ИЗДАТЕЛЬСТВО грамота

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2024. Том 17. Выпуск 5 | 2024. Volume 17. Issue 5
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Кластерный анализ лингвистических профилей скрытых сообществ

Мамаев И. Д.

Аннотация. Цель исследования – представить кластеры профилей скрытых сообществ на основе лингвистических параметров. В статье проводится анализ структуры и связей между атрибутами кластеров профилей сообществ. Научная новизна исследования заключается в том, что комбинация методов иерархического кластерного анализа скрытых сетевых сообществ и дисперсионного анализа позволит выявить однородность/неоднородность создаваемых авторских текстов на грамматическом и лексическом уровнях. С использованием метода Варда было выделено три кластера лингвистических профилей, каждому из которых была дана формальная оценка Silhouette Score. Содержательная оценка профилей представлена в виде соответствующих лингвистических комментариев. В результате исследования установлено, что для онлайн-публикаций характерно варьирование на уровне синтаксиса, но не на уровне морфологии. Предложенный подход кластеризации сообществ можно применять для идентификации потенциально опасных онлайн-субкультур и лидеров мнений в сетевом пространстве. В результате реализации данного подхода лингвистические профили сообществ дополняются цифровой социодемографической информацией.

EN

Cluster analysis of linguistic profiles of hidden communities

I. D. Mamaev

Abstract. The aim of the study is to present clusters of profiles of hidden communities based on linguistic parameters. The article analyzes the structure and relationships between the attributes of clusters of community profiles. The scientific novelty of the study lies in the fact that the combination of methods of hierarchical cluster analysis of hidden network communities and analysis of variance will reveal the uniformity/heterogeneity of the author's texts created at the grammatical and lexical levels. Using the Ward method, three clusters of linguistic profiles were identified, each of which was given a formal Silhouette Score. A meaningful assessment of the profiles is presented in the form of appropriate linguistic comments. As a result of the study, it was found that online publications are characterized by variation at the level of syntax, but not at the level of morphology. The proposed community clustering approach can be used to identify potentially dangerous online subcultures and opinion leaders in the online space. As a result of the implementation of this approach, linguistic profiles of communities are complemented by digital sociodemographic information.

Введение

В современной корпусной лингвистике особое место занимают узкоспециализированные корпусы, в частности тексты социальных сетей. Подобные корпусы имеют ряд характеристик, таких как ограничения на структуру текстов, графическое сочетание различных символов в рамках одной лексемы, наличие эмотиконов, употребление заимствованной лексики, которая претерпела адаптацию на одном или нескольких языковых уровнях, и др. (Тюленева, 2016; Масликова, 2019; Крылова, 2019). Эти особенности делают тексты социальных сетей интересным материалом как для теоретических, так и для прикладных лингвистических исследований. Результаты анализа таких корпусов имеют практическую ценность в лексикографии, информационном поиске и машинном переводе. Стоит отметить, что для получения качественных итоговых данных используется широкий спектр статистических методов интеллектуального анализа данных, в том числе кластеризация. Идея данного метода заключается в разделении набора элементов на группы (*кластеры*) таким образом, чтобы элементы внутри одного кластера были похожи друг на друга, а элементы из разных кластеров были наиболее различны (Булыга, Курейчик, 2021). Проведение кластерного анализа текстовых данных может быть направлено на решение таких задач, как поиск схожих документов, идентификация текстов-дублетов, оптимизация пользовательских запросов в информационно-поисковых системах и пр. В последнее

Научная статья (original research article) | https://doi.org/10.30853/phil20240250

десятилетие кластеризации подвергаются языковые и социодемографические параметры авторов текстов в рамках лингвистического профилирования (Литвинова, Громова, 2020; Литвинова, Котлярова, Заварзина, 2022).

Настоящая статья завершает цепочку исследований (Мамаев, Митрофанова, 2024; Мамаев, 2024), в которых описывается процесс создания корпуса русскоязычного сегмента социальной сети «ВКонтакте», построение модели скрытых сетевых сообществ на основе тематического единства пользовательских публикаций, а также извлечение усредненных языковых параметров для каждого скрытого сообщества. Материалом для данного исследования выступили описанные в предыдущих работах скрытые тематические сообщества, морфосинтаксические корреляции и лексические корреляции, высчитанные на основе автоматически подсчитанных значений в лингвистическом процессоре Profiling-UD (Brunato, Cimino, Dell'Orletta et al., 2020).

Актуальность данного исследования обосновывается потребностью в критическом анализе результатов ряда исследований в быстро развивающейся области лингвистического профилирования текстов методами интеллектуального анализа данных (в частности, кластеризации) и вероятностно-статистическими методами, особенно для текстов, созданных участниками онлайн-коммуникации, поскольку применение традиционных методов фиксации реальной социодемографической информации (например, с помощью анкетирования) не предоставляется возможным. Цифровые следы в социальных сетях не всегда соотносятся с реальными данными о человеке.

