ИЗДАТЕЛЬСТВО грамота

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

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

2024. Том 17. Выпуск 4 | 2024. Volume 17. Issue 4
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Лингвистические профили скрытых сообществ: морфосинтаксический аспект

Мамаев И. Д.

Аннотация. Цель исследования – выявление количественных закономерностей функционирования морфосинтаксических параметров текстов пользователей скрытых сетевых сообществ. В статье предпринята попытка подтвердить статистическими методами «спаянность» основных морфосинтаксических признаков, информация о которых была получена в лингвистическом процессоре Profiling-UD. Научная новизна исследования состоит в том, что на материале русскоязычного корпуса текстов социальных сетей проводится эксперимент по корреляционному анализу морфосинтаксических характеристик, которые могут стать частью будущего лингвистического профиля скрытых сообществ. Подобные профили могут использоваться в современных социальных сетях для улучшения функционала рекомендательных систем. В результате исследования было установлено, что для более 55% скрытых сообществ выделены значимые положительные корреляции со средней силой статистической значимости. Применяя предложенную методику, в дальнейшем лингвистический профиль скрытых сообществ можно расширить синтаксическими и лексическими параметрами, что позволит провести кластерный анализ сообществ и выявить гомогенность/гетерогенность использования характеристик различных языковых уровней в постах пользователей скрытых сообществ.

EN

Linguistic profiles of hidden communities: A morphosyntactic aspect

I. D. Mamaev

Abstract. The aim of the research is to identify quantitative regularities in the functioning of morphosyntactic parameters in the texts by users of hidden online communities. Through statistical methods, the paper attempts to confirm the “cohesion” of the main morphosyntactic features, the information about which was obtained using the Profiling-UD linguistic processor. The scientific novelty of the research lies in the following: based on a corpus of Russian-language social media texts, an experiment is conducted on the correlation analysis of morphosyntactic characteristics, which could become part of the future linguistic profile of hidden communities. Such profiles could be used in modern social media to enhance the functionality of recommendation systems. As a result, the research found that significant positive correlations with moderate statistical significance were identified for over 55% of hidden communities. By applying the proposed methodology, the linguistic profile of hidden communities can be further expanded with syntactic and lexical parameters, allowing for cluster analysis of communities and identification of the homogeneity/heterogeneity of the use of the characteristics across different linguistic levels in user posts from hidden communities.

Введение

Возможность измерения статистических взаимосвязей между различными элементами речи заложена в самой природе этих элементов. Этот вопрос не раз затрагивался в лингвистических трудах. Например, А. А. Потебня в работе «Из записок по русской грамматике» отмечал, что «существительные и прилагательные в тесном смысле… будучи близки к глаголу, еще более близки между собою» (1958, с. 92). Применение количественных методов в лингвистике не только позволяет глубже проникнуть в структуру и функции языка, но также получить объективные знания о нем. Количественные методы широко используются при категоризации текстов, в том числе для решения задач идентификации авторства и атрибуции текстов, для определения специфических черт научных и художественных текстов и др.

Популярность количественных методов сохраняется и сегодня, в том числе и при разработке специализированных текстовых коллекций. Настоящая статья является логическим продолжением исследования (Мамаев, Митрофанова, 2024), в котором описан процесс создания корпуса социальной сети ВКонтакте с целью

Научная статья (original research article) | https://doi.org/10.30853/phil20240168

построения модели скрытых сообществ методами семантической компрессии текстов. Собранная в предыдущей работе текстовая коллекция объемом более 10000 постов, опубликованных не ранее 01.01.2021, и является материалом исследования.

Актуальность нашего исследования обусловлена необходимостью создания критериев отбора внутритекстовых параметров для описания моделей скрытых сетевых сообществ. Современные алгоритмы обнаружения скрытых сообществ основаны на графово-математических или кластерных архитектурах (Baumes, Goldberg, Magdon-Ismail et al., 2004; Mishra, Schreiber, Stanton et al., 2007), а описание итоговых моделей сообществ сводится к детализации формальных характеристик (например, коэффициент кластеризации графов, плотность структур, средний путь графа и пр.), которые не учитывают лингвистические параметры авторов текстов – участников скрытых сообществ, что приводит к невозможности полноценного интегрирования предложенных процедур в современные приложения для онлайн-коммуникации.

Для реализации цели исследования необходимо решить ряд задач:

1. Описать процесс изучения лингвистических корреляций.
2. Выбрать морфологические и синтаксические параметры для описания.
3. Выявить тенденции «спаянности» морфосинтаксических характеристик.

