ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет

«Высшая школа экономики»

Факультет гуманитарных наук
Образовательная программа
«Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

На тему
«Автоматическое извлечение типологической информации из грамматик»
«Automatic extraction of linguistic typological information from grammars»

Студентка 3 курса
группы № 193

Долгодворова Мария
Александровна

Научный руководитель
Толдова Светлана Юрьевна,
доцент

Научный Консультант
Сериков Олег Алексеевич,
приглашенный преподаватель

Москва, 2022 г.

Содержание

1. Введение

1.1. Мотивация исследования и постановка задачи

Лингвистическая типология — это систематическое изучение и сравнение языковых структур (Velupillai 2012: 1). Типология позволяет определять границы возможного в человеческих языках и, таким образом, вносить вклад в универсальную теорию грамматики (Bickel, 2007: 1).

Типологическая информация важна для проведения кросс-лингвистических исследований и изучения устройства языков (Velupillai 2012: 30). Кроме того, типологические признаки играют важную роль в автоматической обработке естественного языка (далее в тексте — NLP, сокращенно от анг. Natural Language Proceccing).

Многие алгоритмы NLP основаны на корпусных данных, аннотированных вручную, и лексических базах данных. Для хорошо изученных языков (английского, китайского и т.д.) такие ресурсы доступны для разрешения ключевых задач NLP, однако для большинства других языков отсутствуют размеченные корпуса (O'Horan, Berzak, Vulic´, Reichart, Korhonen 2016: 1299).

Поскольку создание языковой базы данных требует внушительного количества времени и ресурсов, были разработаны альтернативные решения: языковой перенос, создание универсальных моделей, обучение на нескольких языках (O'Horan, Berzak, Vulic´, Reichart, Korhonen 2016: 1300).

Типологическая информация играет важную роль в каждом из ранее упомянутых способов. Так, при создании модели, обученной на нескольких языках, типологическая информация используется для облегчения сопоставления структурных особенностей между языками; в языковом переносе играет важную роль для решения задач автоматической разметки частей речи (далее в тексте — POS-tagging, сокращенно от анг. Part-of-Speech tagging); использование типологических признаков при создании

универсальных моделей может помочь в создании альтернативы существующим ресурсам, согласованную с идеей универсального моделирования (O'Horan, Berzak, Vulic´, Reichart, Korhonen 2016: 1300-1301).

Так, типологическая информация важна как для теоретических исследований в лингвистике, так и в NLP. Возникает вопрос, где брать соответствующую информацию?

В настоящее время существует 6 основных типологических баз данных: World Atlas of Language Structures (WALS), Syntactic Structures of the World's Languages (SSWL), Atlas of Pidgin and Creole Language Structures (APiCS), PHOIBLE Online, Lyon-Albuquerque Phonological Systems Database (LAPSyD), URIEL Typological Compendium. Каждая из них содержит информацию о типологических особенностях большого количества языков (O'Horan, Berzak, Vulic´, Reichart, Korhonen 2016: 1300-1301).

Корпуса предоставляют широкий функционал, однако доступ к грамматикам с типологической информацией ограничен, что усложняет получение подробной информации о признаке. Тогда, оптимальным вариантом можно считать работу с собственной коллекцией грамматик.

Проводя исследования на нескольких языках и работая с ограниченным набором признаков, вполне реально извлечь необходимую информацию вручную, однако, если кто-то стремится расширить охват с нескольких до нескольких сотен признаков, охватывающих тысячи языков, стратегия ручного извлечения информации и сравнения кажется просто неосуществимой. В таких случаях необходимо автоматизировать поиск по грамматикам.

В базах данных содержится классификация языков по признакам с удобным поиском, а в грамматиках — подробное описание этих признаков по языкам без возможности быстрого извлечения информации. Кроме того, типологические признаки с WALS содержат подробное описание, которое

присуще всем грамматикам. Можно предположить, что использование информации из типологических баз данных повысит эффективность поиска по грамматикам.

Таким образом, *гипотеза* исследования заключается в том, что использование описаний типологических признаков из базы данных WALS положительно скажется на автоматическом извлечении типологической информации из грамматик. *Целью* исследования является реализация подобного алгоритма и анализ полученных результатов. *Новизна* работы заключается в использовании информации из типологических баз данных для реализации алгоритма, а также извлечение не только признаков, но и их описаний.

