Published in the Russian Federation
Oriental Studies (Previous Name: Bulletin of the Kalmyk Institute for Humanities of the Russian Academy of Sciences)
Has been issued as a journal since 2008
ISSN: 2619-0990; E-ISSN: 2619-1008
Vol. 18, Is. 1, Pp. 247–271, 2025
Journal homepage: https://kigiran.elpub.ru

УДК / UDC 81’25:81’42
DOI: 10.22162/2619-0990-2025-77-1-247-271

Сопоставительный анализ вероятностных тематических моделей китайско-русского корпуса политических текстов

Чжу Хуэй[1], *Митрофанова Ольга Александровна*[2]

[1] Даляньский университет иностранных языков (д. 6, западный участок Южной дороги Люйшунь, район Луйшунькоу, 116044 Ляонин, Далянь, Китайская Народная Республика)
аспирант
0009-0003-2922-8156. E-mail: : zhuhui1230[at]qq.com

[2] Санкт-Петербургский государственный университет (д. 7–9, Университетская наб., 199034 Санкт-Петербург, Российская Федерация)
кандидат филологических наук, доцент
0000-0002-3008-5514. E-mail: o.mitrofanova[at]spbu.ru

Аннотация. *Введение*. Статья посвящена сопоставительному анализу вероятностных тематических моделей китайско-русского корпуса параллельных и сопоставимых текстов политической тематики. Разработанный в рамках исследования корпусной ресурс включает в себя три подкорпуса: исходные тексты «Докладов о работе правительства в 2012–2022 гг.» на китайском языке, их переводы на русский язык и сопоставимый подкорпус «Послания Президента Российской Федерации Федеральному Собранию РФ 2011–2021 гг.». *Цель* экспериментов заключается в выявлении и описании общих тем для корпуса, а также тем, специфичных для отдельных текстов. Осуществляется лингвистическая интерпретация тем с помощью генерации меток тем большой языковой моделью YandexGPT, полученные метки тем сопоставляются с результатами экспертной разметки и выделения ключевых выражений. Эксперименты по вероятностному тематическому моделированию проводятся на основе алгоритма LDA с помощью инструмента TMT (Topic Modeling Tool), а для выделения ключевых выражений используется алгоритмы YAKE, mBERT и TF-IDF в библиотеке Orange. В результате были выявлены сходства и различия между словами-тематизаторами в темах подкорпусов, построено семейство вероятностных тематических моделей, описывающих семантическую организацию китайско-русского корпуса параллельных и сопоставимых текстов политической тематики. *Результаты* тематического моделирования были сопоставлены с данными, полученными в ходе автоматического выделения ключевых выражений, и было показано пересечение между наборами слов-тематизаторов и наборами ключевых выражений, сформированными для каждого из подкорпусов. В нашем исследовании также описывается частеречная характеристика слов-тематизаторов в темах. Обнаружено, что тематические модели воспроизводят основные парадигматические и синтагматические отношения в корпусе текстов. Новизна нашего исследования состоит в том, что в ней впервые представлены результаты автоматического постро-

ения тематических моделей для китайско-русского корпуса, что восполняет существующие пробелы в этой области.
Ключевые слова: вероятностная тематическая модель, корпус текстов, параллельный корпус, сопоставимый корпус, политические тексты, автоматическое выделение ключевых выражений, частеречная разметка
Благодарность. Исследование выполнено Чжу Хуэем при поддержке проекта ДУИЯ № YJSCX2022-009; раздел 4 подготовлен О. А. Митрофановой при поддержке проекта СПбГУ № 123042000068-8; раздел 6 подготовлен О. А. Митрофановой при поддержке проекта СПбГУ № 124032900006-1.
Для цитирования: Чжу Хуэй, Митрофанова О. А. Сопоставительный анализ вероятностных тематических моделей китайско-русского корпуса политических текстов // Oriental Studies. 2025. Т. 18. № 1. С. 247–271. DOI: 10.22162/2619-0990-2025-77-1-247-271

Chinese-Russian Corpus of Political Texts: A Comparative Analysis of Probabilistic Topic Models

Zhu Hui[1], *Olga A. Mitrofanova*[2]

[1] Dalian University of Foreign Languages (6, Lushun Nanlu Xiduan, 116044 Dalian, Liaoning Province, People's Republic of China)

Postgraduate Student

0009-0003-2922-8156. E-mail: zhuhui1230[at]qq.com

[2] St. Petersburg State University (9/7, Universitetskaya Emb., 199034 St. Petersburg, Russian Federation)

Cand. Sc. (Philology), Associate Professor

0000-0002-3008-5514. E-mail: o.mitrofanova[at]spbu.ru

Abstract. *Introduction*. The article introduces a comparative analysis of probabilistic topic models derived from a Chinese-Russian corpus of parallel and comparable political texts. The corpus developed hereto includes a total of three sub-corpora: Reports on the Work of the Government in 2012–2022 (original Chinese-language texts), their Russian translations, and Presidential Addresses to the Federal Assembly of Russia in 2011–2021 (a comparable Russian-language sub-corpus). *Goals*. The work aims at identifying and describing topics that prove common within the corpus, as well as ones specific to individual texts. Linguistic interpretations have been conducted with topic labeling tools of the YandexGPT language model, the resulting topic labels be further compared to expert-generated annotations and automatically extracted keyphrases. The conducted probabilistic topic modelling involves the LDA algorithm in TMT (Topic Modeling Tool), as well as the YAKE, mBERT, and TF-IDF algorithms from Orange library for Python. The algorithms are intended to identify keyphrases and find out similarities in topical words across different sub-corpora and between the languages under comparison. *Results*. So, a family of probabilistic topic models that describe semantic organization of the Chinese-Russian parallel and comparable corpus of political texts has been created. The outcomes of our topic modelling are compared to the automatically extracted keyphrases, and reveal certain intersections for each sub-corpus. The study also provides a part-of-speech (POS) tagging analysis of topical words. As is shown, the models reproduce key paradigmatic and syntagmatic relationships in the text corpus. The research is first to present automatically constructed probabilistic topic models for a Chinese-Russian parallel and comparable corpus of political texts, thus filling in some gaps existing in this field.
Keywords: probabilistic topic modelling, text corpus, parallel corpus, comparative corpus, political texts, automatic keyphrase extraction, POS tagging
Acknowledgements. The study was performed by Zhu Hui with the support of the DUFL project No. YJSCX2022-009; section 4 was prepared by O. A. Mitrofanova with the support of St. Petersburg State University, project No. 123042000068-8; section 6 was prepared by O. A. Mitrofanova with the support of St. Petersburg State University, project No. 124032900006-1.

