УДК 81’33 DOI: 10.31249/chel/2025.02.05

Шерстинова Т.Ю., Вепринцева Д.А.

АНАЛИЗ ТЕМАТИКИ ПОВСЕДНЕВНЫХ РАЗГОВОРОВ: ЭКСПЕРТНЫЙ ПОДХОД И АВТОМАТИЧЕСКИЕ МЕТОДЫ[©, 1]

Национальный исследовательский университет «Высшая школа экономики», Россия, Санкт-Петербург, tsherstinova@hse.ru, daveprintseva@edu.hse.ru

Аннотация. В статье рассматриваются три разных подхода к изучению тематики повседневных разговоров: экспертная тематическая разметка и два автоматических метода (тематическое моделирование и кластеризация). Материалом для исследования послужили расшифровки русской устной повседневной речи из корпуса ОРД, подготовленные на основе звукозаписей спонтанных разговоров, выполненных в естественных коммуникативных ситуациях (дома, на работе, в учебном заведении, в магазине, в поликлинике и т.д.). Представлены результаты трех экспериментов, базирующихся на разных методах выявления тематических групп: 1) экспертное тематическое аннотирование транскриптов, дающее подробную картину тематики повседневного общения в динамике, 2) автоматическое тематическое моделирование, позволяющее выявить латентные темы в корпусе расшифровок, и 3) кластеризация, использованная для группировки разговоров по тематике на основе их лексического сходства. Получены статистические данные о распределении тем в повседневной речи на основе пилотной экспертной разметки, автоматически выявлены тематические классы для различных типов коммуникации, таких как общение с коллегами, членами семьи, друзьями и в процессе обучения. Проведенное исследование позволяет оценить эффективность использования автоматизированных методов в сравнении с экспертной разметкой для политематического корпуса неподготовленной повседневной речи.

[1] Публикация подготовлена в результате проведения исследования по проекту «Текст как большие данные: методы и модели работы с большими текстовыми данными», реализованного в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

Ключевые слова: русская повседневная речь; тематика повседневных разговоров; корпусная лингвистика; экспертная разметка; тематическое моделирование; кластеризация.

Поступила: 10.09.2024 Принята к печати: 28.12.2024

Sherstinova T.Y., Veprintseva D.A.

Thematic analysis of everyday conversations: expert approach and automated techniques[©, 1]

National Research University Higher School of Economics,
Russia, St. Petersburg, tsherstinova@hse.ru, daveprintseva@edu.hse.ru

Abstract. The paper concerns three different approaches to studying the topics of everyday conversations: expert thematic annotation, topic modeling, and clustering. The research explores transcripts of the Russian spoken language from the ORD corpus, derived from recordings of spontaneous conversations in natural communicative settings (e.g., at home, work, educational institutions, stores, clinics, etc.). The study presents the results of three experiments, each employing a different method for identifying thematic groups: 1) expert thematic annotation of transcripts, providing a detailed and dynamic picture of everyday communication topics, 2) topic modeling, which uncovers latent themes within the corpus of speech transcript, and 3) clustering, used to group conversations by topic based on lexical similarity. The research provides preliminary statistical data on the distribution of topics in everyday speech through expert annotation and automatically identifies thematic classes for various types of communication, such as interactions with colleagues, family members, friends, and within educational contexts. This study assesses the effectiveness of automated methods compared to expert annotation for analyzing a multi-thematic corpus of unstructured everyday speech.

Keywords: Russian everyday speech; topics of everyday conversations; corpus linguistics; expert annotation; topic modeling; clustering.

Received: 10.09.2024 Accepted: 28.12.2024

Введение

Статья посвящена исследованию тематической организации устной речи – тому, о чем наши современники говорят в повседневной жизни (дома, на работе, в общественных местах). Интерес

[1] The publication was prepared within the framework of a research project "Text as Big Data: Methods and Models for Working with Large Textual Data" implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

к тематическому разнообразию повседневного дискурса определяется важностью этой темы для решения теоретических и практических лингвистических задач (описания лингвистических характеристик для определенных жанров общения, исследований прагматики, лексикографии, обучения русскому языку как иностранному), проведения антропологических, психолингвистических, социологических исследований, а также для разработки максимально приближенных к естественному речевому общению диалоговых систем и языковых моделей. Понимание тематического разнообразия речи будет способствовать более глубокому пониманию когнитивных и культурных процессов, отражаемых в языке, выявлению моделей тематической организации разговорной речи, а также улучшению технологий автоматического анализа и понимания текстов.

Несмотря на то, что мы пользуемся повседневной речью ежедневно, а ее тематический аспект не раз становился объектом внимания отечественных лингвистов (см., например: [Скирдач, 1984; Матвеева, 1990; Сибирякова, 1990; Грудева, 2008; Косицина, 2012] и др.), ее статистические характеристики выпадают из поля исследования ученых из-за объективной сложности получения представительных количественных данных, особенно в отношении тематического разнообразия повседневных разговоров. В данной статье рассматриваются результаты трех экспериментов, посвященных изучению тематического разнообразия устной речи: 1) экспертной тематической разметки, 2) автоматического тематического моделирования и 3) автоматической кластеризации.

Материалом для проведения исследования послужили расшифровки устных спонтанных разговоров из корпуса «Один речевой день» (ОРД корпус) [Корпус русского языка повседневного общения …, 2019], записанные в условиях естественных коммуникативных ситуаций (дома, на работе, в учебном заведении, в магазине, в поликлинике и т.д.) в 2007–2016 гг. [The ORD speech corpus …, 2009; Sociolinguistic extension …, 2016]. Волонтеры, согласившиеся принять участие в сборе звукозаписей для корпуса, должны были прожить в течение дня с включенным диктофоном, записывающим всю их речевую коммуникацию. Информанты подобраны таким образом, что в корпусе хорошо представлены разные социальные группы российского мегаполиса: по полу, возрасту, профессии, уровню образования и др.