1.2 Предыдущие работы

В последние годы набирают популярность исследования, посвященные автоматическому извлечению типологических признаков, и, в частности, исследованию эффективности различных подходов к разрешению этой задачи.

В статье (Virk et al. 2017) описан простой подход, основанный на эмбеддингах слов в качестве единственного источника знаний (Virk et al. 2017: 1485). Такая стратегия показала неплохие результаты, однако для реализации требует глубокого понимания паттернов, которые в определенных случаях могут быть не очень очевидны.

Метод, обратный подходу, основанному на правилах, описан в (Virk et al. 2019). Авторы использовали теорию фреймовой семантики и фреймово-семантического разбора и обучили модель извлечению типологической информации, используя семантические фреймы. В результате оценки система показала хорошие результаты, однако проблема заключается в ограниченном количестве насыщенных событиями фреймов (Virk et al. 2019: 1254).

В работе (Wichmann, Rama 2019) извлечение типологических признаков состоит из двух этапов: сначала алгоритм обнаруживает ту часть текста, которая

с наибольшей вероятностью содержит описание данного признака, а затем решается задача классификации, состоящая в извлечении ровно одного значения признака из целевого текстового фрагмента. Авторы приходят к выводу, что модель показывает хороший результат для неконтролируемого обучения, однако требует дальнейших улучшений (Wichmann, Rama 2019: 2).

Для извлечения типологической информации авторы статьи (Virk, Foster, Muhammad, Saleem 2021) используют рекуррентную нейронную сеть, обученную на эмбеддингах слов и учитывающую семантические фреймы. Работа является продолжением исследования (Virk et al. 2017) и качество работы модели превосходит полученные ранее результаты (Virk, Foster, Muhammad, Saleem 2021: 1486) .

1.3 Материал исследования и методы

Первоначальным этапом работы стал сбор базы данных дескриптивных грамматик и выделение языка, описываемого в грамматике. В результате получилась коллекция 1244 документов по 722 языкам, более подробное распределение грамматик по языкам представлено на следующих графиках:

Рисунок 1. Языки, наиболее часто встречающиеся в грамматиках

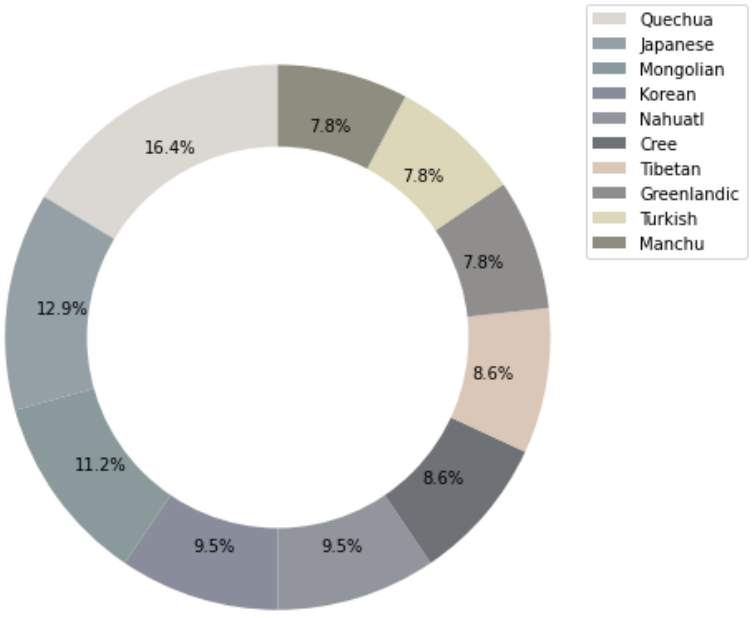
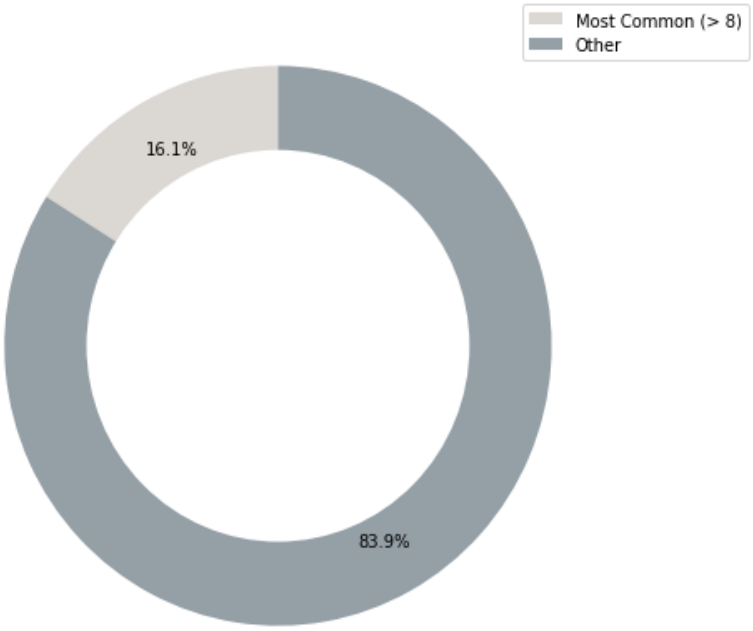