For citation: Zhu Hui, Mitrofanova O. A. Comparative Analysis of Probabilistic Topic Models of the Chinese-Russian Corpus of Political Texts. *Oriental Studies*. 2025; 18(1): 247–271. (In Russ.). DOI: 10.22162/2619-0990-2025-77-1-247-271.

1. Введение

В эпоху цифровых технологий анализа текстовой информации ученые, исследующие функционирование языка в обществе, сталкиваются с необходимостью совмещения традиционных методов лингвистического анализа и алгоритмов компьютерной лингвистики в программных лингвистических комплексах для автоматической обработки корпусов текстов. Процедуры, проводимые на материале корпусных данных, включают многоуровневую разметку корпусов, выделение именованных сущностей, фактов, количественной информации, ключевых слов и словосочетаний, автоматическую классификацию и кластеризацию лексики и документов в корпусе, генерацию заголовков и аннотаций, построение лексических баз данных и формальных онтологий и т. д. [Большакова и др. 2011; Большакова и др. 2017; ПиКЛ 2017].

Особое место в корпусной лингвистике занимают исследования, проводимые на материале параллельных и сопоставимых корпусов текстов [Захаров, Богданова 2020: 61–63]. Китайский язык представлен в ряде корпусных ресурсов, разработка которых ведется с конца XX в.: на сегодняшний день число корпусов текстов для китайского языка достигло нескольких десятков, среди них есть параллельные и сопоставимые корпусы текстов [Колпачкова 2015; Тао, Захаров 2015].

Исследования, проводимые на материале параллельных корпусов, в основном сосредоточены на китайско-английской языковой паре [Dalianis et al 2010; Tian et al. 2014; Zhai et al. 2020], в то же время есть ресурсы, включающие китайский и другие европейские [Cao 2020] и восточные языки [Zhang et al. 2020]. В переводоведении и синологии значительное внимание до сих пор уделялось лингвистическим методам работы с параллельными корпусами текстов политической тематики [Wang, Qin 2009; Li, Hu 2017]. Квантитативные исследования китайско-русских корпусов текстов проводятся не столь часто [Тао, Захаров 2015: 19].

Исследования на материале китайско-русских и русско-китайских параллельных корпусов проводятся с 2010 г. Одним из наиболее значимых ресурсов является русско-китайский подкорпус[1], который появился в 2016 г. в составе Национального корпуса русского языка ‘НКРЯ)[2]. На данный момент подкорпус объемом более 3,5 млн словоупотреблений включает в свой состав свыше 1 тыс. текстов, документов, большая часть которых представляет художественный, публицистический и официально-деловой стили. Вместе с тем китайские ученые разработали несколько китайско-русских параллельных корпусов, таких как многостилевой русско-китайский и китайско-русский параллельный корпус [Cui, Zhang 2014], русско-китайский параллельный корпус научных текстов гуманитарной области [Тао, Захаров 2015], русско-китайский переводческий корпус [Liu, Shao 2016] и другие. В настоящее время происходит становление корпусного направления в китаеведении, что подтверждается большим вниманием, которое уделяется таким вопросам, как создание, разметка и выравнивание параллельных корпусов [Тао, Захаров 2015; Дань 2015; Мухин, Ян 2016; Чэнь, Кукушкина 2018].

В нашем исследовании демонстрируется перспективность применения вероятностного тематического моделирования и процедур автоматического выделения ключевых слов в задачах извлечения информа-

[1] Русско-китайский подкорпус Национального корпуса русского языка [электронный ресурс] // URL: https://ruzhcorp.ruscorpora.ru/?ysclid=lusm9cnfc2229797829 (дата обращения: 15.06.2024).

[2] Национальный корпус русского языка [электронный ресурс] // URL: https://ruscorpora.ru/ (дата обращения: 15.06.2024).

ции из китайско-русских корпусов текстов. В работах [Sun et al. 2010; Huang et al. 2015] описаны результаты тематического моделирования в китайских текстах ‘см. также онлайн-ресурсы[1]), однако применение данных алгоритмов в китайско-русской языковой паре мы, к сожалению, не нашли.

Новизна нашего исследования состоит в том, что в нем впервые представлены результаты автоматического построения тематических моделей для китайско-русского корпуса. Особенностью нашего подхода является то, что эксперименты проводятся на материале принципиально нового типа корпуса, включающего в свой состав как параллельные, так и сопоставимые тексты политической тематики. Цель исследования заключается в том, чтобы в ходе экспериментов выявить темы, общие для корпуса и специфичные для отдельных текстов, осуществить лингвистическую интерпретацию тем, сопоставить их с экспертной разметкой и результатами выделения ключевых выражений.

Гипотеза исследования такова: единство стиля и жанра корпусов текстов является одним из факторов, обеспечивающих общность структуры тематических моделей, обученных на сравниваемых текстовых данных. Тематические модели, построенные с корректно подобранными параметрам, имеют высокую объяснительную способность, в них отражаются как сходства, так и различия между текстами. Наше предположение состоит в том, что различия в результатах тематического моделирования корпуса могут быть объяснены как языковой, так и общественно-политической спецификой источников.

2. Описание китайско-русского корпуса параллельных и сопоставимых текстов политической тематики

Наше исследование основано на материале китайско-русского корпуса параллельных и сопоставимых текстов политической тематики. Данный корпус состоит из трех сегментов: первые два сегмента — это параллельные тексты «Докладов о работе правительства в 2012–2022 гг.» (далее — ДРП), включающие в себя исходный китайский текст (ДРП-К), и перевод на русский язык (ДРП-Р). Корпус содержит политические документы, опубликованные на государственном сайте информационного агентства *Синьхуа*[2] и газеты *Жэньминь жибао*[3]. Ежегодный «Доклад о работе правительства» является важным политическим документом и имеет высокую государственную ценность. Перевод «Докладов о работе правительства» на русский язык был осуществлен и опубликован Научно-исследовательским институтом истории партии и литературы при ЦК КПК, что обеспечило авторитетность переводного текста. Третий сегмент — это сопоставимый корпус «Послания Президента Российской Федерации Федеральному Собранию РФ 2011–2021 гг.» (далее — ППР). «Послание Президента Российской Федерации Федеральному собранию» представляет собой высокую ценность как документ о государственной жизни России, поэтому он рассматривается нами как важный материал для лингвистического исследования.

