УДК 81'27+81'37 DOI: 10.14529/ling190407

СОЦИАЛЬНАЯ СЕТЬ КАК ИСТОЧНИК СОЦИОЛИНГВИСТИЧЕСКИХ ДАННЫХ: ЛЕКСИКО-СТАТИСТИЧЕСКИЙ АНАЛИЗ

М.Ю. Мухин, А.И. Лозовская
Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, г. Екатеринбург, Россия

Статья посвящена обоснованию методики и проведению сопоставительного лексико-статистического анализа социолингвистических данных, полученных из социальной сети «ВКонтакте». Обсуждаются новые возможности исследования вербальной интернет-коммуникации, в том числе социолингвистических признаков пользователей Интернета. Основное внимание в работе уделяется соотнесению возраста пользователей социальных сетей и лексических особенностей их текстов. На основании полученных данных (около 8 млн слов) сформированы корпуса текстов четырех возрастных групп пользователей социальной сети – от 14 до 74 лет. Проведен сопоставительный анализ частотной лексики, и в первую очередь выявлены слова, часто употребляющиеся в текстах пользователей всех четырех групп. Этот массив в общих чертах соответствует данным частотного словаря русского языка, однако имеет и значимые отличия. Сформированы списки частотных слов, характерных для каждой возрастной группы. Спискам дана идеографическая характеристика (выделены доминантные тематические группы слов) и социолингвистические комментарии. Сделаны выводы о лексических и концептуальных различиях между текстами разных возрастных групп пользователей, а также о продуктивности статистического и идеографического анализа текстов социальных медиа.

Ключевые слова: социальная сеть, социолингвистика, возраст, возрастная группа, корпусная лингвистика, лексическая статистика, идеографический анализ.

Введение

Анализ социолингвистических признаков значительно изменился благодаря появлению в Интернете социальных сетей. Социальные сети сегодня становятся уникальным источником различных (в том числе лингвистических) данных, и это открывает беспрецедентные возможности для проведения междисциплинарных исследований на стыке лингвистики, статистики и социологии.

Термин «социальная сеть» исконно использовался в социологических работах, в которых под ним понимали «социальную структуру, состоящую из множества агентов (индивидуальных или коллективных) и определенного множества отношений (совокупности связей между агентами)» [5, с. 4]. С развитием Интернета возникло другое, современное понимание социальных сетей. Они стали интернет-платформами, существующими и изменяющимися в реальном времени. Социальные сети позволяют устанавливать связи между участниками, способствовать их взаимодействию и установлению диалога между ними, а также обмениваться контентом и создавать уникальный медиаконтент.

Если посмотреть на широкий исследовательский контекст, связанный с обработкой информации, извлеченной из социальных медиа, то можно выделить четыре вида анализа контента социальных сетей:

- анализ общей информации с произвольными типами данных;
- анализ текста;
- анализ мультимедиа;
- сенсорный и потоковый анализ [2, с. 14].

Основная часть этого контента (по крайней мере пока) представляет собой вербальную коммуникацию, а значит, является новым объектом лингвистического рассмотрения. Естественно, что исследование лингвиста первым делом обращено к анализу текста, а точнее – текстов, которые представляют собой посты пользователей.

На сегодняшний день выделяют ряд устойчивых направлений исследования вербальной интернет-коммуникации, и этот перечень может быть неполным:

- анализ языка Интернета с точки зрения его соответствия языковым нормам;
- изучение формирования и функционирования видов тематического дискурса (это касается различных чатов, блогов, интернет-форумов);
- анализ систем гипертекста и способов их реализации в языковом формате;
- изучение специфической организации вебстраниц как особого вида текста;
- исследование функционирования средств массовой информации в Интернете;
- обобщение социолингвистических и психологических особенностей пользователей Интернета;
- анализ сетевой художественной литературы, особенностей ее языка и стиля;
- описание компьютерной терминологии и исследование особенностей профессионального компьютерного жаргона [7, с. 11].

Исследования структуры социальных сетей и речевого поведения пользователей сегодня становятся все более популярными, а настоящая работа проведена в контексте направления, которое предполагает обобщение социолингвистических особенностей пользователей Интернета. Конкретизированный предмет связан с возрастными социолингвистическими характеристиками текстов пользователей социальной сети «ВКонтакте».