Рисунок 2. Соотношение наиболее и наименее часто описываемых в грамматиках языков



Следующим этапом в обработке данных стало выделение грамматик с распознанным текстом, так как оптическое распознавание символов (далее в тексте — OCR, сокращенно от анг.. Optical Character Recognition) представляет собой отдельную задачу NLP. В результате было отобрано 672 грамматики с распознанным текстом. Для оценки качества из упомянутых ранее грамматик было отобрано 10, посвященных разным языкам, и написанных на английском языке, для минимизации ошибок, связанных с неверно считанным текстом.

Для реализации поиска грамматикам, основываясь на проведенном ранее исследовании (Scholivet, Dary, Nasr, Favre, Ramisch: 2019), в рамках работы были выбраны аналогичные признаки, связанные с устройством порядка слов в

языке. Ниже представлена таблица с идентификатором признака в WALS и расшифровкой:

Таблица 1. Типологические признаки WALS

ID	Feature
1. **81A**	Order of Subject, Object and Verb
2. **82A**	Order of Subject and Verb
3. **83A**	Order of Object and Verb
4. **85A**	Order of Adposition and Noun Phrase
5. **86A**	Order of Genitive and Noun
6. **87A**	Order of Adjective and Noun
7. **88A**	Order of Demonstrative and Noun
8. **89A**	Order of Numeral and Noun
9. **90A**	Order of Relative Clause and Noun
10. **94A**	
11. **92A**	Position of Polar Question Particles
12. **95A**	Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase
13. **96A**	Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun
14. **97A**	Relationship between the Order of Object and Verb and the Order of Adjective and Noun
15. **143A**	Order of Negative Morpheme and Verb
16. **143E**	Preverbal Negative Morphemes
17. **143F**	Postverbal Negative Morphemes

18. **143G** Minor morphological means of signaling negation

19. **144A** Position of Negative Word With Respect to Subject, Object, and Verb

По упомянутым выше признакам были сформированы описания, взятые из базы данных WALS.

2. Реализация алгоритма по извлечению типологических признаков из грамматик

2.1 Подсчет метрик качества работы алгоритма

Прежде, чем перейти непосредственно к описанию системы поиска, необходимо уточнить метод оценки качества работы алгоритма.

В качестве золотого стандарта были выбраны 10 грамматик. По ним вручную были выделены страницы, содержащие типологические признаки WALS. Кроме того, для некоторых типологических признаков нет отдельного описания. Так, например, в грамматике языка сандаве (Steeman 2011) нет информации о порядке генетива и существительного, как и во многих других грамматиках. Такую информацию следует извлекать по глоссам и подобные случаи будут рассмотрены отдельно.

2.2 Реализация алгоритма базового поиска по грамматикам

На вход подается типологический признак (например, «Order of Numeral and Noun»), информацию по которому программа будет искать по грамматикам. Выполняется предварительная обработка введенного запроса и распознанного текста грамматики: приведение к нижнему регистру, лемматизация, очистка от пунктуации и стоп-слов.