В ходе подготовки китайско-русского корпуса к проведению процедур тематического моделирования тексты были сегментированы по годам, были определены объемы текстов в токенах для каждого сегмента. Результаты представлены на рис. 1.

Сравнение тематических моделей требует предварительной обработки корпусов текстов.

1. Токенизация. Для токенизации китайских текстов использовался инструмент *CorpusWordParser*[4], разработанный группой корпусной лингвистики Пекинского университета иностранных языков, а для токе-

[1] Например, Тематическое моделирование в китайских текстах [электронный ресурс]// URL: https://github.com/sdx0112/Chinese-Topic-Modelling/tree/main (дата обращения: 15.06.2024).

[2] Информационное агентство *Синьхуа* [электронный ресурс] // URL: https://russian.news.cn/index.htm (дата обращения: 15.06.2024).

[3] Газета *Жэньминь жибао* [электронный ресурс] // URL: http://russian.people.com.cn/ (дата обращения: 15.06.2024).

[4] *CorpusWordParser* [электронный ресурс] // URL: https://corpus.bfsu.edu.cn/TOOLS.htm (дата обращения: 15.06.2024).

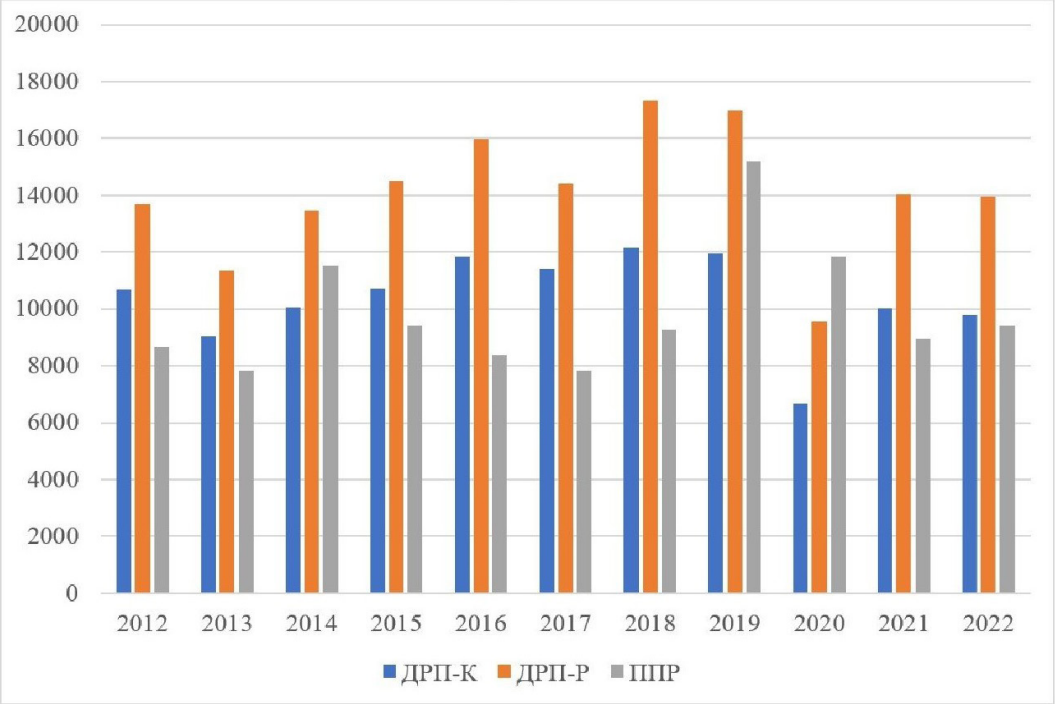

Рис. 1. Распределение текста китайско-русского корпуса по годам, объем в токенах
[*Fig. 1.* Texts of the Chinese-Russian corpus by year (tokens)]

низации русских текстов была применена функция .split(), в Python.

2. Лемматизация. При обработке корпуса на уровне лемматизации была использована библиотека для морфологического анализа spaCy[1], разработанная на языке программирования Python и включающая предобученные модели для различных языков, в том числе для русского и китайского.

3. Подготовка стоп-словарей. Стоп-словарь для русского языка был составлен на основе частотного словаря НКРЯ [Ляшевская, Шаров 2009]. Для работы с китайскими текстами мы использовали стоп-словарь из библиотеки stopwords-iso[2].

3. Тематическое моделирование как метод исследования в корпусной лингвистике

В современной компьютерной лингвистике среди процедур автоматического извлечения информации их корпусов текстов лингвистических задач важную роль играет тематическое моделирование, представляющее собой разновидность машинного обучения без учителя, а именно, нечеткой кластеризации [Воронцов 2023]. Тематическая модель — это один из способов представления семантической структуры корпуса с помощью набора тем, которые включают в себя близкие по значению слова-тематизаторы, характеризующиеся совместной встречаемостью в контекстах корпуса. Темы рассматриваются как скрытые переменные, определяющие внутренние связи в корпусе. Благодаря этому и темы, и слова-тематизаторы соотнесены с текстами корпуса. С вычислительной точки зрения в ходе тематического моделирования производится снижение размерности векторного-матричного представления корпуса в результате восстановления компонент смеси вероятностных распределений, лежащих в основе корпуса текстов. В лингвистическом аспекте тематическое моделирование предполагает семантическую интерпретацию тем, формируемых при обучении тематических моделей.

Тематические модели корпусов текстов могут быть построены различными способами [Daud et al. 2010], с привлечением алгебраических методов ‘латентно-семан-

[1] spaCy [электронный ресурс] // URL: https://spacy.io/ (дата обращения: 15.06.2024).