Теоретической базой исследования послужили труды воронежской школы лингвистики (Litvinova, Sboev, Panicheva, 2018; Panicheva, Litvinova, 2019), в рамках которой предлагается методология лингвистического профилирования, а также работы (Бодрова, Тукмакова, 2012; Тукмакова, 2020; Hengeveld, 2007), посвященные вероятностно-структурной модели языка. Отметим, что в рамках настоящей статьи исследование социопсихологических характеристик представляется невозможным по ряду причин. Во-первых, для выявления психологических параметров требуется провести специализированные тесты с реальными личностями. Необходимо учитывать, что за цифровым образом пользователя может скрываться совершенно другой человек, а сам пользователь может отказаться от выполнения теста. Таким образом, полученные данные могут быть недостоверными. Во-вторых, не все пользователи предоставляют информацию о возрасте, образовании и других параметрах. Поэтому в настоящем исследовании мы выявляем статистические внутритекстовые связи. Мы вводим понятие лингвистического профиля участников скрытого сообщества – набора языковых характеристик, которые выявляются на основе текстов, созданных участниками сообщества. Необходимо отметить, что, в отличие от смежных понятий идиолекта и идиостиля, которые используются по отношению к реальному автору (Корниенко, 2019), понятие лингвистического профиля применяется к цифровому образу.

Мы представим морфосинтаксические профили, а именно: корреляции имен существительных и глаголов с их модификаторами (именами прилагательными и наречиями соответственно), корреляции длины предложения и количества предложных конструкций, а также корреляции длины предложения и длины связей зависимостей. Подобные компоненты вероятностно-структурной модели языка позволяют расширить статическую языковую картину мира. Например, сильная положительная корреляция между именами существительными и именами прилагательными позволяет передать не только понятийное содержание объекта, но и дополнительную информацию о его характеристиках, а сложные структуры зависимостей и длинные предложения могут затруднить понимание тематического поста, что может привести к коммуникативной неудаче при обсуждении текста.

Методы исследования обусловлены целью и совокупностью поставленных задач. Во-первых, метод комбинаторно-статистических расчетов помогает вычислить лингвистические корреляции между морфосинтаксическими параметрами. Во-вторых, метод контекстуального анализа позволяет дать подробный лингвистический комментарий для полученных количественных данных. Наконец, сравнительно-сопоставительный метод способствует выявлению сходств и различий полученных лингвистических профилей.

Практическая значимость исследования заключается в том, что предложенную методику можно внедрить в русскоязычный сегмент современных социальных сетей с целью создания системы модерации пользовательских групп, которая учитывает их предпочтения по ряду лингвистических параметров постов. Полученные данные также могут найти применение в процессе учебно-методической деятельности при создании курсов по компьютерной лингвистике.

Обсуждение и результаты

В качестве инструмента для сбора количественных показателей мы обратились к Profiling-UD (Brunato, Cimino, Dell'Orletta et al., 2020), поскольку это приложение позволяет получить данные о текстах без внедрения программных скриптов. Автоматическая обработка текстов происходит на самом сайте с применением формализмов фреймворка Universal Dependencies. При этом необходимо принимать во внимание тот факт, что предлагаемая морфосинтаксическая разметка обладает особенностями, которые нужно учитывать. Так, все глаголы и вербоиды будут иметь единый тег VERB. Подобное «объединение» наблюдается и в «Русской грамматике» (1980), при этом такой подход имеет доводы «за». Во-первых, все указанные формы, за исключением инфинитива, обладают общими грамматическими категориями: видом, залогом и временем, они являются исключительно глагольными. Во-вторых, у них совпадают синтаксические свойства валентности, т. е. они объединяются, если их рассматривать как главные слова в подчинительных связях. Тем не менее среди доводов «против» подобного подхода можно указать, что синтаксическая роль причастий – зачастую атрибутивная, т. е. их надо поместить в группу прилагательных.

Статистическая процедура состоит из нескольких шагов.

1. Вначале необходимо определить, принадлежат ли количественные параметры нашей выборки нормальному распределению. Для этого воспользуемся критерием – модифицированным тестом Колмогорова-Смирнова (Lilliefors, 1967), который автоматически рассчитывается в калькуляторе Statistics Kingdom (Рисунок 1).

Parameter	Value
P-value	0.00007904
D	0.274
Sample size (n)	66
Average (x̄)	198456993.1515
Median	894938
Sample Standard Deviation (S)	294762548.2053
Sum of Squares	5647522388589830000
K	2.2264
Skewness	1.3936
Skewness Shape	**Asymmetrical**, right/positive (pval=0)
Excess kurtosis	0.7076
Kurtosis Shape	Potentially **Mesokurtic**, normal like tails (pval=0.224)
Outliers	890086774, 999136815, 960009158, 879253006

Рисунок 1. *Пример автоматического определения ненормальности распределения количественных данных об употреблении имен прилагательных в скрытом сообществе «Политика и общественная жизнь»*

2. Для расчета силы корреляции при нормальном распределении используется коэффициент корреляции Пирсона.