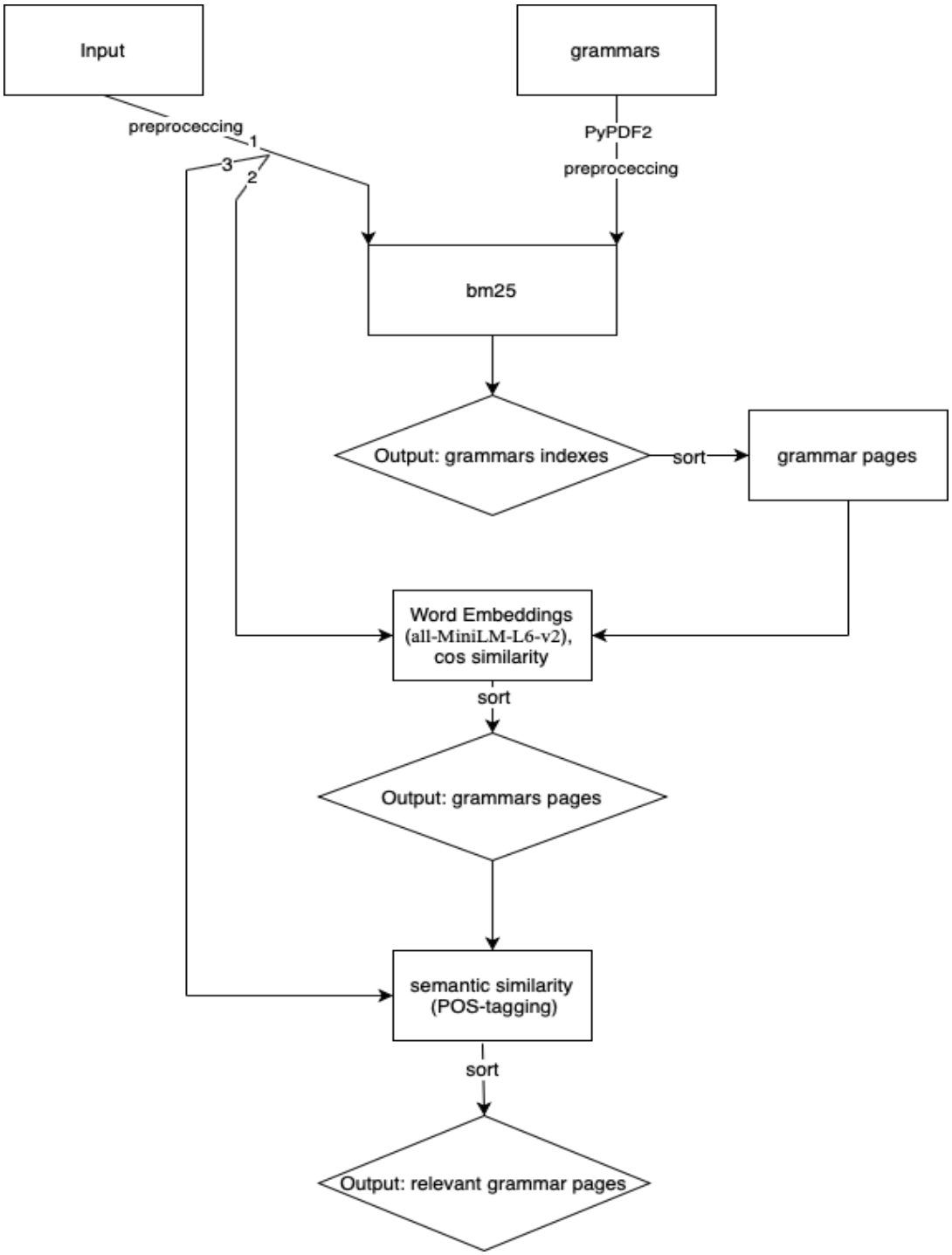
Далее рассчитывается схожесть между страницами текста грамматик и запросом при помощи bm25 (BM25Okapi), по полученным топ-10 результатам возвращаются индексы наиболее релевантных страниц грамматики согласно

формуле. Далее среди ранее полученных страниц вновь выполняется поиск по запросу с использованием эмбеддингов слов (модель all-MiniLM-L6-v2).

Полученные результаты сортируются по значениям косинусного сходства и проходят фильтрацию (в качестве порогового значения косинусного сходства было выбрано 0.39). Релевантные страницы вновь сравниваются с запросом с использованием POS-tagging и синсетов.