Поставленная в исследовании цель достигается решением следующих задач:

1) составление общей базы лингвистических коррелятов на трех уровнях языка;
2) описание полученных кластеров скрытых сообществ на основе близости лингвистических параметров;
3) выявление тех уровней языка, которые позволяют определить количественные отличия между кластерами сообществ.

Теоретической и методологической основами работы послужили как классические, так и современные подходы к исследованию особенностей интернет-дискурса (Сковородников, 2013; Crystal, 2001) и разножанровых текстовых коллекций компьютерными методами (Нокель, Лукашевич, 2015; Litvinova, Litvinova, Panicheva, 2019), которые использовались при создании корпуса и его обработке. Важным аспектом работы являются способы кластеризации лингвистических данных, представленных в трудах (Тулиев, 2019; Прокофьева, Прокофьева, 2013; Белоусов, Дрожжин, Костенчук, 2015). Наконец, в исследованиях (Степаненко, 2017; Савотченко, Проскурина, 2012; Мамина, 2014) описывается используемый статистический аппарат.

Используемые в исследовании методы обусловлены поставленными задачами: это метод кластерного анализа, который позволит выделить группы профилей по общности морфосинтаксических и лексических признаков, метод статистических расчетов, который позволит определить однородность/неоднородность создаваемых авторских текстов, и метод лингвистического анализа, который позволит дать содержательную оценку полученным количественным показателям.

Практическая значимость заключается в том, что данные о лингвистических профилях и результаты их кластеризации могут применяться для усовершенствования методик, в которых необходима идентификация личности на основе анализа ее индивидуальной речевой деятельности. Результаты могут использоваться на занятиях по корпусной лингвистике и социолингвистике.

Обсуждение и результаты

Результаты корреляционного анализа показали, что способы построения публикаций пользователей в скрытых тематических сообществах различаются на морфологическом, синтаксическом, лексическом и структурном уровнях. Количественные профили скрытых сообществ представлены в Таблице 1, в ней используются следующие условные обозначения: NV – корреляция «имя существительное-глагол»; NAdj – корреляция «имя существительное-имя прилагательное»; VAdv – корреляция «глагол-наречие»; AdjAdv – корреляция «имя прилагательное-наречие»; SenLin – корреляция «длина предложения-степень дистантизации»; LinPrep – корреляция «длина предложения-количество предложных конструкций»; TtrDen – корреляция «коэффициент лексической плотности-коэффициент лексического разнообразия».

На следующем этапе полученные данные использовались для определения сходства между сообществами по лингвистическим параметрам. Используя алгоритмы кластерного анализа, мы можем выделить группы пользователей с похожим языковым стилем. Для визуализации лингвистических данных использовался инструмент Orange (Demšar, Zupan, 2013). Мы загрузили таблицу данных в xlsx-формате, которая содержала информацию о 23 сообществах. Отметим, что при создании кластеров скрытых сообществ мы столкнулись с проблемой пропущенных значений, так как некоторые корреляции оказались незначимыми. Для ее решения мы обратились к методу восстановления количественных данных model-based imputation, который доказал свою эффективность в ряде исследований (Kekez, 2021; Chakraborty, Kim, Sudhir, 2022). По умолчанию в методе используется алгоритм 1-NN, который берет значение из наиболее похожего примера данных.

При анализе корпусов и их параметров используются различные методы кластеризации, в том числе иерархические, неиерархические и гибридные методы. Выбор метода зависит от условий лингвистического эксперимента (объема корпуса, структурных особенностей текстов, наличия ограничений на число кластеров и др.). В рамках данного исследования была использована иерархическая кластеризация, поскольку она обладает преимуществом в виде возможности определения оптимального количества кластеров путем изучения характеристик полученного дерева – дендрограммы. Например, можно выделять отдельные ветви дерева в разные группы, если расстояния между ними достаточно велики. Получившаяся структура удобна для поиска кластеров в ней.