[1] Стоп-словарь из библиотеки stopwords-iso [электронный ресурс] // URL: https://github.com/stopwords-iso/stopwords-zh?tab=readme-ov-file (дата обращения: 15.06.2024).

тический анализ (Latent Semantic Analysis, LSA), неотрицательная матричная факторизация (Non-negative Matrix Factorization, NMF), и т. д.), и вероятностных (вероятностный латентно-семантический анализ (Probabilistic Latent Semantic Analysis, pLSA), латентное размещение Дирихле (Latent Dirichlet Allocation, LDA), модель размещения Патинко (Pachinko Allocation Model, PAM, тематическая модель на основе скрытых марковских цепей (Hidden Markov Models Topic Modelling, HMM-TM и т. д.). Для тематического моделирования корпус представляется в виде «мешка слов» (bag of words), при этом не учитывается порядок слов в текстах и порядок слов в корпусе. В результате матрично-векторных преобразований корпус представляется в виде матрицы «слова – тексты», которая в результате линейных преобразований может быть описана через произведение матриц меньшей размерности «слова – темы» и «темы – тексты». В вероятностных тематических моделях тема может быть описана с помощью дискретного распределения на множестве слов, текст — с помощью дискретным распределением на множестве тем. Тем самым корпус текстов формируется из слов, выбранных независимо и случайно из смеси распределений. Формальный способ задания тематической модели таков: наблюдаемая вероятность появления слова w в тексте d определяется как $p(w|d) = \Sigma p(t|d)\, p(w|t)$, где $p(w|d)$ — неизвестная вероятность появления слова w в теме t; $p(t|d)$ — неизвестная вероятность появления темы t в тексте d.

В лингвистической практике более всего востребованы мультимодальные тематические модели [Милкова 2019], в которых учтены различные характеристики корпусов текстов: разметка n-грамм ‘лексико-грамматических конструкций или коллокаций’, [Нокель, Лукашевич 2015; Седова, Митрофанова 2017], введение меток — слов или словосочетаний, обобщающих семантику тем [Mitrofanova et al. 2021], учет авторства [Mitrofanova et al. 2021; Mamaev, Mitrofanova 2020], времени создания текстов [Sherstinova et al. 2020; Митрофанова, Атугодаге 2023], установление иерархии тем, комбинация тематических моделей и моделей распределенных векторных вложений ‘Word2Vec, BERT и т. д.). Отдельной разновидностью мультимодальных тематических моделей являются многоязычные модели, позволяющие выделять параллельные темы в текстах корпуса [Mimno et al. 2009; Vulić, Moens 2012; Митрофанова 2016]. На материале китайско-русских корпусов подобные исследования ранее не проводились, поэтому наша работа восполняет существующий пробел.

4. Эксперимент по тематическому моделированию китайско-русского корпуса параллельных и сопоставимых текстов политической тематики

Из наборов доступных реализаций алгоритмов тематического моделирования в библиотеках MALLET, genism, scikit-learn, tomotopy, BigARTM, OCTIS, BERTopic и т. д. для нашего исследования мы выбрали инструмент Topic Modeling Tool (TMT)[1] [Newman et al. 2009], с помощью которого были построены тематические модели корпусов ДРП-К, ДРП-Р и ППР. Инструмент TMT представляет собой интерфейсное приложение, построенное на основе библиотеки MALLET и реализующее классический вариант алгоритма латентного размещения Дирихле (Latent Dirichlet Allocation, LDA). Выбор TMT обусловлен тем, что данный инструмент допускает обработку многоязычных данных, позволяет настроить кодировки для кириллицы, латиницы, иероглифики, предусматривает подключение стоп-словарей и фоновых корпусов на разных языках.

В ходе экспериментов предварительно обработанные корпусы текстов ДРП-К, ДРП-Р и ППР были импортированы в TMT. Обучение тематических моделей было проведено со следующими параметрами: число тем — 10, объем тем — 20 слов-тематизаторов, число итераций — 400. Подбор параметров производился эмпирически путем оценки перплексии обучения моделей (меры соответствия наблюдаемых

[1] Topic Modeling Tool [электронный ресурс] // URL: https://senderle.github.io/topic-modeling-tool/documentation/2017/01/06/quickstart.html (дата обращения: 15.06.2024).

результатов по отношению к расчетным', и когерентности тем (меры связности). Фильтрация корпусов производилась по пользовательским словарям стоп-слов для русского и китайского языков. В стоп-словари были включены символы латиницы, числовые обозначения, знаки препинания, сокращения, слова служебных частей речи и ряд других. В выдаче ТМТ темы приводятся в случайном порядке, слова-тематизаторы упорядочены по убыванию своего веса внутри тем.

Результаты тематического моделирования продемонстрированы в табл. 1, 2, 3, полное описание тематических моделей содержится в прил. 1. Полученные данные указывают на то, что в процессе обучения моделей инструмент ТМТ генерирует наборы близких по значению слов и формирует на их основе темы, характеризующие содержание корпуса. Чтобы эксплицитно представить связи внутри тематических моделей для подкорпусов, мы провели экспертную и автоматическую разметку тем с помощью меток, обобщающих их семантику.

Назначение меток тем необходимо для того, чтобы повысить интерпретируемость тематических моделей. Экспертная разметка проводится вручную специалистами в той или иной предметной области, в нашем случае разметка проводилась разработчиками корпуса. Из множества вариантов автоматического назначения меток тем, представленных методами генерации меток из выдачи поисковых систем, лексических баз данных, дистрибутивно-семантических моделей, моделей суммаризации и т. д. [Ерофеева, Митрофанова 2019; Mitrofanova et al. 2021], мы выбрали вариант генерации меток средствами большой языковой модели YandexGPT[1]. При взаимодействии с моделью использовались методы промпт-инжиниринга [Liu et al. 2021]. Запросы к модели формулировались с помощью промптов, включающих экспертные метки тем, например: «*Необходимо подобрать три слова или словосочетания, обобщающих значение группы слов, связанных со здравоохранением: медицинский, миллион, город, здравоохранение, населить, строительство, миллиард, пункт, жилищный, доступность, цифровой, здравоохранение, рост, продолжительность, рубль, демографический, предстоящий, стоимость, комплексный*». Здесь назначенная вручную метка темы «здравоохранение» является уточняющим компонентом промпта, а курсивом выделены слова-тематизаторы. Как правило, в качестве ответа на запрос к модели мы получали три связных словосочетания, содержащие в том числе и компоненты промптов, которые выделены полужирным шрифтом в столбце «Метки YandexGPT» табл. 1, 2, 3.