3. Для расчета силы корреляции при ненормальном распределении используется коэффициент ранговой корреляции Спирмена.

Мы рассчитали корреляции для 23 скрытых сообществ из 34, так как в некоторых скрытых сообществах было представлено менее четырех пользователей: «Журналистика», «Астрономия», «Биология», «География», «Информатика», «Техника», «Филология», «Философия», «Машиностроение», «Производство», «Социология». Если в ходе расчетов морфологических корреляций уровень значимости $p>0.05$, то корреляция считалась незначимой, а вместо значения ставился прочерк.

Рассмотрим некоторые примеры морфосинтаксических профилей скрытых сообществ. Так, в сообществе «Бизнес, коммерция, экономика, финансы», включающем 52 пользователя, для имен существительных и глаголов наблюдается обратная взаимосвязь средней силы (Рисунок 2).

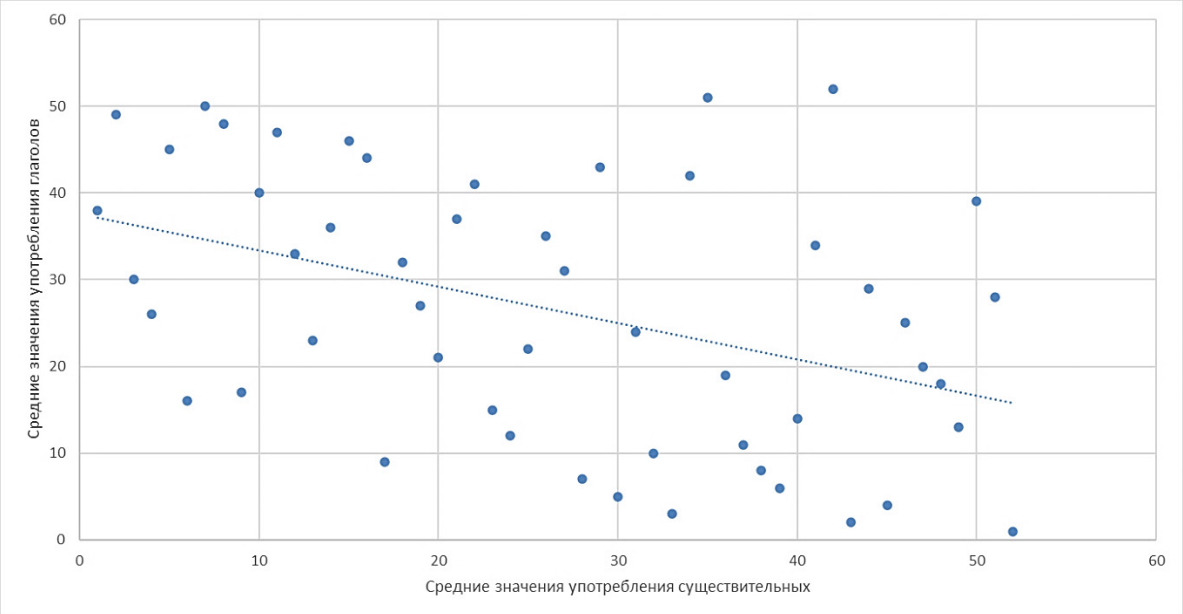

Рисунок 2. *Внутритекстовая корреляция средних значений употреблений имен существительных и глаголов*

Рост употреблений имен существительных при одновременном снижении числа употреблений глаголов связан со способом построения постов, которые пользователь описывает в номинативных конструкциях, например о предстоящих мероприятиях (см. посты пользователей с ID 13977 и 20518 соответственно). В первом отрывке с помощью существительных в рамках составного именного сказуемого перечисляются все профессии и заслуги личности, а во втором – в виде маркированного списка перечисляются идеи, которые предлагает пользователь. Несмотря на использование двух глаголов во втором отрывке («даю», «впишется»), количество имен существительных выше. Подобные колебания в употреблении имен существительных и глаголов указывают на разницу не только в структуре текстов, но и в типе клавиатурно-опосредованного высказывания. Посты с преобладающим количеством имен существительных тяготеют к информационным постам, в то время как преобладание глаголов указывает на принадлежность к постам-повествованиям.