Таблица 1. *Значимые корреляционные параметры тематических публикаций скрытых сообществ*

Скрытое сообщество	Морфологические корреляции				Синтаксические корреляции		Лексические корреляции
	NV	NAdj	VAdv	AdjAdv	SenLin	LinPrep	TtrDen
Армия и государственная безопасность	—	0.7206	0.5245	-0.5196	0.7623	0.87	—
Астрономия	—	—	—	—	—	—	—
Бизнес, коммерция, экономика, финансы	-0.4182	0.4666	0.6419	—	0.6617	0.7257	—
Биология	—	—	—	—	—	—	—
География	—	—	—	—	—	—	—
Дом и домашнее хозяйство	-0.4784	0.4661	0.4008	—	0.6414	0.7081	—
Досуг, зрелища и развлечения	—	0.4044	0.6008	0.1742	0.4394	0.6342	—
Журналистика	—	—	—	—	—	—	—
Здоровье и медицина	-0.4228	0.5434	0.5979	—	0.5239	0.5481	0.286
Информатика	—	—	—	—	—	—	—
Искусство и культура	-0.1924	0.5135	0.6791	—	0.5946	0.7565	—
История	-0.416	—	0.4041	—	0.5197	0.7717	—
Легкая и пищевая промышленность	-0.7069	0.8473	0.6223	-0.5933	—	0.7069	—
Машиностроение	—	—	—	—	—	—	—
Наука и технологии	—	0.6084	—	—	0.6503	0.8392	—
Образование	-0.3633	0.2757	0.4711	—	0.4125	0.5233	0.2262
Политика и общественная жизнь	—	0.4856	0.6715	—	0.4304	0.7179	—
Право	-0.7273	0.8461	0.951	-0.7062	0.6573	0.8182	—
Природа	—	0.6632	0.3301	—	0.5225	0.5713	—
Производство	—	—	—	—	—	—	—
Происшествие	—	0.4637	0.3255	—	0.6093	0.8192	—
Психология	-0.3876	0.4569	0.4512	-0.2244	0.6474	0.6942	—
Путешествие	-0.3515	0.6463	0.6427	-0.3055	0.4886	0.5253	—
Рабочий процесс	-0.3609	0.5776	0.6086	—	0.6018	0.6414	—
Религия	—	—	0.5824	—	0.9077	0.8241	-0.6201
Социология	—	—	—	—	—	—	—
Спорт	-0.2811	0.4523	0.6423	—	0.5205	0.5611	0.2815
Строительство и архитектура	—	0.7	0.8636	—	—	—	—
Техника	—	—	—	—	—	—	—
Транспорт	—	0.7193	0.6636	—	0.548	0.6058	—
Филология	—	—	—	—	—	—	—
Философия	—	—	—	—	—	—	—
Частная жизнь	-0.3672	0.4537	—	0.5034	0.6376	0.7311	—
Эзотерика	—	—	—	—	0.4877	—	—

Для расчета расстояний между рядами данных мы применили евклидову метрику и получили матрицу расстояний, представленную на Рисунке 1. Красные значения указывают на большое удаление сообществ друг от друга на основе лингвистических параметров, что может указывать на их принадлежность к различным кластерам, а синие оттенки указывают на возможность принадлежности сообществ к одному кластеру из-за сходства лингвистических признаков. При построении итоговой дендрограммы не используются значения главной диагонали матрицы расстояний.

Для 23 лингвистических профилей методом Варда было выделено три кластера (см. Рисунок 2), а каждый элемент – профиль сообщества – был оценён с использованием Silhouette Score (см. Рисунок 3). Суть метода Варда заключается в том, что на первом этапе каждый объект считается отдельным кластером. На каждом шаге кластеры объединяются в пары таким образом, чтобы уменьшить сумму квадратов расстояний между объектами внутри кластера (т.е. необходимо минимизировать сумму квадратов расстояний). Процесс объединения кластеров продолжается до тех пор, пока все объекты не будут объединены в один кластер. Значение параметра Silhouette Score показывает, близок ли лингвистический профиль к своей группе по сравнению с другими группами. Отсутствие отрицательных значений указывает на то, что вероятность отнесения лингвистических профилей к «чужим» кластерам снижается.

Представленные на Рисунке 4 данные указывают на то, что разница между морфологическими корреляциями, такими как «имена существительные-глаголы» и «глаголы-наречия», несущественна. Кроме того, в случае медианных значений (желтая линия) для корреляции «глаголы-наречия» различия минимальны. Отчетливые различия между кластерами начинают проявляться на синтаксическом уровне, что можно заметить по асимметрии диаграмм размаха на Рисунке 5.

