарифма

Для каждого глагола мы также собрали все фреймы, которыми он может управлять, и сравнили глаголы по всему корпусу, чтобы выявить общие закономерности. Частоты различных наборов фреймов представлены на рисунке 2.

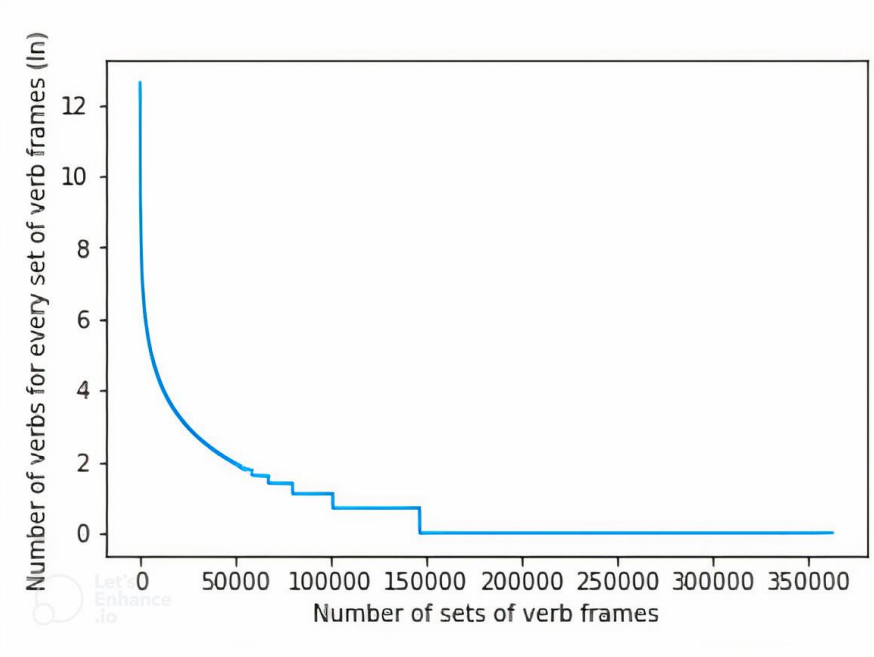

Рис. 2. Частотное распределение наборов глагольных фреймов, представленное в виде натурального логарифма

Как было замечено выше, количество наборов соответствует экспоненциальному распределению. Поскольку уникальных наборов слишком много, чтобы проанализировать их вручную, ниже приводятся некоторые типы ошибок, которые мы извлекли из наших данных с помощью фреймовых фильтров.

Во-первых, в большинстве уникальных фреймов участвуют несколько партиципантов. С одной стороны, это можно объяснить низкой вероятностью использования большого количества партиципантов в одной и той же фразе. С другой стороны, существуют фреймы с несколькими партиципантами, которые являются общими для нескольких глаголов.

Во-вторых, повторы одних и тех же партиципантов (с одним и тем же падежом и предлогом) встречаются в уникальных фреймах чаще, чем во фреймах с несколькими глаголами. Например, структура аргумента с двумя партиципантами именительного падежа (*NO_Gen + на_Loc + NO_Nom + NO_Nom*) встречается только с одним глаголом, а именно *выпучиваться*. Такие экземпляры также могут указывать на ошибки синтаксического анализатора DeepPavlov.

6. Выводы

После расширения массива данных до 3,8 миллиарда слов корпус управления глаголов содержит 25 839 уникальных глаголов и 216 миллионов уникальных глагольных сочетаний. Однако оценка качества корпуса остается сложной задачей из-за отсутствия золотого стандарта. Активный словарь русского языка (Апресян и др., 2014-2017), который является наиболее приемлемым вариантом для сравнения, предоставляет неполные данные, поскольку он охватывает только слова до буквы "З". Точная оценка качества по PDF-версии словаря оказывается практически невозможной из-за его размера в 1867 страниц. Электронная версия словаря существенно упростила бы поиск.

Использование алгоритмов классификации для различения аргументов и дополнений позволило не только извлекать зависимые глаголов, но и классифицировать их как обязательные или необязательные. Однако в будущем необходимо будет извлечь все аргументы, оставшиеся в категории ‘CONSTR’. Использование встраиваемой модели, обученной на НКРЯ, может улучшить результаты.

Для более глубокого понимания глагольной сочетаемости может оказаться полезным дальнейшее изучение информации о группах партиципантов, которые могут встречаться вместе. Опираясь на теоретические лингвистические теории, такие как иерархия чисел Корбетта (2000), мы можем выдвинуть гипотезу о существовании иерархии партиципантов в глагольном управлении. Например, мы могли бы ожидать, что некоторые партиципанты глагола никогда не будут появляться перед другими. Чтобы проверить эту гипотезу, мы можем использовать имеющуюся информацию о глагольных фреймах.