Параметр возраста является одним из ведущих социолингвистических признаков. Как пишут В.И. Беликов и Л.П. Крысин, «возраст и пол являются лингвистически значимыми биосоциальными факторами: представители разных поколений характеризуются неодинаковым использованием средств языка, разными предпочтениями в оценках языковых фактов и т. п., а различия людей по биологическому полу сказываются в их речевых склонностях и неприятиях» [3, с. 156]. В лингвистике и науках, возникших на стыке языкознания и других дисциплин (в первую очередь психо-, социо-, онтолингвистике и т. п.), накоплено большое количество сведений о различиях в языке возрастных групп, о генезисе использования языка по мере взросления человека и развития личности. В то же время сегодня мы располагаем новым по качеству и количеству материалом, который предлагают социальные сети. Его качественные особенности связаны с тем, что это подъязык электронной коммуникации, а количественные особенности определяются объемом этого материала: это пример больших данных, состоящих из огромного количества слов.

В личных профилях пользователи сетей указывают год рождения (или возраст, что фактически то же самое), и как правило эта информация является правдивой. Согласно диссертации С.В. Бондаренко, «информационное неравенство в эпоху формирования информационного общества становится одним из важнейших факторов дифференциации социальных групп». Более того, характерной чертой сетевого общения является то, что «значительная часть информации внутри сетевого сообщества становится информацией только при ее запросе, а объективность информации в значительной мере зависит от поставщиков контента и своевременности ее обновления» [4, с. 19]. Однако опыт обращения к контенту социальных сетей показывает, что возрастная дифференциация большинства пользователей не является фантомом. Мало того, существующие объединения пользователей в группы по интересам, роду занятий и т. д. часто соотносятся с возрастом членов этих сообществ. С нашей точки зрения, это делает параметр возраста системообразующим и усиливает возможности его соотнесения с пользовательским речевым контентом.

Неслучайно к этому параметру обращаются современные лингвисты и социологи. Например, работа М.К. Карповой посвящена самопрезентации пользователя в социальной сети. Согласно выводам исследовательницы, самопрезентация во многом зависит от пола и возраста пользователя [6].

Актуальное для нашей работы исследование было проведено в 2013 году группой ученых Пенсильванского университета (штат Филадельфия, США). Ими было проанализировано 700 млн слов, фраз и тематических примеров, собранных из постов 75 000 добровольцев, которые также прошли стандартные личностные тесты, что позволило обнаружить различия в языке с точки зрения личности, пола и возраста [12]. Например, согласно исследованию, женщины используют больше эмоциональных слов, их посты включают в себя выражение чувств и социальных процессов (к примеру, *люблю тебя* и эмотикон <3 – «сердце»). Мужчины используют больше нецензурных слов. Что касается возрастных характеристик, то ученые отметили различия между молодой группой пользователей, для которой характерно использование жаргона и эмотиконов *(:), idk* – т. е. *«я не знаю»*), и группой пользователей в возрасте от 23 до 29 лет, для которой, например, характерна тематика работы *(на работе, новая работа, рабочее место)*. В возрастной группе от 13 до 18 лет учеными была замечена школьная тематика *(домашнее задание, уроки)*, а для молодых людей в возрасте от 19 до 22 лет ключевыми словами оказались слова *семестр, колледж, поступление* [12, с. 9]. На наш взгляд, упущением данного исследования является отсутствие анализа лексики людей пожилого возраста или же этот пласт материала оказался статистически не значимым.

Сопоставление языковых признаков, характерных для возрастных групп, требует проведения статистического анализа: «понятно, что статистика соотношения разных в возрастном и половом отношении групп говорящих имеет существенное значение в конкретных социолингвистических исследованиях» [3, с. 156]. В то же время можно отметить, что российские социальные сети еще не становились объектом такого анализа – во всяком случае это касается сопоставительного исследования, основанного на обобщении больших данных.

Основная характеристика исследования

Рабочую гипотезу нашего исследования можно сформулировать так: исходя из деления пользователей социальной сети на возрастные группы, можно сформировать отдельные текстовые подкорпуса и на основании сопоставительного анализа частотной лексики выявить их социолингвистические особенности. Теоретически так можно определить не только языковые (лексические), но и социокультурные приоритеты различных поколений и специфику межпоколенческой коммуникации, ее удачи и неудачи.

Базовым методом исследования является сопоставительный лексико-статистический анализ, на основании которого проводится социолингви-

стическая интерпретация данных. Материалом работы являются посты пользователей социальной сети «ВКонтакте», в явном виде указавших в личном разделе аккаунта свой возраст.