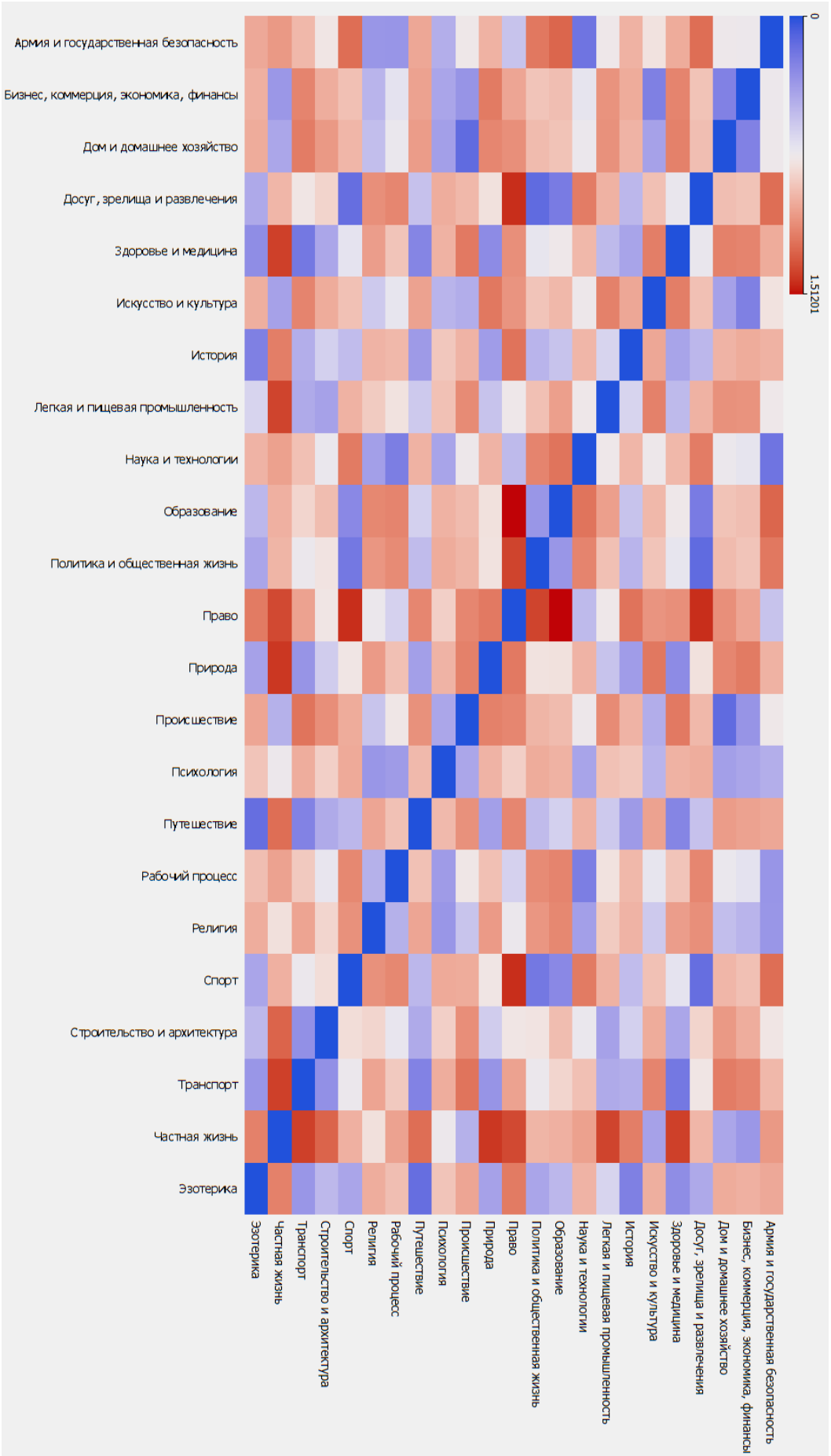

Рисунок 1. *Матрица расстояний для лингвистических профилей*

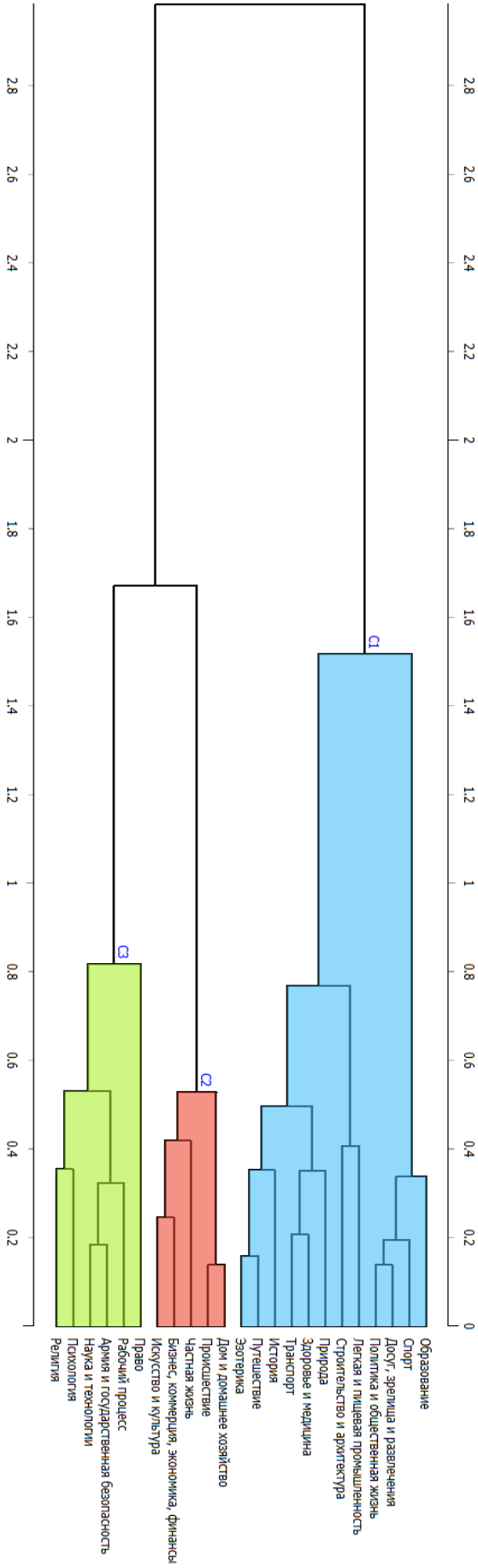

Рисунок 2. *Кластеры лингвистических профилей*

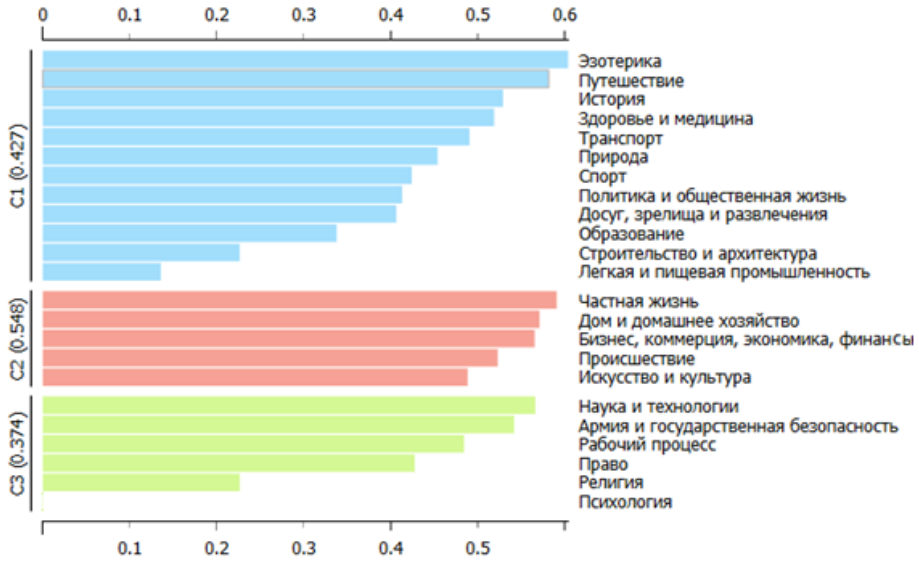

Рисунок 3. *Оценка степени принадлежности профилей к кластерам*

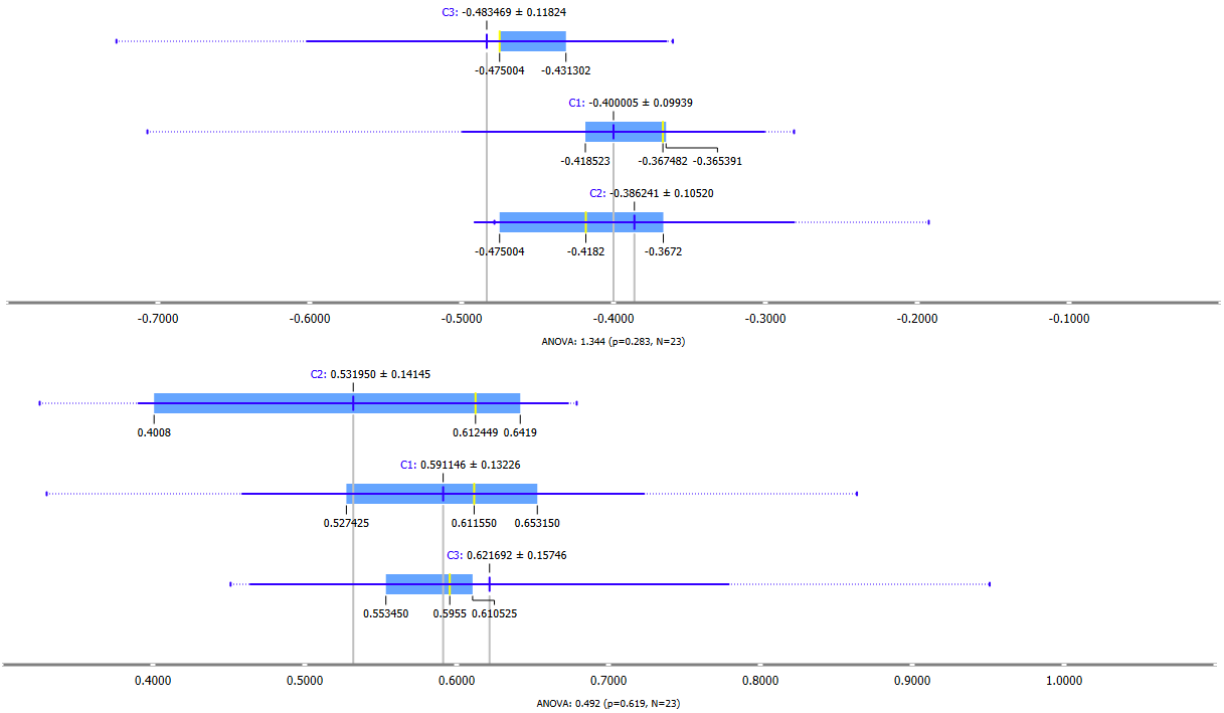

Рисунок 4. *Диаграмма размаха для корреляций «имена существительные – глаголы» и «глаголы – наречия»*

Подтверждение этих результатов находит отражение в исследовательском корпусе. Например, при описании морфологических характеристик нужно отметить, что было выявлено более 50% значимых отрицательных коррелятов, указывающих на употребление имен существительных и глаголов. В предыдущем исследовании (Мамаев, 2024) мы отмечали, что данная особенность связана со структурной организацией публикацией социальных сетей. Наряду с этим необходимо указать, что основой многих текстов социальной сети являются разноплановые ситуации, показанные через некоторый сюжет, благодаря чему история преобразуется в нарратив, который сам по себе характеризуется предикативностью. Наконец, на количество употреблений глаголов влияют и формализмы, заложенные в инструменты автоматической обработки естественного языка: многие алгоритмы при категоризации девербативов в качестве части речи указывают глагол. Самое большое значение рассматриваемой корреляции наблюдается в тематическом скрытом сообществе «Право»: «*...В Гражданском процессе судья не* ***может заставить*** *сторону* ***ответить*** *каким-то определенным образом, на* ***поставленный*** *вопрос оппонента! То есть вы можете* ***отвечать*** *на вопрос оппонента все что угодно. А судья вынуждена будет это* ***слушать*** *и будет* ***злиться*** *в таких случаях только на* ***задающего*** *вопросы, а не на* ***отвечающего****...*» (публикация пользователя 8436). Эта характерная особенность для многих сообществ, в этой связи на уровне морфологии кластеры не сильно отличаются между собой.