7. Заключение

Словарь русского глагольного управления представляет значительную ценность для лексикографов и других исследователей, а также лиц, изучающих русский как второй язык. В данной работе мы усовершенствовали первую версию корпуса управления глаголами, представленную в работе Богдановой и Ладонцева (2022). Наши улучшения включают такие методы, как применение частотного порога, исключение вводных выражений, использование алгоритмов классификации для различия аргументов и дополнений, а также кластеризация глаголов на основе их семантических фреймов. Для оценки результатов мы использовали оцифрованную версию Активного словаря русского языка (Апресян и др., 2014-2017). Расширенный корпус управления глаголов в настоящее время состоит примерно из 25 тысяч уникальных глаголов и более 216 миллионов уникальных комбинаций.

Список литературы

1. Апресян Ю. Активный словарь русского языка (ред.). М.: Языки славянской культуры, 2014-2017. Т. 1-3.
2. Бирюк О., Гусев В., Калинина Е. Словарь глагольной сочетаемости непредметных имен русского языка. 2008. [Электронный ресурс]. http://dict.ruslang.ru/abstr_noun.php
3. Денисов П., Морковкин В. Словарь сочетаемости слов русского языка (2-е изд.). М.: Русский язык, 1983.
4. Клышинский Э., Кочеткова Н. Метод автоматической генерации модели управления глаголов русского языка. // Материалы 13-й Российской конференции по искусственному интеллекту (RCAI-2012). Белгород, 2012. Том 2. С. 227-235.
5. Мельчук И., Жолковский А. Толково-комбинаторный словарь современного русского языка (2-е изд.). М.: Издательский дом ЯСК, 2016.
6. Прокопович Н., Дерибас Л., Прокопович Э. Именное и глагольное управление в современном русском языке (2-е изд.). М.: Русский язык, 1981.
7. Теньер Л. Основы структурного синтаксиса, М.: Прогресс, 1988, 656 с.
8. Bogdanova A., Ladontsev A. Automatic Extraction of Verb Co-occurrences Corpus [Электронный документ] https://drive.google.com/file/d/12yM2xXrEDpR5XXzqZSLwKn7IjKHdYCcD/view?usp=share_link
9. Buchholz S., Marsi E. CoNLL-X shared task on multilingual dependency parsing // Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), 2006. 149-164. https://doi.org/10.3115/1596276.1596305
10. Corbett G.G. Number // Cambridge Textbooks in Linguistics. Cambridge University Press. 2000. https://doi.org/10.1017/CBO9781139164344
11. Dowty D. The dual analysis of adjuncts/complements in Categorial Grammar // Modifying Adjuncts, 33-66. Berlin, Boston: Mouton De Gruyter. 2003. https://doi.org/10.1515/9783110894646.33
12. Klyshinsky E., Lukashevich N., Kobozeva, I. Creating a corpus of syntactic co-occurrences for Russian. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”. Moscow, 2018. 305–316.
13. Klyshinsky E., Bogdanova A. (in press). Generation of Vocabulary of Verbal Governance for the Russian Language. Proceedings of the 12th International Conference “Natural Language Processing and Corpus Linguistics” Slovko 2023. Bratislava, Slovakia.
14. Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R. CoCoCo: Online Extraction of Russian Multiword Expressions. The 5th Workshop on Balto-Slavic Natural Language Processing. 2015. 43-45.
15. Kormacheva D., Pivovarova L., Kopotev M. Automatic collocations extraction and classification of automatically obtained bigrams. Workshop on

Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations. 2014. 27-33.
16. Lyashevskaya O., Kashkin E. FrameBank: A Database of Russian Lexical Constructions. Analysis of Images, Social Networks and Texts. 2015. 542, 337-348.
17. Lyashevskaya O., Kashkin E. Evaluation of frame-semantic role labeling in a case-marking language. Computational linguistics and intellectual technologies. 2014. 13(20), 362-378.
18. Zipf G. The Psychobiology of Language. London: Routledge. 1936.

Приложение А. Вводные выражения, исключенные из корпуса

Без сомнения, в целом, в среднем, в случае, в принципе, в сущности, в итоге, в результате, в частности, для сравнения, к сожалению, к счастью, к слову, по идее, по традиции, по статистике, по сути, по возможности, по словам, по версии, по мнению, по данным, по оценкам, по оценке, по информации, по прогнозам, по прогнозу, по слухам, по счастью, в Х очередь, с Х стороны, в Х случае, к Х моменту, в Х момент, на Х деле, в Х деле, до Х пор, в Х время, на Х взгляд, к Х времени, по большому счету, в конечном счете, в конечном итоге, в конечном результате, в том числе, по всей вероятности, на первый взгляд, тем временем, таким образом, главным образом, иными словами.