Общая аналитическая модель сводится к следующим этапам:

1. Обоснование возрастных групп пользователей социальных сетей.

2. Выборка текстового материала, формирование текстовых корпусов, соответствующих выделенным возрастным группам.

3. Лемматизация (приведение слов к начальной форме) и морфологическая обработка текстов.

4. Формирование массивов частотной лексики из текстовых корпусов и ее сопоставительный анализ. Выборка слов, которые являются частотными во всех корпусах возрастных групп, и лексики, характерной для каждого из них.

5. Социолингвистическая интерпретация полученных данных, предполагающая проведение контекстологического, идеографического и стилистического анализа.

Сходные по принципам сопоставительного лексико-статистического анализа исследования представлены в работах [8, 9].

Опишем в общих чертах технические моменты получения данных из социальной сети.

1. Для сбора данных мы воспользовались публичным потоковым API (application programming interface). Поток состоит из нескольких типов событий (запись, комментарий, репост, закрепление поста на странице) и действия (создание, редактирование, удаление). Нами были выбраны только личные записи пользователей, размещенные на страницах, т. е. мы не учитывали репосты и комментарии.

2. Сам пост содержит в себе множество полей – к примеру, имя пользователя, фамилию и прочие атрибуты, но для извлечения данных необходим только ID (идентификационный номер) автора сообщения, по которому можно получить информацию о пользователе (текст записи, пол и возраст, если он указан).

3. На основании идентификационного номера дополнительная программа запрашивает информацию об авторе поста, используя метод API ВКонтакте users.get. После этого записи отправляются на вывод программы, где сохраняются в файл.

Необходимо отметить определенные недостатки выборки данных из социальных сетей. Веб-интерфейсы социальных сетей не всецело оптимизированы для автоматического анализа по следующему ряду причин:

1) приватность данных – пользователи социальных сетей могут закрывать доступ к своим данным;

2) социальные сети содержат «нечистые» данные, которые могут включать в себя рекламу, неинформативные сообщения, состоящие исключительно из хэштегов или эмодзи;

3) невозможность проверить правдивость данных: пользователи могут неверно указывать свой возраст, город проживания и прочие социальные атрибуты [1, с. 441].

Итак, на начало сентября 2019 года нами всего получено 643 433 записи пользователей. С каждой записью связаны параметры пола автора и его точного возраста.

Для социолингвистического исследования данные разделены на следующие возрастные группы:

группа 1: 14–18 лет (1 033 878 слов);
группа 2: 19–35 лет (4 171 035 слов);
группа 3: 36–50 лет (2 173 848 слов);
группа 4: 51–74 года (1 170 089 слов).

Таким образом, общий объем материала на сегодняшний день превысил 8,5 млн слов.

Разделение именно на такие возрастные группы основывается на социо-культурологической теории «аналогового» и «цифрового» поколений – см. об этом, например, в работе А.С. Сумской и А.И. Лозовской [13]. Группы 1 (в большей степени) и 2 составляют представители исключительно цифрового поколения, эта аудитория социальной сети моложе 36 лет. Группа 3 относится к так называемому эхо-поколению – являющемуся промежуточным, переходным от «цифрового» поколения к «аналоговому». Последняя, четвертая группа относится к «аналоговому» поколению, которое привыкло к традиционным средствам массовой информации и значительной частью не является активным пользователем Интернета.

Выгруженные записи составили базу четырех текстовых корпусов возрастных групп. Был проведен морфологический анализ, и из каждого корпуса извлечено по 2000 самых частотных слов, которые сопоставлены. Выявлен массив слов, которые часто встречаются во всех корпусах. Приведем верхнюю часть этого списка (60 самых частотных слов):

Быть, цена, человек, размер, день, год, новый, мочь, жизнь, друг, получать, хотеть, любить, время, хороший, очень, уже, можно, руб., еще, работа, много, становиться, давать, игра, знать, мир, качество, первый, сделать, делать, заказ, писать, кожа, приходить, место, ребенок, рубль, наличие, раз, сегодня, сказать, говорить, слово, идти, любовь, хорошо, вопрос, уровень, начинать, сейчас, понимать, работать, группа, жить, помогать, большой, дом, деньги, ждать.