Единицей хранения и обработки данных в корпусе ОРД являются звуковые и текстовые файлы, соотносящиеся с одним макро-

эпизодом повседневной речи [Шерстинова, 2013]. Макроэпизод является однородным по локусу, участникам и прагматической задаче фрагментом коммуникации (например, «ужин в семейном кругу» или «совещание на работе»), при этом в одном макроэпизоде может происходить обсуждение достаточно большого количества тем.

Экспертное тематическое аннотирование расшифровок

Классическим подходом к выделению и аннотированию фрагментов речи на корпусном материале является их экспертная разметка, выполняемая вручную. В случае работы с мультимедийным материалом такое аннотирование может осуществляться непосредственно в среде корпусного менеджера, поддерживающего экспертное аннотирование данных, которым для корпуса ОРД является ELAN [ELAN], а также проводиться в любом компьютерном приложении, позволяющем работать с экспортированными из базы данных табличными значениями. Для экспертного аннотирования расшифровок по тематике в нашем первом эксперименте был использован MS Excel.

Для проведения исследования было отобрано 60 макроэпизодов, полученных от информантов – представителей разных социальных групп. Каждый эпизод был прослушан, поделен на тематически однородные фрагменты, которые были описаны по следующей схеме: 1) код макроэпизода по корпусу ОРД; 2) место коммуникации; 3) начало микротемы (в минутах); 4) конец микротемы (в минутах); 5) микротема; 6) вид смены микротемы; 7) возрастная категория говорящих [Akinshina, Sherstinova, 2022]. Пример заполнения таблицы для одного из макроэпизодов представлен в табл. 1.

В результате проведенной работы было получен список из 355 тематически проаннотированных фрагментов, то есть в среднем почти шесть разных тем на один макроэпизод. После этого была проведена нормализация полученных данных: выявлено 16 макротем, которые можно считать обобщающими для полученных микротем [Akinshina, Sherstinova, 2022]: 1) наука; 2) искусство; 3) природа; 4) бытовые вопросы; 5) обсуждение прошлого и планы на будущее; 6) еда; 7) здоровье; 8) работа; 9) семья и друзья; 10) покупки; 11) социальные вопросы; 12) государство и политика; 13) образование; 14) хобби; 15) погода; 16) автомобили. Итоговая статистика по распределению макротем по частоте и длительности обсуждения на исследуемой выборке представлена в табл. 2.

Таблица 1

Пример экспертного аннотирования эпизодов по тематике (для эпизода ordS08-01 локус: в автомобиле, собеседники: две молодые девушки)

Макроэпизод	Номер фрагмента	Время начала	Время окончания	Микротема	Вид смены темы
ordS08-01	1	01:02	01:19	выражение эмоций	обстоятельственная
ordS08-01	2	06:37	07:57	романтические отношения	резкая
ordS08-01	3	07:58	08:25	беременность подруги	цепная
ordS08-01	4	08:57	09:29	ПДД, поведение пешеходов на улице	обстоятельственная
ordS08-01	5	09:29	09:59	беременность подруги	кольцевая
ordS08-01	6	11:42	12:16	времяпрепровождение с молодым человеком	обстоятельствен-ная
ordS08-01	7	14:52	15:25	физическая активность в разном возрасте	обстоятельственная
ordS08-01	8	15:35	16:34	чтения, выступления	резкая

Таблица 2

Дистрибуция макротем по частоте встречаемости и длительности обсуждения

Ранг по частоте	Макротема	Доля от общего количества тем, %	Длительность обсуждения, %	Ранг по длительности
1	работа	12,5	13,9	1
2	прошлое и будущее	12,5	10,2	3
3	здоровье	11,5	11,7	2
4	искусство	10	8,8	5
5	бытовые вопросы	9,6	10,1	4
6	социальные вопросы	8,4	7,9	6
7	семья и друзья	6,7	7,4	7
8	еда	5,6	5,8	9
9	наука	5,2	6,7	8
10	государство и политика	3,8	2,6	13
11	природа	3,6	3,5	10
12	образование	3,4	2,6	14
13	автомобили	2,8	3,4	11
14	покупки	2,2	2,8	12
15	хобби	2	2,3	15
16	погода	0,2	0,3	16

Эти результаты показывают, что в целом по выборке наиболее часто и длительно обсуждаемыми темами в повседневной жизни являются: *работа*, *здоровье*, *прошлое и будущее*, популярными также являются разговоры на *бытовые и социальные темы*, тему *искусства*, о жизни *семьи и друзей*.

Преимуществом экспертного аннотирования можно считать относительно точные данные о количестве времени, которое ушло у собеседников на обсуждение той или иной темы. Недостатками экспертного подхода являются прежде всего его высокая трудоемкость, а также зависимость получаемых результатов от используемой при аннотировании классификации макротем.

Тематическое моделирование устной речи

Тематическое моделирование – специализированный метод для выявления латентных тем в заданном корпусе документов. Изначально этот метод разрабатывался для обработки научных и новостных текстов, однако есть успешные свидетельства его применения и к текстам иных жанров. Для проведения тематического моделирования было отобрано 246 макроэпизодов, которые впоследствии группировались по трем выборкам: «коллеги» (общение с сослуживцами; 99 эпизодов; 263 063 слов), «семья» (общение с домашними и членами семьи; 88 эпизодов; 211 515 слов) и «друзья» (общение в среде друзей; 77 эпизодов; 178 284 слов).

В этап предобработки входило удаление предлогов, местоимений, частиц, союзов, междометий и стоп-слов.

Тематическое моделирование проводилось посредством LDA (латентного размещения Дирихле), специализированного метода для выявления тематических структур в текстовых документах [Kherwa, Bansal, 2018]. Потенциальное количество тем для выборок определялось с помощью показателя перплексии и сравнительного анализа тематических моделей с разным количеством тем. Каждая тема представлена десятью наиболее релевантными для нее словами.

Полученные тематические модели для каждой выборки были проинтерпретированы, и каждая тема получила свое обозначение[1]. Так, в выборке «коллеги» (табл. 3) были обнаружены темы *литература* и *кино* (t_1), *работа за компьютером* (t_3), *искусство* (t_4),

[1] Наименования тем определены экспертным образом.