В результате работы алгоритма выводится список страниц грамматики, которые содержат искомую типологическую информацию. Подробная реализация алгоритма представлена в виде схемы (см. рис. 3)

Рисунок 3. Схема реализации базового алгоритма поиска по грамматикам



Оценка качества работы алгоритма представлена в следующей таблице:

Таблица 2. Показатели метрик качества работы базового алгоритма поиска

Precision	**Recall**	**F1**
0.44	0.57	**0.497**

2.3 Алгоритм поиска по грамматикам с использованием типологических описаний

Для использования типологических описаний все шаги работы алгоритма остаются прежними, однако меняется формат ввода и сам обрабатываемый запрос. Так, на вход подается идентификатор типологического признака WALS. По нему выделяется типологическое описание, по которому осуществляется дальнейший поиск.

В результате оценки качества работы модифицированного алгоритма получились следующие значения метрик:

Таблица 3. Показатели метрик качества работы алгоритма поиска с использованием типологических описаний

Precision	**Recall**	**F1**
0.28	0.66	**0.39**

В результате поиска по типологическим описаниям точность (precision) работы алгоритма значительно снизилась, однако показатель полноты улучшился. Вероятно, в типологических описаниях содержится лишняя информация, что приводит к выдачи не релевантных страниц поисковиком. Однако, эту проблему можно решить, используя информацию об устройстве типологических признаков в языках.

2.4 Алгоритм поиска по грамматикам с использованием ключевых фраз

В работе (Акинина, Бонч-Осмоловская, Кузнецов, Клинцов, Толдова: 2012) показано, что добавление общей и специфической лексики способно улучшить алгоритм по извлечению информации, хоть и требует дополнительных решений, повышающих точность.

База Данных WALS предоставляет информацию о том, как устроены признаки в различных языках. Используя эти знания, можно модифицировать составленные типологические описания с использованием знаний о том, как устроен тот или иной признак в языке.

Например, в юкагирском языке прилагательное стоит перед существительным (Maslova 1999). Используя эту информацию, можно оптимизировать базовый алгоритм поиска, производя поиск по названию признака (например, Order of Adjective and Noun) и дополняя его ключевой фразой, определяемой в зависимости от устройства признака в искомом языке.

Гипотеза заключается в том, что такой подход положительно скажется на точности работы алгоритма, за счет исключения лишней информации из поискового запроса, ведущей к ошибке 1 рода.

Таблица 4. Фрагмент типологических признаков и ключевых фраз

ID	**Feature**	**Feature Type**	**Key Phrases**
87A	Order of Adjective and Noun	1	adjective preceding the noun
		2	adjective following the noun
		3	adjectives more frequently follow the noun, both orders with neither order dominant
88A	Order of Demonstrative	1	the demonstrative is a separate word which precedes the noun

	and Noun		
		2	the demonstrative is a separate word which follows the noun
		3	the demonstrative is a suffix on the noun
		4	a demonstrative word or affix preceding the noun occurring simultaneously with a demonstrative word or affix following the noun
		5	two or more of the above constructions occur without either being dominant
		6	the demonstrative is a prefix on the noun
89A	Order of Numeral and Noun	1	the numeral precedes the noun
		2	the numeral follows the noun
		3	both orders of numeral and noun occur with neither order dominant

Как предполагалось ранее, такой способ положительно сказался на точности работы алгоритма, при этом полнота осталась прежней. В результате, общее качество работы алгоритма значительно повысилось (ср. с *Таблица 2*):

Таблица 5. Показатели метрик качества работы алгоритма поиска с использованием ключевых фраз

Precision	**Recall**	**F1**
0.69	0.57	**0.62**

2.5 Поиск типологических признаков не имеющих описания в грамматиках

Использование типологической информации для поиска по грамматикам следует из наблюдения о том, что искомая страница грамматики содержит описание, схожее с описанием признака в WALS. Для большинства типологических признаков это действительно так, однако, есть исключеня.

Так, например, информация о порядке генетива и существительного доступна из глосс, в то время как ее текстовое описание отсутствует. В таких случаях использование ключевых фраз не улучшит результат работы модели.

Следующие признаки попадают в данную категорию: Order of Genitive and Noun, Minor morphological means of signaling negation, Relationship between the Order of Object and Verb and the Order of Adjective and Noun, Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun, Relationship between the Order of Object and Verb and the Order of Adjective and Noun.

Наиболее оптимальным решением кажется исключение ключевых фраз для вышеупомянутых признаков и реализация поиска непосредственно по запросу.