(1) ID 13977: «*…15.50-16.10 “ЛИЧНЫЙ БРЕНД В ПОМОГАЮЩИХ ПРОФЕССИЯХ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ”*

*Солдатова Светлана (Симферополь) – **бизнес-консультант**, **коуч**, управляющий **партнёр** консалтинговой компании “S&D Group”, действительный **член** ППЛ…*».

(2) ID 20518: «*…Вот **парочка** идей, я <u>даю</u> очень много пользы и возможностей безвозмездно!*

- *Дополнительный **доход**, высокая маржинальность.*
- *Отлично <u>впишется</u> в интерьер празднично лофта, ресторана или островка в ТЦ.*
- *Дополнительная **фотозона** при проведении праздников.*
- *Дополнительная развлекательная **программа** для детей.*
- *Дополнительная **витрина** для размещения товаров (игрушки, воздушные шарики, сладости)…*».

На синтаксическом уровне выявлена сильная положительная корреляция среднего количества используемых предложных конструкций и длины предложения (Рисунок 3), что также подтверждается рядом проведенных ранее исследований. Так, согласно (Конюшкевич, 2013), корреляция предикативной части сложного предложения и соответствующей именной синтаксемы с предлогом обладает сильной связью.

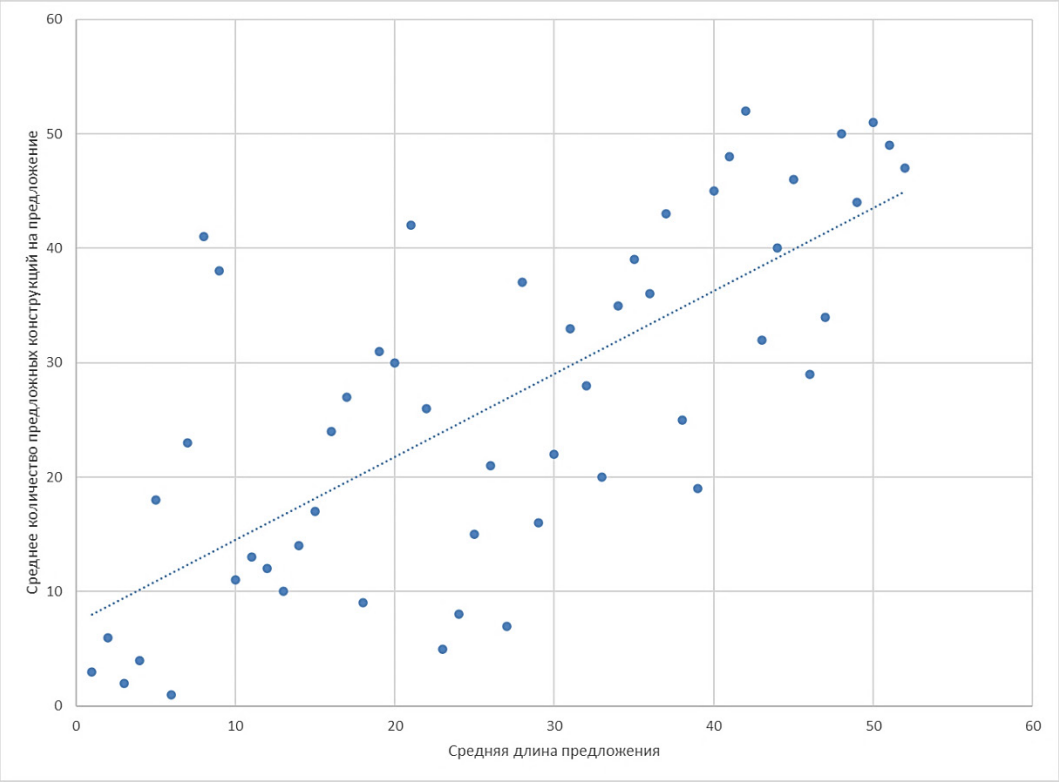

Рисунок 3. *Внутритекстовая корреляция средних длин предложений и среднего количества предложных конструкций*

Отметим, что на эту корреляцию может повлиять и длина предложных конструкций, что подтверждено исследованиями разножанровых русскоязычных и англоязычных корпусов (Хохлова, Рубинер, 2019; Curtotti, McCreath, 2011). В частности, на длину некоторых предложных конструкций повлияло использование нескольких атрибутивов перед зависимым именем существительным. Отметим, что синтаксические конструкции «имя прилагательное + имя существительное» приводят не только к усложнению предложных групп, но и к образованию устойчивых сочетаний, которые активно воспроизводятся в речи участников скрытых сообществ.

1. Предложная конструкция длиной в шесть единиц (пользователь с ID 32764): «*…**В любой непонятной или спорной ситуации** он будет замирать и ждать того, кому видней. Внимание, вопрос: разве сможет такой человек быть эффективным сотрудником, особенно в 21 веке, где постоянно нужно принимать решения и перепроверять информацию?…*».