закупки / *химические вещества* (t_5), *путешествие* (t_6), *документация* (t_7), *бронирование жилья* (t_8), *стоимость ремонта* (t_9) и *промышленные изделия* (t_10). Для темы (t_2) содержательная интерпретация оказалась затруднена, но можно предположить, что речь идет *о потере работы* (*лицами женского пола*), с подтемой, характеризующейся отношениями между *русскими и китайцами.*

Таблица 3

Тематическая модель: выборка «коллеги»

t_1	понимать фильм обзор книжка интересный балл информация подумать турецкий черепаха[1]
t_2	хороший думать женщина потерять девушка работа русский сумка масло китаец
t_3	компьютер работа писать точки быстро открыть галлий полчаса выходить файл
t_4	история смотреть видео язык говорить мир думать спектакль делать сцена
t_5	белый плюс завтра спирт вариант потеря каталог подумать силикон хороший
t_6	море лавировали процесс кусок каша ждать вкусно общение есть вода
t_7	счёт документы договор письмо деньги пятница дорога семьсот позвонить протокол
t_8	деньги номер чек касса паспорт бронь утро позвонить сутки сдача
t_9	тридцать пятьдесят семьдесят девяносто стоимость ремонт ноль понедельник бюджет здание
t_10	насос сделать труба быстро задание часть поставить план расходомер вода

Для выборки «семья» (табл. 4) основными темами стали: *еда* (t_1 и t_4), *поездка* (t_2, t_5 и t_10), *покупки* (t_3), *выполнение школьных домашних заданий* (t_8), *праздник* (t_6) и *день рождения* (t_9). Топик (t_7) объединил две разные темы: *лечение* и *отдых на природе*.

[1] Серым цветом выделены слова, очевидным образом «выбивающиеся» из темы.

Таблица 4

Тематическая модель: выборка «семья»

t_1	мясо есть сосиски яйца ребёнок молоко триста бизнес курица рыба
t_2	номер поезд минуты станция ёлки сторона сделать платформа внимание приезжать
t_3	знать идти рубли смотреть восемьдесят посмотреть стоить обои купить мама
t_4	чай мама идти пить кушать телефон нравиться суп есть хотеть
t_5	понятно ездить кофе вечер погода посмотреть позвонить разговаривать ехать видеть
t_6	шампанское брют чай кошка домик маленький пиво пирожные пино пить
t_7	зубы драть кола пятый лечить голова лагерь смотреть быстро мангал
t_8	сто метр компьютер мамочка икс окно квадратный семьдесят гектары номер
t_9	вкусный рождение купить сыр чай концерт кусочек вино бокал балет
t_10	идти дом ключи камни мама ехать обратно тихо машина пара

Таблица 5

Тематическая модель: выборка «друзья»

t_1	говорить платье ткани нравиться классно ручной фишка продавать тканный шкаф
t_2	мама машина родители выпить дешевле деньги читать гости двери пиво
t_3	деньги шутка дела вместе квартира покупать дом рождение звонить порядок
t_4	цветы боль препарат мигрень посадить приезжать пригодиться сосуды лечение красивые
t_5	деньги давать неделя найти пиво водка думать чувак сосиски комната
t_6	свадьба театр понять выпить свадебный диабло петь танцы прикольный невеста

Тематическая модель для выборки «друзья» (табл. 5) выявила всего шесть тем, среди которых: *одежда* (t_1), *наведение порядка в доме*, *покупка недвижимости* (t_3), *свадьба* (t_6). В этой выборке прослеживается выявление разных топиков по возрастам собеседников. Так, два топика (t_2 и t_5) описывают «заботы» молодых людей, связанные с поиском денег и желанием выпить, в то время как топик (t_4), характеризует, по-видимому, общение друзей / подруг старшего возраста, фокусирующихся на двух семантических зонах: *здоровье* и *работа на даче*.

Таким образом, применение методов тематического моделирования на материале расшифровок спонтанной речи показало не столь убедительный результат, как при использовании для специальных текстов, но тем не менее вполне интерпретируемый, сопоставимый с применением этих методов для анализа художественной литературы [Митрофанова, 2019; Шерстинова, Кирина, Москвина, 2024]. Повтор отдельных слов в разных построенных топиках наводит на мысль, что список стоп-слов стоит еще больше расширить, что может улучшить интерпретируемость построенных моделей. Имеет смысл также провести подобные эксперименты с использованием других тематических моделей и с разными вариантами предобработки, в том числе и не на лемматизированных текстах, с целью выявления оптимальной методики тематического моделирования для анализа политематической устной речи.

Тематическая кластеризация транскриптов устной речи

Кластеризация – еще один из методов автоматической группировки текстовых данных, используемых для выявления тематических классов документов. В качестве материала для следующего эксперимента были использованы три выборки макроэпизодов из корпуса ОРД, отличающиеся по составу участников диалога: «семья» (разговоры с домашними; 178 эпизодов; 417 649 слов), «коллеги» (разговоры с коллегами на профессиональные и бытовые темы; 127 эпизодов; 319 675 слов) и «образование» (учебная коммуникация; 57 эпизодов; 185 541 слов).

Предобработка текстов включала удаление всех цифр, небуквенных знаков, иностранных слов, имен собственных, союзов, частиц, междометий, местоимений, предлогов и стоп-слов. По результатам тестового эксперимента было решено очистить тексты и от

глаголов, которые негативно влияли на интерпретируемость полученных кластеров.

Кластеризация осуществлялась с помощью метода k-средних, который позволяет объединить *n* наблюдений в k кластеров таким образом, чтобы каждое наблюдение входило в кластер с ближайшим средним [Text clustering using K-mean, 2021] на основе векторизации текстов, проведенной с помощью метода TF-IDF, рассчитывающего важность слова в текстовом документе [Khan, Yurong, Sajid, 2019]. Для определения оптимального числа кластеров был использован метод силуэтов [Saputra, Saputra, Oswari, 2019]. Проверка коэффициента силуэта проходила для количества кластеров от 2 до 20; рассматривалось значение *k* с наиболее высоким коэффициентом силуэта [ibid.]. Для анализа интерпретируемости полученных кластеров был проведен сравнительный анализ результатов кластеризаций с разным *k*, поскольку метод силуэтов не всегда корректно отражает наиболее предпочтительное количество кластеров. Каждый кластер представлен десятью словами с указанием их веса.