[1] YandexGPT [электронный ресурс] // URL: https://ya.ru/ai/gpt-3 (дата обращения: 15.06.2024).

5. Анализ результатов тематического моделирования

Основываясь на результатах тематического моделирования, мы обнаружили как сходства, так и различия в тематике текстов в корпусах ДРП-К, ДРП-Р и ППР, что указывает на существование общих тенденций в освещении общественно-политической и социально-экономический жизни Китая и России, а также на то, что рассматриваемые корпусы текстов имеют свою специфику.

Для всех трех корпусов, сравниваемых с точки зрения выявленных тем, общими являются вопросы экономического, научно-технического развития и инноваций, развития стран в целом и их регионов, развития общественной сферы, рынка, предприятий, международного сотрудничества, партнерских взаимоотношений и путей взаимодействия стран.

В корпусах ДРП-К и ДРП-Р на первый план выходят масштабные темы, связанные с экономической стабилизацией и с инновациями в стране.

К особенностям ДРП-К следует отнести то, что в текстах данного корпуса доминирует упоминание различных общественно-политических и экономических мер и реформ, делается акцент на таких вопросах, как партнерство и совместная работа на государственном уровне, укрепление благосостояния народа, обеспечение жильем, борьба с бедностью, жизнь села. В связи с проблематикой здравоохранения особо выделяется тема борьбы с эпидемией.

Таблица 1. Примеры тем из тематической модели корпуса ДРП-К
[*Table 1.* Reports on the Work of the Government (Chinese). Sample topics from a topic model]

№	Темы	Метки YandexGPT
1	稳 ‘стабильность’, 新 ‘новый’ 创新 ‘инновация’, 创业 ‘предпринимательство’, 治理 ‘управлять / управление’ 建设 ‘строить / строительство’, 微 ‘микро-’, 培育 ‘воспитывать / воспитание’, 加大 ‘наращивать / наращивание’, 互联网 ‘Интернет’, 化 ‘превращать’, 防治 ‘профилактика’, 更多 ‘многочисленный’, 准 ‘точный’, 就业 ‘трудоустройство’, 提升 ‘продвигать / продвижение’, 科技 ‘наука и техника’, 区间 ‘интервал’, 坚决 ‘решительно’, 强化 ‘усиливать / усиление’	Научно-технические инновации. Развитие науки и техники. Технологические преобразования.
2	经济 ‘экономика’ 推进 ‘продвигать / продвижение’, 政策 ‘политика’, 建设 ‘строить / строительство’, 社会 ‘общество’, 加快 ‘ускорять’, 实施 ‘претворять’, 服务 ‘услуга’, 创新 ‘инновация’, 提高 ‘повышать’, 支持 ‘поддерживать / поддержка’, 推动 ‘способствовать’, 就业 ‘трудоустройство’, 扩大 ‘расширять’, 国家 ‘государство’, 新 ‘новый’, 继续 ‘продолжать’, 坚持 ‘настаивать’, 产业 ‘производство’, 稳定 ‘стабильность / стабилизировать’	Экономическое развитие. Инновационная экономика. Меры государственной поддержки.
	...	...

Таблица 2. Примеры тем из тематической модели корпуса ДРП-Р
[*Table 2*. Reports on the Work of the Government (Russian). Sample topics from a topic model]

№	Темы	Метки YandexGPT
1	новый, механизм, область, реализация, технический, распределение, год, система, регион, продвигать, повышение, стратегический, производство, деятельность, усиливать, вид, единый, увеличение, достигнуть	Научно-техническое развитие. Стратегические инновации. Повышение технологического уровня.
2	развитие, местный, год, процент, главный, экономика, система, трансформация, страна, экономический, путь, обеспечение, регулирование, международный, процент, хороший, население, расти, ряд	Экономическое развитие. Трансформация системы. Международный путь.
	...	...

Таблица 3. Примеры тем из тематической модели корпуса ППР
[*Table 3*. Presidential Addresses to the Federal Assembly. Sample topics from a topic model]

№	Темы	Метки YandexGPT
1	научный, центр, технологический, исследовательский, передовой, запустить, подготовка, инфраструктура, средний, школьник, нацелить, цифровой, технология, рабочий, рост, поколение, банковский, бизнес, компания, искусственный	Научно-технологическое развитие. Подготовка кадров. Цифровая трансформация.

2	деловой, рост, рынок, налоговый, процент, бизнес, экспорт, просить, промышленный, иностранный, фонд, миллиард, свобода бизнеса, доход, производство, орган, производство, хозяйство	Экономическое развитие. Бизнес-процессы. Инвестиционная деятельность.
	...	...

Отличительной особенностью корпуса ДРП-Р является внимание к экономическим реформам, улучшению жизни страны, сложности экономической ситуации, формулируется созидательность, рациональность, инновационность, последовательность процессов в стране. В экономической сфере подчеркивается развитие и поддержка микропредприятий.

Наблюдаемая асимметрия тем оригинального корпуса на китайском языке и корпуса переводов на русский язык может быть объяснена типологическими различиями между двумя языками, а также невозможностью подбора точных переводных эквивалентов.

Темы, характерные для корпуса ППР, отличаются значительным разнообразием по сравнению с темами корпусов ДРП-К и ДРП-Р. В текстах корпуса ППР уделяется большее внимание инфраструктуре города по сравнению с селом, подчеркиваются вопросы подготовки кадров, поддержки семьи, здравоохранения, образования, прав и свобод граждан, социальных и некоммерческих проектов. Специфичными являются темы, связанные с военной тематикой, законодательной и судебной властью.

Общими для корпусов ППР и ДРП-К являются темы инвестиционной политики, в то время как корпусы ППР и ДРП-Р объединяют темы государственных структур и общественного контроля.