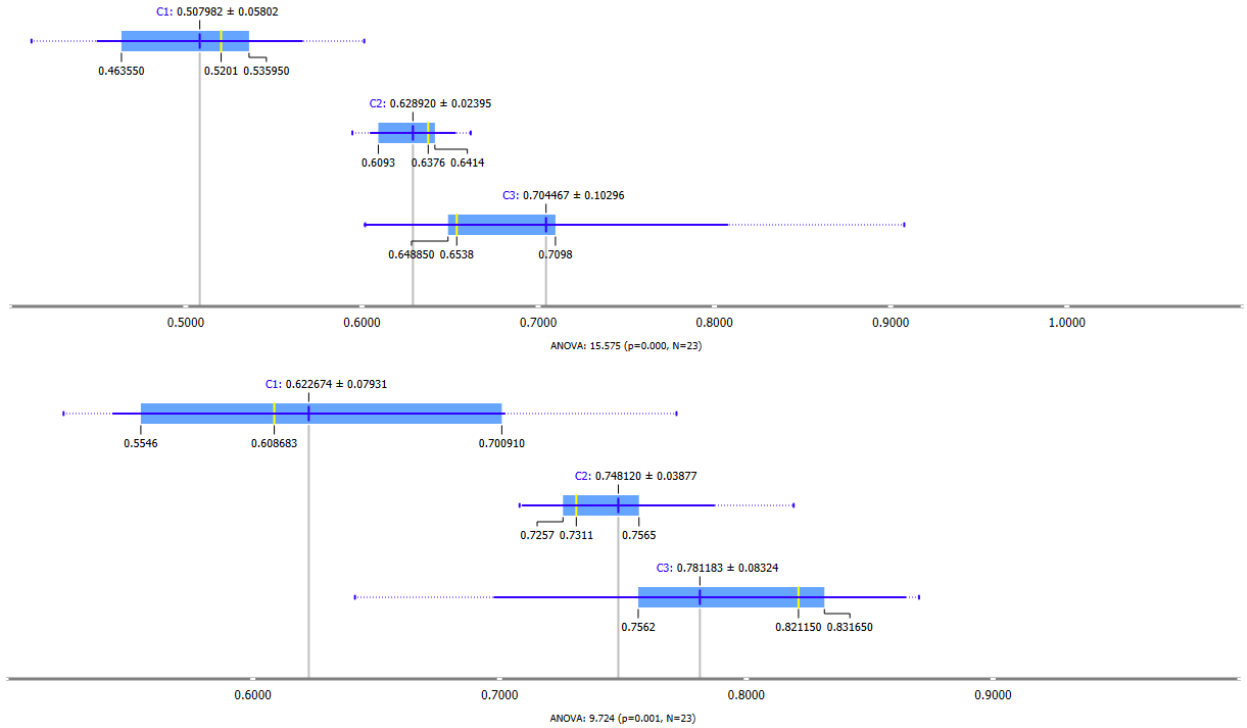

Рисунок 5. *Диаграмма размаха для синтаксических корреляций*

Для сообществ «Происшествие» и «Искусство и культура» на основании Таблицы 1 выводятся близкие количественные значения для синтаксических параметров. Для коррелятов «длина предложения-степень дистантизации» (см. Рисунок 5, верхняя диаграмма размаха), выявленных в этих двух тематических сообществах, длина предложения увеличивается за счёт расширения зависимых слов именных и/или предложных групп дополнениями, определениями или их комбинациями.

1. *«…В Москве суд отправил под **домашний арест** Антона Астахова – сына **бывшего уполномоченного** по правам ребенка в РФ Павла Астахова…»* (скрытое сообщество «Происшествие», публикация пользователя 57170).

2. *«Арнольд Шварценеггер и Дэнни ДеВито снова сыграют братьев в продолжении комедии “Близнецы”. Сиквел, съёмки которого начнутся в январе 2022 года, будет называться “Тройняшки”. Роль **третьего брата** исполнит актёр и комик Трейси Морган. Съёмками займётся Айван Райтман, снявший оригинальный фильм…»* (скрытое сообщество «Искусство и культура», публикация пользователя 6086).

В пользовательских публикациях первого кластера в основном отсутствуют распространители рассматриваемых конструкций: *«А сегодня я не сплю в 2 часа **ночи**, потому что у **кое-кого** отит и температура 39+ …Подскажите, кого можно вызвать, чтобы сдать кровь на **дому**, не ожидая полдня?!»* (скрытое сообщество «Здоровье и медицина», публикация пользователя 46028).

Обратим внимание на то, что внедрение лексической корреляции, указывающей на связь лексического разнообразия и лексической плотности текстов тематических групп скрытых сообществ, не повлияло на принадлежность сообществ к определенному кластеру. Во-первых, значимые коэффициенты были выявлены лишь для четырех сообществ – «Спорт», «Здоровье и медицина», «Религия» и «Образование». Во-вторых, лексическая плотность для трех сообществ проявляет низкую корреляцию ($k \leqslant 0.3$) с лексической насыщенностью. Лексическая плотность почти не зависит от употребляемых лексических единиц. В исследовании (Стрельников, Воробьева, 2022) на материале корпуса выпускных квалификационных работ получены похожие результаты, что позволяет сделать вывод о нечувствительности исследуемой корреляции к жанровой принадлежности корпуса.

Заключение

Проведенное исследование позволяет сформулировать ряд выводов. Из 23 лингвистических профилей скрытых сетевых сообществ ни одно не было представлено всеми семью типами корреляций. Наблюдаемое число коррелирующих параметров принимает значения в промежутке от 1 до 6, что может быть связано с объемом данных в анализируемой выборке. Лингвистическое профилирование было проведено с целью создания функциональной модели, отражающей текущие языковые тенденции в текстах социальных сетей, объединенных общей темой. Пользовательские публикации, объединенные общей темой, наиболее полно охарактеризованы с точки зрения морфологии и синтаксиса, а значимые лексические корреляции представлены в небольшом количестве профилей. Итоговая модель может найти практическое применение

при разработке систем автоматической модерации групп. Также необходимо отметить, что лингвистические профили могут использоваться для отслеживания пользовательских тенденций. На этом основании рекламные группы смогут видоизменять синтаксические конструкции своих постов для привлечения большего количества клиентов. Таким образом, все задачи работы решены, а цель достигнута.

Отметим, что эта серия работ может быть продолжена в нескольких направлениях: возможно внедрение алгоритмов динамического тематического моделирования для анализа языковых тенденций и проведение экспериментов по созданию лингвистических профилей на материале других русскоязычных социальных сетей.