Наиболее интерпретируемыми оказались результаты кластеризации образовательной коммуникации, что легко объясняется тематической однородностью разговоров такого типа. Так, для выборки «образование» было образовано 18 кластеров (табл. 6), которые соотносятся с 16 основными темами: *пение* (c_1), *естественные науки* / *измерения* (c_2), *музыка* (c_3), *экзамен* / *статистика* (c_4), *биология* (c_5), *свадьба* (c_6), *защита диссертации* (c_7), *увольнительная* (c_8), *балет* / *спорт* (c_9), *школа* (c_10), *медицина* (c_11), *искусство* / *театр* (c_12 и c_17), *загородная жизнь* / *природа* (c_14), *компьютерные технологии* (c_15) и *сталелитейное производство* (c_16). Однозначная интерпретация кластера (c_18) вызывает затруднение – здесь присутствует лексика, связанная с военным делом, игрой и литературой.

Для выборки, отражающей коммуникацию с коллегами, было сформировано 18 кластеров (табл. 7). Эти классы отличает бóльшая неоднородность, чем при образовательной коммуникации, что вполне ожидаемо. 17 из полученных кластеров можно отнести к следующим темам: *спорт* / *футбол* (c_1), *отпуск* (c_2), *искусство* (c_3), *ремонтные работы* / *жилье* (c_4), *наука и медицина* (c_6), *строительство* (c_7), *поездка* (c_8), *текущий ремонт* (c_9), *социальные сети* (c_10), *работа* (c_11 и c_12), *закупки* (c_13), *доставка* (c_14), *покраска* (c_15), *краска* (c_16), *пластиче-*

ская хирургия (c_17) и *работа на компьютере* (c_18). Кластер (c_5) объединяет лексику, *связанную с едой* и *разными национальностями* (*русский*, *китаец*, *айзер*(*байжданец*).

Таблица 6

Результат кластеризации для выборки «образование»

c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
Расщепление: 0.4255	метр: 0.2422	нота: 0.1941	ноль: 0.2787	слово: 0.1636	букет: 0.3302	отзыв: 0.3564	Рапорт 0.4006	ручка: 0.2525
связка: 0.3932	секунда: 0.2027	нук*: 0.1328	статистика: 0.2477	экология: 0.1584	невеста: 0.2185	страница: 0.1588	час: 0.2013	меч: 0.2418
опора: 0.2379	нулевой: 0.1765	мажор: 0.1115	макр*: 0.2397	индетный: 0.1487	платье: 0.1605	телефон: 0.1198	йогурт: 0.1372	пуант: 0.2235
голос: 0.1926	квадрат: 0.1489	минор: 0.1095	балл: 0.2235	кофе: 0.1266	свадьба: 0.1582	свидание: 0.1189	утро: 0.1333	лезвие: 0.1957
звук: 0.1649	давление: 0.1388	левый: 0.1039	доля: 0.2189	человек: 0.0981	брак: 0.1518	кафедра: 0.1168	пацан: 0.1250	кинжал: 0.1957
пение: 0.1392	воздух: 0.1074	романс: 0.1017	экзамен: 0.1959	молоко: 0.0938	спортивный: 0.1302	научный: 0.1130	выходной: 0.0924	стрела: 0.1842
трендовый: 0.1126	объём: 0.0967	разочек: 0.0971	п**ц: 0.1125	окружающий: 0.0900	классика: 0.1256	упорство: 0.1057	увольнение: 0.0898	полкилограмма: 0.1596
человек: 0.1093	площадь: 0.0964	внимание: 0.0927	половина: 0.1125	биология: 0.0880	человек: 0.1188	диссертация: 0.1023	пачка: 0.0873	собачка: 0.1434
подзязочный: 0.0876	ноль: 0.0894	концерт: 0.0855	курс: 0.0907	происхождение: 0.0850	глагол: 0.1135	собранный: 0.1019	магазин: 0.0853	штучка: 0.0975
язык: 0.0872	энергия : 0.0882	любовь: 0.0823	оценка: 0.0839	коробочка: 0.0834	жених: 0.0969	библиотека: 0.0952	сигарета: 0.0851	мамочка: 0.0869

c_10	c_11	c_12	c_13	c_14	c_15	c_16	c_17	c_18
год: 0.1605	иммуноглобулин: 0.2675	зависимый: 0.1783	реклама: 0.3554	дерево: 0.3665	текст: 0.2173	чугун: 0.5761	история: 0.1142	группа: 0.1430
чай: 0.0889	бактерия: 0.2634	вопрос: 0.1724	от**ля: 0.3554	клумба: 0.1283	устройство: 0.1797	сталь: 0.2315	интерьер: 0.1028	именной: 0.1342
запах: 0.0831	клетка: 0.2202	культура: 0.1208	с**а: 0.2808	лес: 0.1250	практика: 0.1033	ковкий: 0.1600	хокку: 0.0977	тип: 0.0838
свинья: 0.0797	лимфоцит: 0.2194	причина: 0.1075	наушник: 0.2808	огород: 0.1174	технология: 0.0950	серый: 0.1314	архитектор: 0.0923	игра: 0.0837
перерыв: 0.0791	тимус: 0.2107	человек: 0.0919	бурят: 0.2369	крыша: 0.1116	мультимедиа: 0.0939	марк*: 0.1172	актёр: 0.0871	рота: 0.0810
школа: 0.0737	антиген: 0.1774	вода: 0.0911	айзер: 0.2369	тёмный: 0.1072	изображение: 0.0928	миллиметр: 0.1081	искусство: 0.0847	книга: 0.0760
мама: 0.0698	макрофаг: 0.1662	актёр: 0.0871	минута: 0.2031	кот: 0.1061	производственный: 0.0902	углерод: 0.1027	пьеса: 0.0798	партия: 0.0719
петух: 0.0659	антитело: 0.1535	пьеса: 0.0800	пацан: 0.1773	ёжик: 0.1013	пароль: 0.0892	бронза: 0.1000	государство: 0.0785	улица: 0.0600
класс: 0.0622	ткань: 0.1462	речь: 0.0781	ячмень: 0.1185	охра: 0.0943	общий: 0.0810	углеродистый: 0.0992	год: 0.0761	пространство: 0.0575
файл: 0.0594	иммунный: 0.1348	имя: 0.0750	мадмас: 0.1185	дом: 0.0940	видео: 0.0747	качественный: 0.0987	человек: 0.0761	винтовка: 0.0553