В ходе экспертной и автоматической разметки тем с точки зрения меток мы полагались на результаты, полученные в исследовании [Чжу, Захаров 2024: 125–126], где описаны восемь тематических классов применительно к рассматриваемым корпусам: 1), «Государственное управление и регулирование, инвестиции и инновации», 2), «Экономическое развитие», 3), «Военное дело и внешняя политика», 4), «Здоровье и экология», 5), «Социальная сфера и благосостояние населения», 6), «Торговля и международные отношения», 7), «Сельское хозяйство», 8), «Общество, образование, молодежь». Тематическое моделирование позволило конкретизировать данные классы с учетом специфичности слов-тематизаторов, формирующих темы. Распределение меток тем по тематическим классам представлено в табл. 4. В столбце «Тематические классы» в скобках указано число меток тем, соответствующих каждому классу в корпусах. В столбце «Метки тем» в скобках приводится число меток, обозначающих темы в каждом из трех корпусов.

Анализ результатов классификации меток тем ChatGPT показывает, что с точки зрения детализации объема и содержания понятий доминирующими являются три тематических класса: 1), «Государственное управление и регулирование, инвестиции и инновации» (33 метки), 2), «Экономическое развитие» (22 метки), 5), «Социальная сфера и благосостояние населения» (12 меток). Остальные пять классов являются более частными и включают в себя от двух до семи меток, при этом темы могут быть представлены не во всех трех рассматриваемых корпусах: например, в корпусе ППР явно не отражена тема «сельское хозяйство», в корпусах ДРП-К и ДРП-Р отсутствуют темы, связанные с военным делом. Полученные результаты могут быть представлены в виде схемы тематической разметки и в дальнейшем применены для рубрикации документов корпусов общественно-политических текстов как на китайском, так и на русском языках.

6. Сравнение наборов ключевых выражений и слов-тематизаторов в китайско-русском корпусе параллельных и сопоставимых текстов политической тематики

Организация экспериментов, представленных в данной статье, позволила провести

Таблица 4. Сравнение тематических классов и меток тем для ДРП-К, ДРП-Р и ППР
[*Table 4*. Comparing topic categories and labels for RWG-C, RWG-R and PAFA]

№	Тематические классы	Метки тем
1	Государственное управление и регулирование, инвестиции и инновации (33)	**ДРП-К (10)**: меры государственной поддержки, реформы в Китае, региональное развитие, научно-технические инновации, развитие науки и техники, технологические преобразования, инвестиционная деятельность, жилищная политика государства, экономическая политика, борьба с бедностью. **ДРП-Р (13)**: инновационное развитие, последовательная интенсификация, научно-техническое развитие, стратегические инновации, повышение технологического уровня, рациональное развитие, инновационная сфера, налоговое регулирование, государственный контроль, развитие страны, правительство и программа, развитие и реформа, управление и работа. **ППР (10)**: обеспечение прав и свобод граждан, научно-технологическое развитие, подготовка кадров, цифровая трансформация, инвестиционная деятельность, развитие страны, необходимые изменения, планы на будущее, законодательная и судебная власть, государственные институты.
2	Экономическое развитие (22)	**ДРП-К (7)**: экономическое развитие, инновационная экономика, экономическая стабильность, экономические преобразования, социально-экономическая поддержка, развитие рынка, социально-экономические преобразования. **ДРП-Р (10)**: экономические реформы, меры стимулирования, повышение качества, экономическая стабилизация, развитие микропредприятий, сложная экономическая ситуация, экономическое развитие, трансформация системы, экономический рост, реформа и стратегия. **ППР (5)**: социально-экономическое развитие региона, финансирование бюджетных организаций, экономическое развитие, бизнес-процессы, социально-экономическое развитие.
3	Военное дело и внешняя политика (5)	**ДРП-К** (0) **ДРП-Р** (0) **ППР (5)**: военная политика, угроза безопасности, ядерное оружие, военная техника, оборонительные системы.
4	Здоровье и экология (2)	**ДРП-К (1)**: меры по борьбе с эпидемией. **ДРП-Р** (0) **ППР (1)**: здравоохранение и социальная сфера.
5	Социальная сфера и благосостояние населения (12)	**ДРП-К (5)**: социальное обеспечение, меры социальной поддержки, укрепление благосостояния народа, реформы в жилищной сфере, меры правительства по улучшению жилищных условий. **ДРП-Р (3)**: улучшение жизни, создание и улучшение, общественное развитие. **ППР (4)**: развитие общественной сферы, социальное обеспечение, поддержка семей с детьми, меры социальной поддержки.

6	Торговля и международные отношения (5)	**ДРП-К (3)**: международное сотрудничество, партнерские отношения, совместная работа. **ДРП-Р (1)**: международный путь. **ППР (1)**: международные отношения.
7	Сельское хозяйство (4)	**ДРП-К (3)**: реформы в сельском хозяйстве, развитие сельских территорий, культурная и экономическая жизнь села. **ДРП-Р (1)**: производство и система. **ППР** (0)
8	Общество, образование, молодежь (7)	**ДРП-К (1)**: реформы в сфере образования. **ДРП-Р (2)**: высококачественные услуги, поддержание стабильности. **ППР (4)**: некоммерческие организации, социальные проекты, инфраструктура города, общественно-политическая сфера.

сравнение результатов тематического моделирования и выделения ключевых выражений в китайско-русском корпусе параллельных и сопоставимых текстов политической тематики. Возможность сравнения обеспечивается тем, что как тематические модели, так и наборы ключевых выражений представляют семантическую свертку исходных текстов [Чжу, Захаров 2024]. В данном исследовании была выдвинута гипотеза о том, что ключевые выражения, понимаемые как «текстовые единицы разной структуры ‘как слова, так и словосочетания), способные в сжатом виде представить основные компоненты семантической структуры текста» [Гусева, Митрофанова 2024: 23], могут совпадать со словами-тематизаторами, являющимися наиболее значимыми для определения тематики текстов корпуса. Для выделения ключевых выражений в русскоязычных подкорпусах использовался инструмент SketchEngine[1], для работы и с китайским, и с русскими текстами было принято решение применить статистический алгоритм TF-IDF, гибридный алгоритм YAKE [Campos et al. 2020] и алгоритм на основе многоязычной модели семейства Трансформер mBERT [Wu, Dredze 2020] в библиотеке для автоматической обработки текста Orange[2]. Стоит отметить, что списки ключевых выражений, полученные с помощью указанных алгоритмов, различаются: алгоритм TF-IDF дает преимущество словам, специфичным для отдельных документов корпуса, алгоритм YAKE присваивает больший вес словам и словосочетаниям, тяготеющим к началу текста, часто встречающимся в разных предложениях, специфичным для того или иного контекста и т. д. Алгоритм, опирающийся на модель mBERT, выбирает в качестве ключевых выражений слова или словосочетания, вектора которых близки к вектору текста. Наборы ключевых выражений, полученные с помощью алгоритмов YAKE и mBERT для русскоязычных политических текстов, обладают высокой интерпретируемостью, в то время как для китайских текстов более информативны результаты алгоритмов YAKE и TF-IDF. В табл. 5 представлены наборы из 20 важнейших ключевых выражений для подкорпусов **ДРП-К**, **ДРП-Р** и **ППР**.