2. Предложная конструкция длиной в пять единиц (пользователь с ID 32764): «*…А потом* ***в одно не прекрасное утро*** *понимает, что “всё не то”. Ничто не изменилось, но он перестал быть счастливым…*».

Тем не менее пользователи скрытого сообщества публикуют посты, в которых содержатся сложные предложения с большим количеством коротких предложных конструкций (пользователь с ID 243192): «*…ВАЖНО: оповещайте* ***об отмене*** *сеанса* ***за 1-2 суток****, ситуации* ***у всех*** *бывают разные, поэтому сообщайте, пожалуйста,* ***обо всех изменениях*** *заранее…*».

В сообществе «Эзотерика» (17 участников) оказалось, что всего одна корреляция является значимой – корреляция длины предложения и степени дистантизации с прямой силой связи $r=0.4877$ при уровне значимости $p=0.047$ (Рисунок 4).

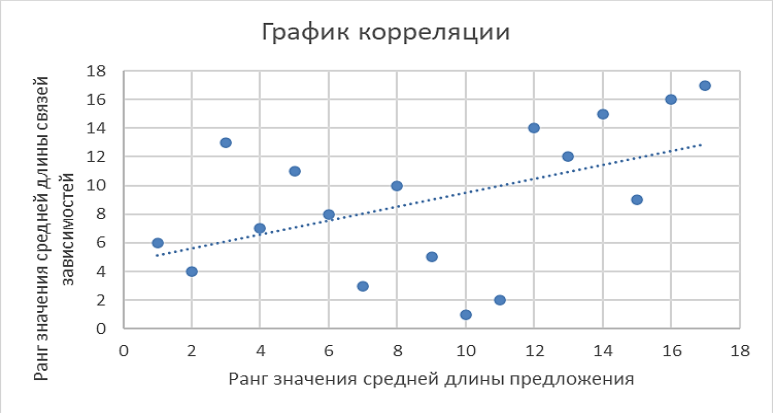

Рисунок 4. *Внутритекстовая корреляция средних значений длины предложений и длины структур зависимостей*

Согласно пособию (Мартыненко, Гребенников, 2018), в современных русскоязычных корпусах степень дистантизации относится к мерам, основанным на усложнении линейного порядка предложения (мера линейной сложности). У пользователя с ID 2644 встречаются примеры инверсионной структуры предложения вида OVS с расширенным дополнением, что приводит к увеличению расстояния между главным и зависимым узлом. На Рисунке 5 степень дистантизации между лексемой-предикатом «является» и левостоящей зависимой лексемой-дополнением «дверью» равняется трем при общей длине предложения в 10 словоформ: расширение структурной схемы конструкции N_5V_{fin} обусловлено использованием предложной группы.

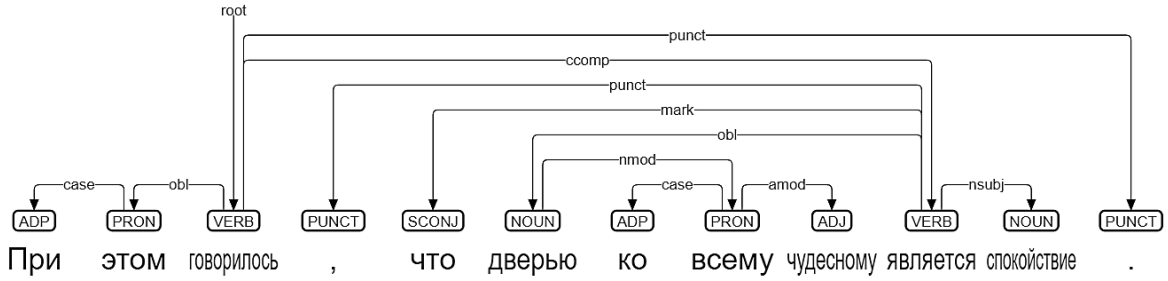

Рисунок 5. *Дерево зависимостей для предложения из поста пользователя с ID 2644*

В постах с прямым порядком слов степень дистантизации может увеличиваться при подчинении нескольких зависимых дополнений при единой глагольной вершине. В одном из постов пользователя с ID 9074 встречается предложение, в главной части которого расстояние между инфинитивом и косвенным дополнением увеличено за счет стоящего между ними прямого дополнения (Рисунок 6).