Этот перечень слов соотносится с данными Нового частотного словаря русской лексики [11]. В целом в него вошли слова, являющиеся частотными в русскоязычных текстах. Однако в списке заметны слова типа *цена* (ранг № 2!), *руб.* и *рубль* (сокращения не приводились нами к полным формам), *размер*, *кожа*, *качество* и другие, которые в частотном словаре (т. е. по данным Национального корпуса русского языка [10]) занимают совсем другие частотные позиции. Здесь, безус-

ловно, проявляется лексическая специфика социальных сетей, однако более детальная интерпретация этого ряда может быть предметом отдельной статьи.

Теперь рассмотрим лексику, ставшую оригинальной для выделенных возрастных групп пользователей. Четыре списка слов были сформированы на основании сопоставления, т. е. были выбраны слова, которые часто встречаются в текстах каждой группы и одновременно не характерные для других (в разбросе 150–400 для каждой группы). Приведем примеры таких слов (табл. 1).

Напомним, что в этих столбцах приведены далеко не все лексемы, а только некоторые, наиболее частотные. Первоначальный взгляд на материал позволяет заметить следующее. Для первой возрастной группы характерны слова, связанные с времяпрепровождением подростков: здесь наблюдаются и увлечения *(аниме, Винкс* – популярный мультсериал), и тематика социальных сетей *(коммент, проголосовать)*, частотным словом оказалось и слово *учеба*, что первым делом характерно только для этой возрастной группы (очевидно, студенты этого слова избегают). Стоит отметить, что для данной возрастной группы более чем для других характерны слова-эмотивы *(обижать, грустный, полюбить, скучно)*.

Многие частотные слова, которые вошли во вторую группу, связаны с реакцией на рекламу, которая присутствует в большом объеме в социальной сети. Избыточность рекламы можно обосновать тем, что аудитория с 19 до 35 лет является целевой аудиторией, которая готова покупать одежду *(рубашка, куртка, блузка)*, парфюмерию и косметику *(духи, пудра)*, а также активно заботится о своем здоровье и внешнем виде *(целлюлит, питать, шугаринг)*. Именно здесь в число частотных попадает слово *секс*.

Для возрастной группы 36–50 лет характерна тематика работы и получения новых квалификаций *(диплом, вебинар, программирование)*, начинает актуализироваться тематика здоровья *(иммунитет, медицина, травма)*, встречается и лексика, связанная с отдыхом, развлечениями *(бассейн, дача)*, и, очевидно, специфичный для этого возраста *биткоин*.

Рассуждая о текстах последней возрастной группы, стоит отметить, что многие записи пользователей содержат стихи, различные цитаты – в том числе из Библии и Корана (т. е. не тексты самих авторов постов). Достаточно много здесь рассуждений на политическую тематику, что определяет появление слов *Сталин*, *столица*, *Кремль* и др. Кроме того, весьма значительно представлена

Таблица 1

1 (14–18 лет)	2 (19–35 лет)	3 (36–50 лет)	4 (51–74 года)
Винкс	Парфюмерия	Отчет	Депозит
Поклонник	Туалетный	Получатель	Убеждение
Продвигать	Композиция	Программирование	Русь
Учеба	Рукав	Диплом	Эмоциональный
Копировать	Цветочный	Фишка	Столица
Спускаться	Лайт	Рабочий	Консультант
Стройка	Питать	Ипотека	Дервиш
Тупой	Поставщик	Уборка	Сталин
Братишкин	Шорты	Биржа	Солдат
Перевозка	Духи	Бассейн	Божественный
Проголосовать	Дисконт	Предприятие	Крест
Выставлять	Целлюлит	Вебинар	Религия
Курить	Штаны	Системный	Инвестировать
Ловушка	Резинка	Выезд	Цивилизация
Скучно	Шампунь	Ограничение	Молиться
Полюбить	Жидкость	Стрижка	Политический
Грустный	Брюки	Полиция	Ментальный
Сдача	Кофта	Снг	Царь
Мститель	Рубашка	Федеральный	Кремль
Пацан	Куртка	Владелец	Внук
Кончаться	Шугаринг	Рак	Доктор
Плевать	Фабричный	Расход	Христов
Возрождение	Блузка	Кафе	Благословение
Туризм	Пояс	Биткоин	Старик
Аниме	Знакомство	Травма	Глядеть
Магический	Кроссовки	Дача	Виртуальный
Коммент	Комфорт	Плата	Хозяйство
Познакомить	Пудра	Медицина	Господи
Валентинка	Ресничка	Вероятность	Директор
Обижать	Секс	Иммунитет	Заявление