3. Анализ полученных результатов

В результате работы было выявлено, что полное описание типологических признаков повышает полноту работы алгоритма, однако значительно ухудшает точность, что негативно сказывается на работе системы в целом (см. Таблица 2, Таблица 3, Таблица 5).

Использование ключевых фраз улучшает качество работы алгоритма в сравнении с базовой реализацией, хоть и требует трудоемкой ручной разметки.

Типологические описания не следует брать в качестве основного критерия по улучшению поиска. Однако, эти признаки можно использовать для улучшения систем, основанных на обучении без учителя.

4. Перспективы исследования

Прежде всего, стоит отметить, что в рамках исследования была рассмотрена лишь часть типологических признаков, тогда как их полный список намного обширнее. Для получения более точных результатов следует расширить коллекцию признаков и провести анализ на большей выборке грамматик.

При добавлении ключевых слов и паттернов улучшается полнота работы алгоритма (Ермакова 2012). Так, же разработанные алгоритмы могут быть улучшены путем добавления ключевых фраз.

Кроме того, отдельную задачу представляет поиск по грамматикам типологических признаков, не имеющих отдельного описания, а представленных в виде глосс.

5. Заключение

В результате работы был реализован базовый алгоритм, осуществляющий поиск по грамматикам, использующий BM25, эмбеддинги слов и информацию о семантической близости.

Результат, полученный при добавлении типологических описаний в систему поиска, подтвердил исследование роли общей и специфической лексики в извлечении информации из текста (Акинина, Бонч-Осмоловская, Кузнецов, Клинцов, Толдова: 2012). Так, показатель полноты увеличился, однако значительно понизилась точность.

Для улучшения метрик качества вместо полных типологических признаков учитывались ключевые фразы, основанные на имеющихся знаниях об устройстве типологических признаков в языках, доступных в типологической базе данных WALS.

Поиск по типологическому признаку с использованием дополнительной информации значительно повысил общее качество работы модели. Таким

образом, можно предположить, что существующие алгоритмы могут быть улучшены путем добавления информации из типологических корпусов.

Литература

Акинина Ю.С., Бонч-Осмоловская А.А., Кузнецов И.О., Клинцов В.П., Толдова С.Ю. (2012). *Роль общей и специфической лексики при извлечении информации из текста на примере анализа события «Ввод новых технологий»*

Ермакова Л. М., Методы извлечения информации из текста // Вестник Пермского университета, серия Математика. Механика. Информатика. - 2012. - Выпуск № 1(9). - С. 77-84.

Balthasar Bickel. 2007. Typology in the 21st century: Major current developments. Linguistic Typology, 11(1): 239–251

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. *Survey on the Use of Typological Information in Natural Language Processing.* In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.

Manon Scholivet, Franck Dary, Alexis Nasr, Benoit Favre, and Carlos Ramisch 2019. Typological Features for Multilingual Delexicalised Dependency Parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3919–3930, Minneapolis, Minnesota. Association for Computational Linguistics.

Shafqat Mumtaz Virk, Azam Sheikh Muhammad, Lars Borin, Muhammad Irfan Aslam, Saania Iqbal, and Nazia Khurram. 2019. *Exploiting Frame-Semantics and Frame-Semantic Parsing for Automatic Extraction of Typological*

Information from Descriptive Grammars of Natural Language. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pages 1247–1256, Varna, Bulgaria. INCOMA Ltd.

Shafqat Mumtaz Virk, Azam Sheikh Muhammad, Lars Borin, Muhammad Irfan Aslam, Saania Iqbal, and Nazia Khurram. 2019. Exploiting Frame-Semantics and Frame-Semantic Parsing for Automatic Extraction of Typological Information from Descriptive Grammars of Natural Language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1247–1256, Varna, Bulgaria. INCOMA Ltd.

Vivenka Velupillai 2012. *An Introduction to Linguistic Typology.*

Virk, Shafqat Mumtaz & Borin, Lars & Saxena, Anju & Hammarström, Harald. 2017. *Automatic Extraction of Typological Linguistic Features from Descriptive Grammars.* 111-119.

Wichmann Søren and Taraka Rama. 2019. *Towards unsupervised extraction of linguistic typological features from language descriptions*. In First Work-shop on Typology for Polyglot NLP, Florence.

Приложение

<https://github.com/knapweedss/typological-information-extraction>