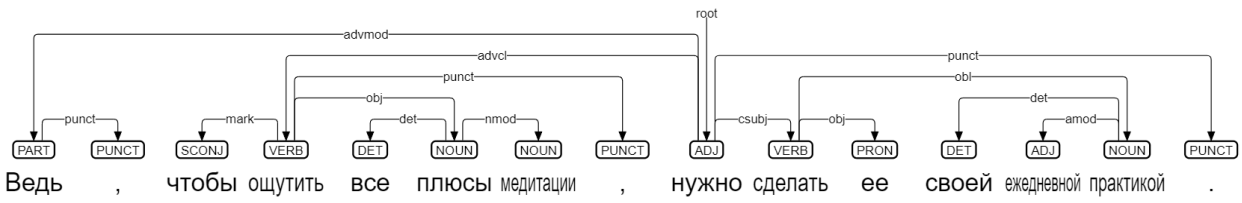

Рисунок 6. *Дерево зависимостей для предложения из поста пользователя с ID 9074*

Наконец, увеличение степени дистантизации связано с правосторонним расширением подлежащих. На Рисунке 7 степень дистантизации между подлежащим и сказуемым в главной части предложения равняется четырем при общей длине предложения в 14 словоформ.

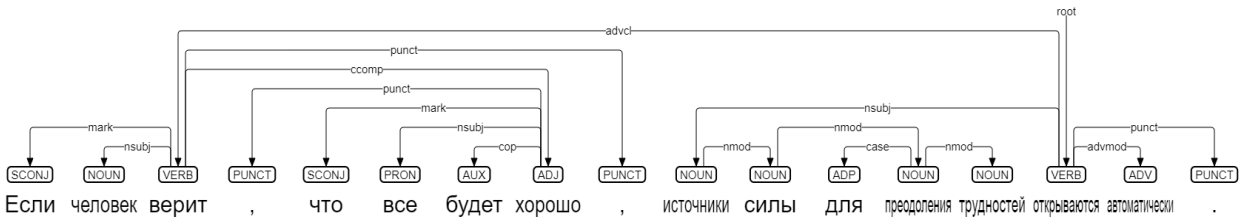

Рисунок 7. *Дерево зависимостей для предложения из поста пользователя с ID 51196*

Скрытое сообщество «Искусство и культура» представлено 115 пользователями, в нем выявлено пять значимых морфосинтаксических корреляций (Рисунок 8).