Таблица 2

1 (14–18 лет)	2 (19–35 лет)	3 (36–50 лет)	4 (51–74 года)
Социальные сети	Одежда	Работа	Религия
Эмотивы	Косметика и парфюмерия	Быт	Быт
Увлечения	Процедуры	Финансы	Финансы
Личные имена	Здоровье	Здоровье	Политика
Бранная и обсценная лексика	Покупки	Отдых	

в этой группе религиозная лексика *(благословение, крест, божественный, молиться)*, причем контексты частотных слов отражают взгляды пользователей разных конфессий (в первую очередь христиан и мусульман).

Кроме того, заметим, что две «старшие» группы демонстрируют знание более абстрактной и книжной лексики, а «младшие» – наоборот, т. е. конкретной и сниженной. В особенности это касается первой группы (14–18 лет), в частотный список которой попали грубо-просторечные слова и мат, которые по понятным причинам в список выше мы не включили.

Идеографический анализ показывает, что в каждой возрастной группе выделяются следующие доминантные группы лексики (табл. 2).

Наиболее рельефными с точки зрения соответствия пользователям определенного возраста, с нашей точки зрения, являются группы 1 и 3. Лексические списки этих групп основаны на более чистых данных (т. е. большей доле оригинальных пользовательских текстов), и, кроме того, они более точно отражают поколенческие интересы.

Заключение

Таким образом, сравнительный лексико-статистический и идеографический анализ выявил, с одной стороны, предполагаемые признаки возрастных групп. С другой стороны, в частотном лексическом массиве каждой группы мы обнаружили неожиданные для себя единицы. Важно понимать, что при существующей лексической многозначности контексты многих слов необходимо проверять, чтобы избежать превратного взгляда на материал.

Данное исследование пока нельзя назвать абсолютно завершенным по двум причинам. Во-первых, выборка данных запланирована как минимум на целый год, чтобы она могла отразить весь годичный цикл существования социальной сети. На лексическую статистику существенное влияние оказывают сезонные и локальные причины – в частности, различные праздники, события национального масштаба и т. п. Во-вторых, большой объем текстовых корпусов затрудняет проведение качественного контекстологического анализа. В ближайшее время мы планируем завершить выборку материала, а также улучшить качество как машинной, так и ручной обработки полученных больших данных.

Перспективы исследований такого рода очевидны: новый для лингвистики и к тому же столь объемный материал дает результаты не только семантико-стилистического характера. Он обогащает научные изыскания в смежных с лингвистикой сферах: социологии, психологии, политологии и в любых социо-гуманитарных направлениях, для которых надежным источником данных может стать лингвистический корпус текстов.

Благодарности

Исследование выполнено при финансовой поддержке гранта РНФ № 19-18-00264 в рамках научного проекта «Цифровизация коммуникативно-культурной памяти и проблемы ее межпоколенческой трансляции».

Авторы выражают благодарность Морозову Алексею Дмитриевичу за помощь в сборе данных из социальной сети.

Литература

1. Анализ социальных сетей: методы и приложения [Электронный ресурс] / А.Коршунов, И.Белобородов, Н.Бузун и др. // Труды ИСП РАН. – 2014. – № 1. – URL: http://cyberleninka.ru/article/n/analiz-sotsialnyhseteymetody-i-prilozheniya (дата обращения: 15.01.2019).

2. Батура, Т.В. Методы анализа данных из социальных сетей / Т.В. Батура, Н.С. Копылова, Ф.А. Мурзин, А.В. Проскуряков // Вестн. Новосиб. гос. ун-та. Сер.: Информационные технологии. – 2013. – Т. 11. – Вып. 3. – С. 5–21.

3. Беликов, В.И. Социолингвистика: учебник для вузов / В.И. Беликов, Л.П. Крысин. – М.: РГГУ, 2001. – 315 с.

4. Бондаренко, С.В. Социальная структура виртуальных сетевых сообществ: автореф. дис. ... д-ра соц. наук / С.В. Бондаренко. – Ростов н/Д: Ростовский гос. пед. ун-т, 2001. – 28 с.

5. Губанов, Д.А. Социальные сети: модели информационного влияния, управления и противоборства / Д.А. Губанов, Д.А. Новиков, А.Г. Чхартишвили. – М.: Физматлит, 2010. – 228 с.