Таблица 7

Результат кластеризации для выборки «коллеги»

c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
год: 0.2170	путёвка: 0.2154	история: 0.1432	труба: 0.0842	китаец: 0.2180	балл: 0.0852	насос: 0.2293	чек: 0.1491	человек: 0.0918
навес: 0.1420	море: 0.1793	спектакль: 0.1242	вода: 0.0755	масло: 0.1971	налёт: 0.0799	решётка: 0.0969	номер: 0.1358	год: 0.0901
дед: 0.1351	питание: 0.1418	сцена: 0.0831	заявка: 0.0684	скумбрия: 0.1444	лакунарный: 0.0799	труба: 0.0737	телефон: 0.1031	книжка: 0.0728
середина: 0.1342 (поля)	табличка: 0.0929	ширма: 0.0815	утро: 0.0651	китайский: 0.1382	ноябрь: 0.0762	должный: 0.0572	бронь: 0.0971	игра: 0.0726
матч: 0.1230	филологический: 0.0903	фильм: 0.0780	подвал: 0.0581	сумка: 0.1204	живот: 0.0737	теплообменник: 0.0529	семёрка: 0.0953	банк: 0.0703
экзамен: 0.1219	сиделка: 0.0801	французский: 0.0697	лючок: 0.0555	айзер: 0.1160	папочка: 0.0668	метр: 0.0527	кондиционер: 0.0820	бюджет: 0.0654
четверг: 0.1207	свободный: 0.0771	человек: 0.0654	маленький: 0.0546	подсолнечный: 0.1160	защита: 0.0654	воронка: 0.0527	касса: 0.0705	доллар: 0.0648
водолей: 0.1173	жёлтенький: 0.0738	театр: 0.0636	добрый: 0.0539	кольцо: 0.1115	статья: 0.0608	вода: 0.0524	сдача: 0.0701	нужный: 0.0611
отец: 0.1164	аллё: 0.0721	режиссёр: 0.0626	тётя: 0.0526	двор: 0.0991	иллюзия: 0.0605	затвор: 0.0516	паспорт: 0.0633	текущий: 0.0588
центр: 0.1007	искусство: 0.0708	библия: 0.0585	общежитие: 0.0522	русский: 0.0922	учёный: 0.0598	насосный: 0.0510	сумка: 0.0632	ремонт: 0.0548
c_10	c_11	c_12	c_13	c_14	c_15	c_16	c_17	c_18
тип: 0.1142	деньга: 0.1267	человек: 0.0556	рубль: 0.0702	заказ: 0.2272	краска: 0.1639	стибиум: 0.2261	палец: 0.2695	компьютер: 0.1330
хороший: 0.0721	месяц: 0.0778	дело: 0.0443	лист: 0.0619	прыщ*: 0.1786	белый: 0.1169	раствор: 0.1586	силикон: 0.1653	файл: 0.1151
картина: 0.0681	сапог: 0.0675	работа: 0.0406	безнал: 0.0603	тачка: 0.1594	глянцевый: 0.1008	галлий: 0.1413	след: 0.1207	программа: 0.1008
человек: 0.0681	октябрь: 0.0595	должный: 0.0397	договор: 0.0566	бом*: 0.1305	право: 0.0992	выход: 0.1394	спирт: 0.1109	инструкция: 0.0808
мир: 0.0631	ядовитый: 0.0584	день: 0.0378	доставка: 0.0532	курьер: 0.1208	бабка: 0.0951	образец: 0.1277	носик: 0.1074	оперативный: 0.0804

c_10	c_11	c_12	c_13	c_14	c_15	c_16	c_17	c_18
потеря: 0.0611	человек: 0.0566	год: 0.0333	кило-грамм: 0.0529	бампер: 0.1184	деньга: 0.0825	пассивация: 0.1007	апрель: 0.0771	ноль: 0.0633
карта: 0.0596	договор: 0.0565	ладный: 0.0329	массажик: 0.0471	чтоль*: 0.1152	борт: 0.0823	коллекция: 0.1005	ветошь: 0.0769	почта: 0.0625
пост: 0.0531	сотруд-ник: 0.0559	телефон: 0.0311	прайс: 0.0463	замена: 0.0942	жёлтый: 0.0813	работа: 0.0911	дочка: 0.0704	уведомле-ние: 0.0624
паблик: 0.0530	служебок: 0.0534	большой: 0.0301	машина: 0.0451	доставка: 0.0927	камень: 0.0784	спиртовой: 0.0779	июль: 0.0689	аналитик: 0.0620
картинка: 0.0522	плата: 0.0492	свидание: 0.0299	дыщ*: 0.0405	полкило-грамм: 0.0893	правый: 0.0662	трилон: 0.0779	ошейник: 0.0651	видеокарта: 0.0616

Результаты, полученные для домашних разговоров внутри семьи, отличаются тем, что здесь наряду с легко атрибутируемыми темами на выходе получается большее количество неустановленных тем. Как показано в табл. 8, для этой выборки было сформировано десять кластеров, семь из которых можно считать вполне интерпретируемыми: *семейный обед* (c_1), *работа* (c_2 и c_4), *природа* (c_3), *завтрак* (c_6), *детский спорт* / *здоровье* (c_7), *выполнение школьной домашней работы* (c_8), *покупки* (c_9) и *поездка* (c_10). Кластер (c_5) содержит две семантические группы: *фотоателье* и *животные*.