Источники | References

1. Белоусов Р. Л., Дрожжин Н. А., Костенчук М. И. Построение нечетких лингвистических переменных с использованием методов кластерного анализа данных // Прикладная информатика. 2015. № 1 (55).
2. Булыга Ф. С., Курейчик В. М. Алгоритмы агломеративной кластеризации применительно к задачам анализа лингвистической экспертной информации // Известия Южного федерального университета. Технические науки. 2021. № 6 (223).
3. Крылова М. Н. Язык современного интернет-общения (на материале интеллектуального контента социальной сети «ВКонтакте») // Актуальные проблемы филологии и педагогической лингвистики. 2019. № 1.
4. Литвинова Т. А., Громова А. В. Компьютерные технологии в судебной автороведческой экспертизе: проблемы и перспективы использования // Вестник Волгоградского государственного университета. Серия 2: Языкознание. 2020. Т. 19. № 1.
5. Литвинова Т. А., Котлярова Е. С., Заварзина В. А. Фактор гендера в ассоциативных связях слов: данные словаря и дистрибутивно-семантической модели // Научный диалог. 2022. Т. 11. № 5.
6. Мамаев И. Д. Лингвистические профили скрытых сообществ: морфосинтаксический аспект // Филологические науки. Вопросы теории и практики. 2024. Т. 17. Вып. 4.
7. Мамаев И. Д., Митрофанова О. А. Лингвистические параметры для идентификации скрытых сетевых сообществ // Terra Linguistica. 2024. Т. 15. № 1.
8. Мамина Т. М. Принципиальная многозначность информации // Вестник Санкт-Петербургского университета. Социология. 2014. № 2.
9. Масликова О. С. Языковые особенности общения в интернет-пространстве // Инновационная наука. 2019. № 9.
10. Нокель М. А., Лукашевич Н. В. Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами // Вычислительные методы и программирование. 2015. Т. 16.
11. Прокофьева Е. В., Прокофьева О. Ю. Сравнительный обзор идентификационных возможностей кластерного, корреляционного и структурно-лингвистического анализа в распознавании образов // Судебная экспертиза. 2013. № 4.
12. Савотченко С. Е., Проскурина Е. А. Корреляционный и дисперсионный анализ лингвистических особенностей поиска в Интернете // Среднее профессиональное образование. 2012. № 12.
13. Сковородников А. П. О предмете эколингвистики применительно к состоянию современного русского языка // Экология языка и коммуникативная практика. 2013. № 1.
14. Степаненко А. А. Гендерная атрибуция текстов компьютерной коммуникации: статистический анализ использования местоимений // Вестник Томского государственного университета. 2017. № 415.
15. Стрельников А. И., Воробьева М. С. Исследование методов анализа информационной и лексической насыщенности научных текстов // Математическое и информационное моделирование: материалы всероссийской конференции молодых ученых (г. Тюмень, 18-23 мая 2022 г.) / Министерство науки и высшего образования РФ; Тюменский государственный университет; Институт математики и компьютерных наук; ред. колл.: Е. П. Вдовин и др. Тюмень: ТюмГУ-Press, 2022. Вып. 20.
16. Тулиев У. Ю. Кластерный анализ текстовых документов по отношению их связности // Проблемы вычислительной и прикладной математики. 2019. № 6.
17. Тюленева В. Н. Принципы адаптации заимствованной лексики в русском и китайском языках (на примере интернет-обзоров электронной техники) // Педагогическое образование в России. 2016. № 11.
18. Brunato D., Cimino A., Dell'Orletta F., Venturi G., Montemagni S. Profiling-UD: A tool for linguistic profiling of texts // Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, 2020.
19. Chakraborty I., Kim M., Sudhir K. Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes // Journal of Marketing Research. 2022. Vol. 59. Iss. 3.
20. Crystal D. Language and the Internet. Cambridge: Cambridge University Press, 2001.
21. Demšar J., Zupan B. Orange: Data mining fruitful and fun-a historical perspective // Informatica. 2013. Vol. 37. Iss. 1.
22. Kekez M. Model-based imputation of sound level data at thoroughfare using computational intelligence // Open Engineering. 2021. Vol. 11. Iss. 1.
23. Litvinova T., Litvinova O., Panicheva P. Authorship attribution of Russian forum posts with different types of n-gram features // Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval. N. Y., 2019.

Информация об авторах | Author information

RU

Мамаев Иван Дмитриевич[1]
[1] Балтийский государственный технический университет «Военмех» имени Д. Ф. Устинова; Санкт-Петербургский государственный университет, г. Санкт-Петербург

EN

Ivan Dmitrievich Mamaev[1]
[1] Baltic State Technical University “Voenmeh” named after D. F. Ustinov; Saint Petersburg State University, St. Petersburg

[1] *mamaev_id@voenmeh.ru; i.mamaev@spbu.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 04.05.2024; опубликовано online (published online): 30.05.2024.

Ключевые слова (keywords): кластерный анализ; скрытые сообщества социальных сетей; лингвистическое профилирование; морфосинтаксические характеристики постов; cluster analysis; hidden communities of social networks; linguistic profiling; morphosyntactic characteristics of posts.