Во-первых, эксперименты показали, что наборы ключевых слов отражают разные аспекты общественной жизни в Китае и России. Например, в ДРП-К ключевые слова, извлеченные алгоритмом YAKE, отражают политику и инициативы национального развития, такие как 小康社会 ‘общество средней зажиточности’, 改革开放 ‘политика реформы и открытости’, 现代化 ‘модернизация’ и т. д. В то же время семантика ключевых выражений, извлеченных алгоритмом TF-IDF, боль-

[1] SketchEngin [электронный ресурс] // URL: https://www.sketchengine.eu/ (дата обращения: 15.06.2024).

[2] Orange [электронный ресурс] // URL: https://orangedatamining.com/ (дата обращения: 15.06.2024).

Таблица 5. Наборы ключевых выражений для ДРП-К, ДРП-Р и ППР
[*Table 5*. Keyphrase sets for RWG-C, RWG-R and PAFA]

№	Набор ключевых выражений в ДРП-К	
	YAKE	**TF-IDF**
1	结构性 ‘структурность’	疫情 ‘эпидемия’
2	领导人 ‘лидер’	脱贫 ‘ликвидация бедности’
3	国务院 ‘Госсовет’	公里 ‘километр’
4	服务业 ‘сфера услуг’	危机 ‘кризис’
5	互联网 ‘Интернет’	降 ‘снизить’
6	小康社会 ‘общество средней зажиточности’	累计 ‘итог / суммировать’
7	市场化 ‘коммерциализация’	覆盖 ‘охватывать’
8	人民币 ‘китайский юань’	疫 ‘эпидемия’
9	人民群众 ‘народ’	年均 ‘среднегодовой’
10	改革开放 ‘политика реформы и открытости’	时间 ‘время’
11	党中央 ‘ЦК КПК’	十四五 ‘14-я пятилетка’
12	现代化 ‘модернизация’	物价 ‘цена товара’
13	社会主义 ‘социализм’	企 ‘предприятие’
14	总书记 ‘генеральный секретарь’	微型 ‘микро-’
15	养老金 ‘пенсия по старости’	侧 ‘предложение’
16	农民工 ‘рабочие из крестьян’	丝绸之路 ‘Великий шёлковый путь’
17	高水平 ‘высококачественный’	安居工程 ‘программа достойного жилья’
18	试验区 ‘пилотная зона’	小型 ‘малый тип’
19	进一步 ‘шаг за шагом’	碳 ‘углерод’
20	房地产 ‘недвижимость’	直达 ‘прямой’
№	Набор ключевых выражений в ДРП-Р	
	YAKE	**mBERT**
1	**страна**	трудоустройство
2	**экономика**	урбанизация
3	**экономический**	**проект**
4	**население**	роль
5	**уровень**	интернет
6	**реформа**	ликвидация
7	**система**	реконструкция
8	**сельский**	Китай
9	**работа**	**реализация**
10	**сфера**	предпринимательство
11	**предприятие**	потребление
12	**услуга**	модернизация
13	**новый**	**реформа**
14	**обеспечение**	**сельский**
15	**основный**	внешний
16	Китай	хозяйство
17	**рост**	практика

18	**мера**	сервис
19	**стимулировать**	научно-технический
20	**политика**	финансирование
№	Набор ключевых выражений в ППР	
	YAKE	**mBERT**
1	**регион**	**Россия**
2	свой	проект
3	гражданин	**регион**
4	**развитие**	конкуренция
5	**необходимый**	эксперимент
6	решение	интернет
7	**система**	**рубль**
8	нужный	**детский**
9	задача	сад
10	**работа**	гражданство
11	**правительство**	капитал
12	**страна**	**материнский**
13	**новый**	реализация
14	число	механизм
15	коллега	реальность
16	**сделать**	**свобода**
17	**государственный**	паспорт
18	создать	**здравоохранение**
19	вопрос	канал
20	говорить	поддержка

ше сосредоточена на социальных вопросах, которые тесно связаны с жизнью граждан, в частности, 物价 ‘цена товара’, 脱贫 ‘ликвидация бедности’ и 安居工程 ‘программа достойного жилья’. Для ДРП-Р и ППР в выдаче алгоритма YAKE преобладает общеполитическая лексика текстов, например: *обеспечение*, *Китай*, *политика (ДРП)*; *гражданин*, *решение*, *коллега (ППР)*, в то время как алгоритм mBERT выделил ключевые слова, связанные с конкретными темами. В ДРП-Р это лексика, связанная с трудоустройством, урбанизацией, сельским хозяйством, научно-техническим развитием, в ППР внимание акцентируется на темах регионального развития, поддержки детей и семьи, здравоохранения и т. д.

Во-вторых, после сравнения наборов ключевых выражений и слов-тематизаторов в подкорпусах мы обнаружили, что эти списки частично пересекаются. Ключевые выражения, совпадающие со словами-тематизаторами, отмечены в табл. 5 полужирным шрифтом. Преимущество тематического моделирования по сравнению с алгоритмами выделения ключевых выражений заключается в том, что оно предоставляет читателю больше информации о каждой теме, автоматическая кластеризация тем может оказать большую помощь в быстром поиске в корпусе.