Таблица 8

Результат кластеризации для выборки
«семейная коммуникация»

c_1	c_2	c_3	c_4	c_5
папа: 0.1839	договор: 0.0644	кот: 0.1157	человек: 0.0587	фотография: 0.1562
мама: 0.0801	пакет: 0.0567	мясо: 0.0706	машина: 0.0447	тигр: 0.1284
суп: 0.0624	понедельник: 0.0539	карельский: 0.0657	хороший: 0.0439	козёл: 0.1119
бабушка: 0.0562	работа: 0.0537	домик: 0.0648	день: 0.0392	енот: 0.1006
конфета: 0.0440	кафедра: 0.0494	берёза: 0.0629	деньга: 0.0359	собака: 0.0761

c_1	c_2	c_3	c_4	c_5
ключ: 0.0429	книжка: 0.0429	холд: 0.0592	вода: 0.0316	паспорт: 0.0615
дедушка: 0.0425	большой: 0.0426	дубль: 0.0541	дело: 0.0311	пила: 0.0602
плохой: 0.0388	плитка: 0.0413	ёлка: 0.0496	работа: 0.0297	дело: 0.0571
большой: 0.0375	мама: 0.0402	хороший: 0.0493	время: 0.0297	фотоателье: 0.0541
рука: 0.0374	чай: 0.0353	небо: 0.0484	год: 0.0278	ачетырить*: 0.0501
c_6	c_7	c_8	c_9	c_10
хороший: 0.0359	бассейн: 0.3160	мама: 0.0687	рюкзак: 0.1442	машина: 0.0511
мама: 0.0356	ладненький: 0.0958	солнышко: 0.0588	обои: 0.1010	улица: 0.0471
чай: 0.0332	папа: 0.0518	год: 0.0488	скидка: 0.0885	мама: 0.0440
большой: 0.0328	группа: 0.0505	депутат: 0.0485	магазин: 0.0742	общий: 0.0423
утро: 0.0327	мальчик: 0.0491	день: 0.0442	дешёвый: 0.0709	дело: 0.0389
день: 0.0317	детский: 0.0487	молодец: 0.0414	цена: 0.0686	место: 0.0320
сыр: 0.0298	бронх: 0.0461	подлежащее: 0.0365	хороший: 0.0628	дорога: 0.0313
яйцо: 0.0291	бабушка: 0.0443	остаток: 0.0347	метр: 0.0620	карта: 0.0286
каша: 0.0291	поздний: 0.0410	сказуемое: 0.0345	лекало: 0.0580	хороший: 0.0256
время: 0.0277	поездка: 0.0390	дело: 0.0341	коридор: 0.0565	тая: 0.0255

Что касается тематического состава для всех трех выборок в целом, то наиболее нейтральные (общеупотребительные) темы характерны для выборки «семья», в то время как для выборок «образование» и «коллеги» кластеризация показала эффективность в выявлении специфических тематик. Тем не менее в полученных моделях прослеживаются и общие темы, к которым можно отнести тему *спорт* (для всех выборок), темы *искусство* и *техника* (общие

темы для выборок «коллеги» и «образование»), темы *здоровье* и *поездка* (общие для выборок «коллеги» и «семья») и тему *природа* (для выборок «образование» и «семья»). В целом, метод кластеризации можно считать вполне эффективным для выявления тематического состава документов. Что касается интерпретируемости кластеров и их соответствия расшифровкам, то выборочный анализ тематик и соответствующих им текстов продемонстрировал, что обозначенные темы действительно присутствуют в транскриптах аудиоэпизодов.

В качестве недостатков построенных лексических кластеров можно отметить как повторы частотной лексики (*мама*, *папа*, *хороший*, *человек* и др.), так и появление в кластерах одиночных неправильных форм, выделенных в таблице символом звездочка. Стоит рассмотреть возможность переноса такой лексики в словарь стоп-слов и проведение повторной кластеризации.

В целом, сравнение результатов кластеризации и тематического моделирования продемонстрировало, что оба подхода могут быть использованы для изучения тематического разнообразия устной русской спонтанной речи, дополняя друг друга в зависимости от поставленных задач.

Заключение

Проведенные эксперименты показали, что все задействованные методы извлечения тематических данных применимы для использования при исследовании тематики повседневных разговоров, при этом каждый имеет как свои плюсы, так и минусы. Экспертная разметка хороша тем, что она является «экспертной» в буквальном смысле и позволяет выделять темы с привязкой к временной шкале. Более того, ее результатом становится важный лингвистический материал, на основе которого можно проводить дальнейшие исследования (например, выделить все реплики, относящиеся к теме «погода» или «взаимоотношения»), а также проводить их количественный анализ (сколько времени занимает обсуждение той или иной темы). Недостатком экспертного метода является субъективность оценки в плане того, какие темы и подтемы выделяются, насколько дробно макроэпизоды членятся на темы, но самым главным ограничением этого метода является высокая трудоемкость

такой работы[1]. Обработка больших объемов данных вручную невозможна, поэтому нужно искать оптимальные методы автоматизации.

Два таких метода впервые апробированы на материале расшифровок русской спонтанной речи в данной работе – это тематическое моделирование и кластеризация. И тот, и другой метод лучше всего работают на специальных текстах, характеризующихся высокой долей терминов, выступающих надежными маркерами для выявления темы. Подмножества расшифровок образовательной коммуникации и, отчасти, разговоры с коллегами, в которых речь идет о профессиональных вопросах, тоже можно условно считать «специальными» текстами, и, как результат, они показывают довольно высокий процент интерпретируемых классов. Что же касается бытовых частных разговоров, их тематическое разнообразие настолько велико, а сама речевая коммуникация настолько неоднородна в тематическом плане, что эффективность применения этих методов к этому классу расшифровок на данный момент не очевидна. К тому же результаты автоматической кластеризации и тематического моделирования в большой степени зависят от предобработки исходных текстов (лемматизации или ее отсутствия, списка стоп-слов) и разных моделей для выполнения процедуры, поэтому эксперименты по выявлению оптимальных методов предобработки устных текстов однозначно стоит продолжать. Следует также заметить, что у обоих этих методов есть существенный недостаток – рассмотрение моделью каждого макроэпизода как тематически однородного объекта. Эксперимент с ручной разметкой наглядно показал, что макроэпизоды устной речи, как правило, весьма неоднородны с тематической точки зрения и их сегментация на темы представляет собой отдельную задачу.