6. Карпова, М.К. Социальные сети как особый канал самопрезентации индивида [Электронный ресурс] / М.К. Карпова, М.А. Монина // Электронный научный журнал «Наука. Общество. Государство». – 2018. – Т. 6. – № 1. – URL: http://esj.pnzgu.ru (дата обращения: 2.10.2019).

7. Морослин, П.В. Язык Интернета как объект лингвистических исследований / П.В. Морос-

лин // Вестник РУДН. Сер.: Лингвистика. – 2009. – № 3. – С. 10–17.

8. Мухин, М.Ю. Индивидуальная лексическая сочетаемость и ее корпусная формализация / М.Ю. Мухин // Язык, культура, ментальность: проблемы и перспективы филологических исследований. – Курск, 2019. – С. 310–317.

9. Мухин, М.Ю. Лексическая статистика и концептуальная система автора: М. Булгаков, В. Набоков, А. Платонов, М. Шолохов / М.Ю. Мухин. – Екатеринбург: Изд-во Урал. ун-та, 2010. – 232 с.

10. Национальный корпус русского языка: http://www.ruscorpora.ru (дата обращения: 2.10.2019).

11. Новый частотный словарь русской лексики / под ред. О.Н. Ляшевской, С.А. Шарова. – URL: http://dict.ruslang.ru/freq.php (дата обращения: 02.10.2019).

12. Schwartz, H.A. Personality, Gender, and Age in the Language of Social Media: The Open Vocabulary Approach / H.A. Schwartz, J.C. Eichstaedt, M.L. Kern et al. // PLOS One. – 2013. – https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073791.

13. Sumskaya, A. "Analogue" and "digital" generation of mass media audience / A. Sumskaya, A. Lozovskaya // 6 SWS International Scientists Conference on Social Science (SGEM 2019). – 2019. – Vol. VI, Is. 5. – P. 113–120.

Мухин Михаил Юрьевич, доктор филологических наук, директор департамента лингвистики, профессор кафедры фундаментальной и прикладной лингвистики и текстоведения, Уральский федеральный университет им. первого Президента России Б.Н. Ельцина (Екатеринбург), mu-hi@ya.ru

Лозовская Алина Ильясовна, магистрант кафедры фундаментальной и прикладной лингвистики и текстоведения, лаборант-исследователь лаборатории компьютерной лексикографии, Уральский федеральный университет им. первого Президента России Б.Н. Ельцина (Екатеринбург), a.i.lozovskaya@yandex.ru.

Поступила в редакцию 2 сентября 2019 г.

DOI: 10.14529/ling190407

SOCIAL NETWORK AS A SOURCE OF SOCIOLINGUISTIC DATA: LEXICOSTATISTICAL ANALYSIS

M.Yu. Mukhin, *mu-hi@ya.ru*
A.I. Lozovskaya, *a.i.lozovskaya@yandex.ru*
Ural Federal University named after B.N. Yeltsin, Ekaterinburg, Russian Federation

The article is devoted to the substantiation of the methodology and comparative lexicostatistical analysis of sociolinguistic data obtained from the social network "VK". New possibilities of research of verbal Internet communication, including sociolinguistic features of Internet users, are discussed. The primary attention in the paper is paid to the correlation between the age of social network users and lexical features of their texts. Based on the data obtained (about 8 million words), the text corpora of four age groups of social network users aged from 14 to 74 years old have been compiled. The authors have conducted a comparative analysis of the frequency wordlists and, first of all, identified the words that are often used in the texts of users of all four groups. This array generally correlates with the data from the frequency dictionary of the Russian language, although it also has significant differences. The article presents the lists of frequency words typical of each age group; ideographic characteristics of the wordlists (dominant thematic groups of words) and sociolinguistic comments are provided. Conclusions are made about lexical and conceptual differences between texts of different age groups of users, as well as about the productivity of statistical and ideographic analysis of social media texts.

Keywords: social network, sociolinguistics, age, age group, corpus linguistics, lexical statistics, ideographic analysis.

References

1. Korshunov A., Beloborodov I., Buzun N. et al. Social *Media Analysis: Methods and Applications* [Analiz socialnyh setej: metody i prilozhenija]. URL: http://cyberleninka.ru/article/n/analiz-sotsialnyhseteymetody-i-prilozheniya (accessed: 15.01.2019)

2. Batura T.V., Kopylova N.S., Murzin F.A., Proskurjakov A.V. [Methods of Data Analysis From Social Networks]. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Serija "Informacionnye tehnologii"* [Bulletin of Novosibirsk State University. Series: "Information Technologies"]. 2013, no. 11, vol. 3, pp. 5–21.