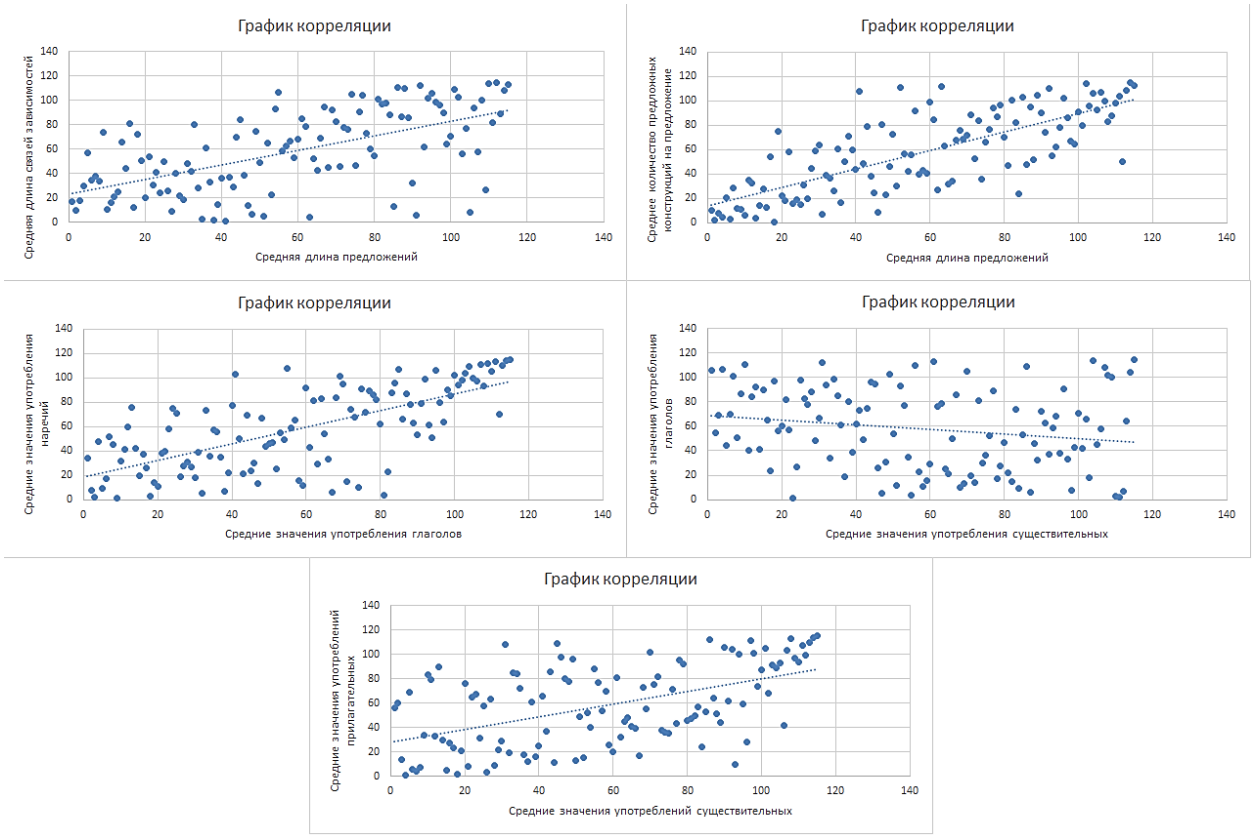

Рисунок 8. *Сводные данные о внутритекстовых корреляциях*

На первом месте по силе связи – корреляция длины предложения с количеством предложных конструкций (первый ряд, правая колонка): *r=0.7565*.

(1) Пользователь с ID 871024: «*…Vasya Ve – Fiesta the Pogues – я впервые* ***за много лет*** *услышал песню, которая написана не* ***для чартов*** *и чтобы кого-то удивить или высказаться* ***о боли****, а просто весёлый балаган,* ***в котором*** *творчество* ***ради творчества****…*» (5 предложных конструкций на 34 словоформы).

(2) Пользователь с ID 68068: «*…Александр Лециус и Кристина Карпышева* ***за последние несколько лет*** *показали свои знаменитые аудио-визуальные перформансы* ***на Таймс-сквер, в Музее искусств Лос-Анджелеса, на фестивале медиа искусств в Японии, на монреальском фестивале MUTEK*** *– Каннах* ***в области цифрового искусства и электронной музыки****…*» (7 предложных конструкций на 38 словоформ).

(3) Пользователь с ID 22242: «*…Кульминация* ***в сценарии****, должна быть зримой* ***на сцене****, она не должна быть словами* ***в тексте ведущего****…*» (3 предложные конструкции на 16 словоформ).

Положительная сила взаимосвязи также выявлена у глаголов и наречий-модификаторов (второй ряд, левая колонка): *r=0.6791*. Коэффициент *r=0.5946* установлен для корреляции длины предложения и степени дистантизации (первый ряд, левая колонка). Для предложения длиной в 18 словоформ из поста пользователя с ID 69784 степень дистантизации между главным узлом «из» и подчиненным узлом «звука» равна четырем, при этом в иерархической структуре предложения (Рисунок 9) за счет сочинительных связей между исследуемыми узлами наблюдается несколько ветвей составляющих.

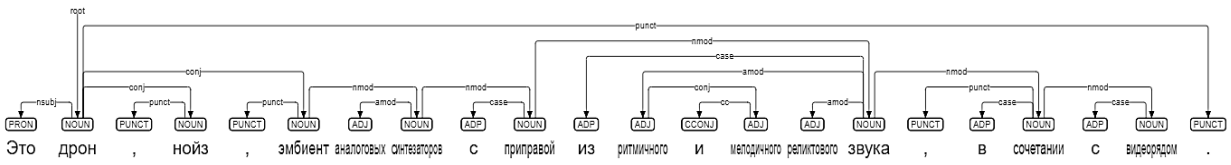

Рисунок 9. *Дерево зависимостей для предложения из поста пользователя с ID 69784*

Единственная отрицательная корреляция представлена парой имен существительных и глаголов (второй ряд, правая колонка), однако в связи со слабой силой связи (*r=-0.1924*) не наблюдается четкая обратная взаимосвязь морфологических параметров.

Так, полученные результаты свидетельствуют о вариациях стиля пользователей тематических скрытых сообществ на морфологическом и синтаксическом уровнях, а также о специфической организации постов социальных сетей.