Для более тонкого выявления тематических фрагментов можно предложить два способа: 1) подготовку экспертной разметки на представительном текстовом объеме звукозаписей и использование методов машинного обучения для получения тематических данных по всему массиву расшифровок корпуса, 2) использование для тематической разметки больших языковых моделей, применение которых для решения задач автоматической разметки лин-

[1] Например, чтобы получить тематическое аннотирование всех звукозаписей ОРД, насчитывающих более 1400 часов звучания, потребовалось бы 175 рабочих дней.

гвистических корпусов представляется весьма перспективным (см., например: [Automated speech act annotation …, 2024]).

Список литературы

Грудева Е.В. Способы введения и определения темы в русском языке и стратегии носителей языка в ее определении (экспериментальное исследование) // Известия Российского государственного педагогического университета им. А.И. Герцена: общественные и гуманитарные науки (философия, языкознание, литературоведение, культурология, экономика, право, история, социология, педагогика, психология) : научный журнал. – 2008. – № 11(71). – С. 36–44.

Корпус русского языка повседневного общения «Один речевой день»: текущее состояние и перспективы / Богданова-Бегларян Н.В., Блинова О.В., Мартыненко Г.Я., Шерстинова Т.Ю. // Труды Института русского языка им. В.В. Виноградова. Вып. 21. Национальный корпус русского языка: исследования и разработки. – Москва , 2019. – С. 101–110.

Косицина Ю.В. Текущее тематическое развитие: когерентный и когезивный аспекты // Вестник Кемеровского государственного университета. – 2012. – Вып. 1 – № 52. – С. 281–284.

Матвеева Т.В. Тематическое развертывание разговорного текста // Языковой облик уральского города : сб. науч. трудов. – Свердловск : УрГУ, 1990. – С. 46–54.

Митрофанова О.А. Исследование структурной организации художественного произведения с помощью тематического моделирования: опыт работы с текстом романа «Мастер и Маргарита» М.А. Булгакова // Корпусная лингвистика 2019. – Сакнт-Петербург, 2019. – С. 387–394.

Сибирякова И.Г. Опыт тематического анализа диалогического разговорного текста // Языковой облик уральского города : сб. науч. трудов. – Свердловск : УрГУ, 1990. – С. 61–71.

Скирдач О.М. Динамика развития темы в тексте // Лексико-фразеологическая система немецкого языка и ее реализация в тексте : сб. науч. трудов МГПИИЯ им. М. Тореза. – Москва, 1984. – Вып. 232. – С. 149–165.

Шерстинова Т.Ю. Коммуникативные макроэпизоды в корпусе повседневной русской речи «Один речевой день»: принципы аннотирования и результаты статистической обработки // Труды международной конференции «Корпусная лингвистика – 2013». – Санкт-Петербург, 2013. – С. 449–456.

Шерстинова Т.Ю., *Кирина М.А.*, *Москвина А.Д.* Тематическое моделирование художественной прозы: оценка и интерпретируемость результатов (на примере русского рассказа 1900–1930 гг.) // Вестник Томского государственного университета. Филология. – 2024. – № 89. – С. 127–151. DOI: 10.17223/19986645/89/6. – URL: https://pureportal.spbu.ru/files/122949387/_89_.pdf?ysclid=m4quz23l4t639599053

Akinshina E., *Sherstinova T.* Thematic diversity of everyday Russian discourse: a case study based on the ORD corpus // Mahadeva Prasanna et al. (eds.) SPECOM 2022, LNCS. – Springer Nature, 2022. – Vol. 13721. – P. 1–9.

Automated speech act annotation in a Russian spoken corpus using large language models: a comparative study / Sherstinova T., Firsanova V., Novoseltseva A., Megre M., Savchenko E. // Proceedings of The 36th Conference on FRUCT Associacion. – 2024. –

P. 912–920. – URL: https://fruct.org/publications/volume-36/acm36/files/She.pdf?ysclid=m4qspqw9xv562608432

ELAN = Linguistic annotator. Version 4.9.3 / Hellwig B., van Uytvanck D., Hulsbosch M. et al. – URL: http://tla.mpi.nl/tools/tla-tools/elan/

Khan R., *Yurong Q.*, *Sajid N.* Extractive based text summarization using KMeans and TF-IDF // International Journal of Information Engineering and Electronic Business. – 2019. – Vol. 11. – P. 33–44.

Kherwa P., *Bansal P.* Topic modeling: a comprehensive review // ICST Transactions on Scalable Information Systems. – 2018. – Vol. 7(24). – P. 1–17.

Saputra D.M., *Saputra D.*, *Oswari L.D.* Effect of distance metrics in setermining K-Value in K-Means clustering using Elbow and Silhouette Method // Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019). – Atlantis Press, 2019. – P. 341–346. DOI: 10.2991/aisr.k.200424.051. Retrieved from: https://www.atlantis-press.com/proceedings/siconian-19/125939938

Sociolinguistic extension of the ORD corpus of Russian everyday speech / Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A. // Ronzhin A. et al. (eds.) SPECOM 2016, LNAI. – Springer, Switzerland, 2016. – Vol. 9811. DOI: https://doi.org/10.1007/978-3-319-43958-7_80

Text clustering using K-mean / Lal C., Ahmed A., Siyal R., Kumar S. // International Journal of Advanced Trends in Computer Science and Engineering. – 2021. – Vol. 10. – P. 2892–2897.

The ORD speech corpus of Russian everyday communication "One Speaker's Day": creation principles and annotation / Asinovsky A., Bogdanova N., Rusakova M., Stepanova S., Ryko A., Sherstinova T. // Text, Speech and Dialogue, LNCS. – Springer, Switzerland, 2009. – Vol. 5729. – P. 250–257. DOI: https://doi.org/10.1007/978-3-642-04208-9_36

References

Grudeva, E.V. (2008). Ways of introducing and defining a topic in the Russian language and the strategy of native speakers in its definition (an experimental study). *Proceedings of the Russian State Pedagogical University. A.I. Herzen*: *Social and human sciences* (*philosophy*, *linguistics*, *literary criticism*, *cultural studies*, *economics*, *law*, *history*, *sociology*, *pedagogy*, *psychology*): *Scientific journal*, *11*(*71*), 36–44.