3. Belikov V.I., Krysin L.P. Sociolingvistika: uchebnik dlja vuzov [Sociolinguistics: A textbook for universities]. Moscow, Publishing of the Russian State University for the Humanities, 2001. 315 p.

4. Bondarenko S.V. *Socialnaja struktura virtualnyh setevyh soobshhestv*. Avtoreferat of dokt. diss. [Social Structure of Virtual Network Communities. Abstract of Doct. Diss.]. Rostov-on-Don. 2001. 28 p.

5. Gubanov D.A., Novikov D.A., Chhartishvili A.G. *Socialnye seti: modeli informacionnogo vlijanija, upravlenija i protivoborstva* [Social Networks: Models of Information Influence, Management and Confrontation]. Moscow, Publishing of Fizmatlit, 2010. 228 p.

6. Karpova M.K., Monina M.A. *Social Networks as a Special Channel for Individuals' Self-Presentation* [Socialnye seti kak osobyj kanal samoprezentacii individa]. URL: http://esj.pnzgu.ru (accessed: 2.10.2019).

7. Moroslin P.V. [Language of the Internet as an object of linguistic research]. *Vestnik Rossijskogo universiteta druzhby narodov. Serija: Lingvistika* [Bulletin of the Peoples' Friendship University of Russia. Series: Linguistics]. 2009, no. 3, pp. 10–17.

8. Muhin M.Ju. Individualnaja leksicheskaja sochetaemost i ee korpusnaja formalizacija [Individual Lexical Compatibility and its Corpus Formalization]. *Jazyk, kul'tura, mentalnost: problemy i perspektivy filologicheskih issledovanij* [Language, Culture, Mentality: Problems and prospects of philological research]. 2019, pp. 310–317.

9. Mukhin M.Ju. *Leksicheskaja statistika i konceptual'naja sistema avtora: M. Bulgakov, V. Nabokov, A. Platonov, M. Sholohov* [Lexical Statistics and Conceptual System of the Author: M. Bulgakov, V. Nabokov, A. Platonov, M. Sholokhov]. Ekaterinburg, Publishing of the Ural State University, 2010. 232 p.

10. *National Corps of Russian Language*: http://www.ruscorpora.ru (accessed: 2.10.2019).

11. Lyashevskaya O.N., Sharov S.A. (eds.) *A Frequency Dictionary of the Russian Language*. URL: http://dict.ruslang.ru/freq.php (accessed: 02.10.2019).

12. Schwartz H.A., Eichstaedt J.C., Kern M.L., Dziurzynski L., Ramones S.M. et al. *Personality, Gender, and Age in the Language of Social Media: The Open Vocabulary Approach*. PLOS One. 2013. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073791.

13. Sumskaya. A., Lozovskaya, A. "Analogue" and "digital" generation of mass media audience. *6 SWS International Scientists Conference on Social Science (SGEM 2019)*. 2019, vol. VI, is. 5, pp. 113–120.

Mikhail Yu. Mukhin, Doctor of Philology, Head of the Linguistics Department, professor of the Department of Fundamental and Applied Linguistics and Textual Studies, Ural Federal University named after B.N. Yeltsin (Ekaterinburg), mu-hi@ya.ru.

Alina I. Lozovskaya, undergraduate student, Department of Fundamental and Applied Linguistics and Textual Studies, research assistant, the Laboratory of Computer Lexicography, Ural Federal University named after B.N. Yeltsin (Ekaterinburg), a.i.lozovskaya@yandex.ru.

Received 2 September 2019

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Мухин, М.Ю. Социальная сеть как источник социолингвистических данных: лексико-статистический анализ / М.Ю. Мухин, А.И. Лозовская // Вестник ЮУрГУ. Серия «Лингвистика». – 2019. – Т. 16, № 4. – С. 38–44. DOI: 10.14529/ling190407

FOR CITATION

Mukhin M.Yu., Lozovskaya A.I. Social Network as a Source of Sociolinguistic Data: Lexicostatistical Analysis. *Bulletin of the South Ural State University. Ser. Linguistics.* 2019, vol. 16, no. 4, pp. 38–44. (in Russ.). DOI: 10.14529/ling190407