Заключение

Таким образом, для достижения поставленной в работе цели были решены все обозначенные задачи. В рамках первой задачи описан процесс адаптации традиционного корреляционного анализа к лингвистическому материалу. В ходе решения второй задачи был выбран ряд морфосинтаксических параметров для профилирования скрытых сетевых сообществ. Наконец, были представлены значимые количественные показатели корреляций для отобранных пар лингвистических параметров и подтверждающие примеры. На данный момент можно сделать вывод, что для пользователей скрытых сообществ социальных сетей характерны следующие тенденции «спаянности». Во-первых, наблюдается тесная положительная связь имен существительных и прилагательных, при этом для имен существительных и глаголов выявлены отрицательные коэффициенты корреляции, что указывает на преобладание номинативной структуры текста над предикативной. Во-вторых, подтверждается предположение о синтаксической сложности текстов скрытого сообщества: с увеличением длины предложной конструкции и/или длин связей зависимости увеличивается и длина предложения. Подобные корреляции могут использоваться для разработки лингвистических процедур выявления ключевых влиятельных лиц в социальных сетях наравне с формальными показателями (количество комментариев, реакций и т. д.). Тексты социальных сетей с длинными синтаксическими конструкциями с большей вероятностью будут использоваться лидерами мнений и влиятельными лицами в определенной нише.

Тем не менее отметим ряд перспектив дальнейшего исследования. Так, помимо морфосинтаксических параметров в лингвистический профиль можно добавить проверку гипотезы о зависимости критериев информационной насыщенности от метрик лексической насыщенности (type-token ratio). Полученные количественные профили можно будет разбить на кластеры для того, чтобы изучить лингвистическую близость скрытых сообществ и гомогенность/гетерогенность сообществ на каждом из языковых уровней: морфологическом, синтаксическом и лексическом.

Источники | References

1. Бодрова Т., Тукмакова Н. Определение коэффициента ранговой корреляции частей речи в русских и чувашских газетных текстах // Мовознавчий вісник. 2012. № 14-15.
2. Конюшкевич М. Преобразование предложно-падежной синтаксемы в предикативную единицу: корреляция предлога и показателя связи сложного предложения // Лінгвістичні студії. 2013. № 26.
3. Корниенко Е. Р. Идиолект и идиостиль: к вопросу о соотнесении понятий // Филология: научные исследования. 2019. № 1.
4. Мамаев И. Д., Митрофанова О. А. Лингвистические параметры для идентификации скрытых сетевых сообществ // Terra Linguistica. 2024. Т. 15. № 1.
5. Мартыненко Г. Я., Гребенников А. О. Основы стилеметрии: учеб.-метод. пособие. СПб.: Изд-во С.-Петерб. ун-та, 2018.
6. Потебня А. А. Из записок по русской грамматике: в 4-х т. М.: Учпедгиз, 1958. Т. 1-2.
7. Русская грамматика / гл. ред. Н. Ю. Шведова. М.: Наука, 1980. Т. 1. Фонетика. Фонология. Ударение. Интонация. Словообразование. Морфология.
8. Тукмакова Н. П. Определение коэффициента взаимной сопряженности в русских и чувашских газетных текстах // Филологические науки. Вопросы теории и практики. 2020. Т. 13. Вып. 7.
9. Хохлова М. В., Рубинер В. И. К вопросу о количественном анализе предложно-падежных сочетаний в русском языке на примере законодательных текстов // Корпусная лингвистика – 2019: труды международной конференции. СПб., 2019.

10. Baumes J., Goldberg M., Magdon-Ismail M., Wallace W. A. Discovering hidden groups in communication networks // International Conference on Intelligence and Security Informatics. Berlin – Heidelberg: Springer Berlin Heidelberg, 2004.
11. 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.
12. Curtotti M., McCreath E. C. A corpus of Australian Contract Language: Description, profiling and analysis // Proceedings of the 13th International Conference on Artificial Intelligence and Law. 2011. http://dx.doi.org/10.2139/ssrn.2304652
13. Hengeveld K. Parts-of-speech systems and morphological types // ACLC Working Papers. 2007. Vol. 2.
14. Lilliefors H. W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown // Journal of the American Statistical Association. 1967. Vol. 62. No. 318.
15. Litvinova T., Sboev A., Panicheva P. Profiling the age of Russian bloggers // Conference on Artificial Intelligence and Natural Language. Cham: Springer International Publishing, 2018.
16. Mishra N., Schreiber R., Stanton I., Tarjan R. E. Clustering social networks // International Workshop on Algorithms and Models for the Web-Graph. Berlin – Heidelberg: Springer Berlin Heidelberg, 2007.
17. Panicheva P., Litvinova T. Authorship attribution in Russian in real-world forensics scenario // International Conference on Statistical Language and Speech Processing. Cham: Springer International Publishing, 2019.

Информация об авторах | Author information

RU

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

EN

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

[1] *mamaev_id@voenmeh.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 23.02.2024; опубликовано online (published online): 16.04.2024.

Ключевые слова (keywords): лингвистическое профилирование; корпус русскоязычных социальных сетей; морфосинтаксические характеристики постов; скрытые сообщества; linguistic profiling; corpus of Russian-language social media; morphosyntactic characteristics of posts; hidden communities.