Bogdanova-Beglaryan, N.V., Blinova, O.V., Martynenko, G.Ya., Sherstinova, T.Yu. (2019). Corpus of the Russian language of everyday communication "One day of speech" (ORD corpus): current state and prospects. In *Proceedings of the institute Russian language them. V.V. Vinogradov Russian Academy of Sciences* (pp. 100–110). Moscow

Kositsina, Yu.V. (2012). Current thematic development: coherent and cohesive aspects. *Bulletin of the Kemerovo State University*, *4*(*52*), 281–284.

Matveeva, T.V. (1990). Thematic deployment of spoken text. In *Linguistic appearance of the Ural city*: *a collection of scientific papers* (pp. 46–54). Sverdlovsk: Ural State University.

Mitrofanova, O.A. (2019). Study of the structural organization of a work of art using thematic modeling: experience of working with the text of the novel "The Master and

Margarita" by M.A. Bulgakov. In *Korpusnaya lingvistika – 2019. Proceedings of the International Conference* (pp. 387–394). Saint-Petersburg: Saint-Petersburg State University.

Sibiryakova, I.G. (1990). Experience in thematic analysis of a dialogic conversational text. In *Linguistic appearance of the Ural city: a collection of scientific papers* (pp. 61–71). Sverdlovsk: UrGU.

Skirdach, O.M. (1984). Dinamika razvitiya temy v tekste. In *Sbornik nauchnykh trudov MGPIIYa imeni Morisa Toreza*, 232: *Leksiko-frazeologicheskaya Sistema nemetsskogo yazyka i ee realizatsiya v tekste*, (pp. 149–165).

Sherstinova, T.Yu. (2013). Communicative macro episodes in the ORD corpus of Russian everyday communication: annotation principles and statistics. In *Proceedings of the International Conference* "*Corpus Linguistics – 2013*" (pp. 449–456). Saint-Petersburg.

Sherstinova, T.Yu., Kirina, M.A., Moskvina, A.D. (2024). Topic modeling of prose fiction: model assessment and interpretability (the case of Russian short stories of the 1900s–1930s). *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya. Tomsk State University Journal of Philology*, *89*, 127–151. DOI: 10.17223/19986645/89/6. Retrieved from: https://pureportal.spbu.ru/files/122949387/_89_.pdf?ysclid=m4quz23l4t639599053

Akinshina, E., Sherstinova, T. (2022). Thematic diversity of everyday Russian discourse: a case study based on the ORD corpus. *SPECOM 2022*, *LNCS*, 13721, 1–9.

Sherstinova, T., Firsanova, V., Novoseltseva, A., Megre, M., Savchenko, E. (2024). Automated speech act annotation in a Russian spoken corpus using large language models: a comparative study. In *Proceedings of The 36th Conference on FRUCT Associacion* (pp. 912–920). Retrieved from: https://fruct.org/publications/volume-36/acm36/files/She.pdf?ysclid=m4qspqw9xv562608432

Hellwig, B., van Uytvanck, D., Hulsbosch, M. et al. ELAN – Linguistic Annotator. Version 4.9.3. Retrieved from: http://tla.mpi.nl/tools/tla-tools/elan/

Khan, R., Yurong, Q., Sajid, N. (2019). Extractive based text summarization using KMeans and TF-IDF. *International Journal of Information Engineering and Electronic Business*, *11*, 33–44.

Kherwa, P., Bansal, P. (2018). Topic modeling: a comprehensive review. *ICST Transactions on Scalable Information Systems*, *7*(*24*), 1–17.

Saputra, D.M., Saputra, D., Oswari, L.D. (2019). Effect of distance metrics in determining K-value in K-means clustering using Elbow and Silhouette Method. In *Sriwijaya International Conference on Information Technology and its Applications* (*SICONIAN 2019*) (pp. 341–346). Atlantis Press. DOI: 10.2991/aisr.k.200424.051. Retrieved from: https://www.atlantis-press.com/proceedings/siconian-19/125939938

Sherstinova, T.Yu., Kirina, M.A., Moskvina, A.D. (2024). Topic modeling of prose fiction: model assessment and interpretability (the case of Russian short stories of the 1900s–1930s). *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya*, *89*, 127–151.

Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Ermolova, O., Baeva, E., Martynenko, G., Ryko, A. (2016). Sociolinguistic extension of the ORD corpus of Russian everyday speech. *SPECOM 2016, LNAI*, 9811. DOI: https://doi.org/10.1007/978-3-319-43958-7_80

Lal, C., Ahmed, A., Siyal, R., Kumar, S. (2021). Text clustering using K-mean. *International Journal of Advanced Trends in Computer Science and Engineering*, *10*, 2892–2897.

Asinovsky, A., Bogdanova, N., Rusakova, M., Stepanova, S., Ryko, A., Sherstinova, T. (2009). The ORD speech corpus of Russian everyday communication "One Speaker's Day": Creation Principles and Annotation. *Text*, *Speech and Dialogue*, *LNCS*, *5729*, 250–257. DOI: https://doi.org/10.1007/978-3-642-04208-9_36

Об авторах

Шерстинова Татьяна Юрьевна – кандидат филологических наук, доцент департамента филологии, Национальный исследовательский университет «Высшая Школа Экономики», Россия, Санкт-Петербург, tsherstinova@hse.ru

Вепринцева Дарья Александровна – магистрант, Санкт-Петербургский государственный университет, Россия, Санкт-Петербург, daveprintseva@edu.hse.ru

About the authors

Sherstinova Tatiana Yurievna – PhD in Philology, Associate Professor of the Department of Philology, National Research University "Higher School of Economics", Russia, Saint-Petersburg, tsherstinova@hse.ru

Veprintseva Daria Aleksandrovna – Magister Student, Saint Petersburg State University, Russia, Saint-Petersburg, daveprintseva@edu.hse.ru