7. Сравнение частеречных характеристик слов-тематизаторов в темах

На основе полученных данных в нашей работе было установлено, что наборы слов-тематизаторов, формирующих темы корпусов ДРП-К, ДРП-Р и ППР, существенно различаются по своим частеречным характеристикам. Используя в качестве инструмента морфологической разметки

Таблица 6. Статистические данные о частеречных характеристиках тем
[*Table 6*. Statistical data on POS tagging]

	ДРП-К		ДРП-Р		**ППР**	
Существительное / **NOUN**	**115**	**58,38 %**	98	49,49 %	106	52,48 %
Прилагательное / ADJ	15	7,61%	51	25,76 %	**59**	**29,21 %**
Глагол / VERB	**54**	**27,41%**	38	19,19 %	20	9,90 %
Наречие / ADV	**13**	**6,60 %**	6	3,03 %	6	2,97 %
Иные части речи	**0**	**0**	5	2,53 %	11	5,44 %

текстов библиотеку *SpaCy*, мы определили соотношение слов различных частей речи в темах исследуемых корпусов ‘см. табл. 6.). Следует отметить, что китайский язык отличается от русского высоким уровнем морфологической неоднозначности [Колпачкова 2015: 7; Тао, Захаров 2015: 80], прежде всего, омоформии, под которой обычно понимается совпадение отдельных форм слов, относящихся к разным словоизменительным парадигмам (например, необходимо — предикативное наречие или краткое прилагательное в форме среднего рода), / совпадающих в пределах одной парадигмы (например, регион — существительное в форме именительного / винительного падежей). Разрешение неоднозначности на морфологическом уровне представляет большую проблему, которая в компьютерной лингвистике решается различными способами: с использованием контекстных правил, с привлечением дистрибутивно-статистических методов и моделей машинного обучения [Manning, Schütze 2000; Захаров, Богданова 2020: 43]. Рассмотрим случаи омоформии в составе тем. К примеру, слово 发展 в китайском языке может быть как существительным, так и глаголом, и, соответственно, при переводе на русский язык для него будет два варианта разметки: 1), существительное развитие; 2), глагол развить/развивать. В корпусе ДРП-К много слов такого типа, как 支持 ‘поддерживать / поддержка’, 强化 ‘усиливать / усиление’, 累计 ‘итог / суммировать’ и т. д. Всего во множестве слов-тематизаторов ДРП-К потенциально неоднозначными оказываются 36 вхождений лексических единиц, из них 33 случая приходятся на омоформию существительного и глагола (см. выше), два — на омоформию прилагательного и существительного (в частности 基本 ‘основной / основа’), один — на омоформию прилагательного и наречия ‘ 积极 ‘активный / активно’). Это наблюдение следует учитывать при анализе результатов.

Полученные данные указывают на то, что с точки зрения частеречной принадлежности слов-тематизаторов в тематических моделях ДРП-К, ДРП-Р и ППР преобладают существительные, прилагательные и глаголы. Данная закономерность характерна для тематических моделей, которые обучаются на корпусах текстов без фильтрации словаря по частям речи: соотношение номинативных, глагольных и адъективных тем для русского языка описано в исследовании [Кольцов и др. 2014: 138–139]. На материале китайского языка данные такого типа были получены впервые, поэтому их анализ представляется нам очень важным.

Как показано в табл. 6, состав тем, выделенных в китайском корпусе ДРП-К, по своим количественным характеристикам существенно отличается от того, что было получено для русскоязычных корпусов. Так, корпус ДРП-Р, содержащий переводы с китайского, и корпус ППР, включающий оригинальные тексты на русском языке, близки друг к другу по частеречным характеристикам, что можно объяснить морфологической спецификой русского языка. Корпус ДРП-К отличается самыми высокими значениями частоты существительных и глаголов среди слов-тематизаторов. Больше всего прилагательных обнаруживается в темах корпуса ППР, тогда как в китайском корпусе ДРП-К их, наоборот меньше всего. Кроме того, было обнаружено, что в темах корпуса ДРП-К доля наречий выше, чем в ДРП-Р и ППР. Корпус ДРП-Р, занимающий промежуточное положение между ДРП-К и ППР, демонстрирует взаимовлияние систем китайского и русского языков в процессе перевода.

Сделанные нами наблюдения позволяют сформулировать гипотезу о том, что смещение частотности существительных, прилагательных и глаголов в корпусе ДРК-Р по сравнению с оригинальными китайскими (ДРК-К), и русскоязычными (ППР), текстами является специфической чертой переводных политических текстов. Для проверки этой гипотезы мы использовали статистические данные лингвистического процессора *Sketch Engine* [Чжу, Захаров 2024: 119] и сравнили частотность слов различных частей речи не только в тематических моделях, но и в полных текстах корпусов ДРП-К и ДРП-Р. Результаты анализа представлены в табл. 7.

Сходство частеречных характеристик тематических моделей и корпусов ДРП-Р и ППР заключается в том, что существительные, прилагательные и глаголы занимают доминирующие позиции. Распределение глаголов в корпусах и моделях более специфично. Для ДРП-Р доля глаголов в темах выше, чем в корпусе в целом, в то время как для корпуса ППР наблюдается обратная ситуация. В части служебных частей речи закономерно сокращение их долей в темах по сравнению с корпусами вследствие фильтрации по стоп-словарю.

Тем не менее доли трех основных знаменательных частей речи в тематических моделях выше, чем в корпусах текстов: это можно объяснить тем, что тематическая модель воспроизводит основные парадигматические и синтагматические отношения в корпусе текстов, при этом «усиливает» связи, наиболее важные с точки зрения структуры и содержания текстов. Глаголы, прилагательные и существительные составляют ядро предложения как синтаксически связной единицы текста, и сохранение этих частеречных классов при переходе от корпуса к тематической модели воспроизводит в ней формальную и содержательную связность исходного текста. Для тематических моделей связность тем — это один из критериев оценки качества (а именно, когерентности). В конечном итоге, сохранение соотношения основных частей речи в тематической модели — показатель ее состоятельности по отношению к корпусу текстов.

8. Заключение

В результате проведенного исследования было построено семейство вероятностных тематических моделей, описывающих семантическую организацию китайско-русского корпуса параллельных и сопоставимых текстов политической тематики. Полученные данные подтвердили высокий потенциал тематического моделирования в задаче сопоставительного анализа многоязычных текстовых коллекций.

Описание наборов слов-тематизаторов внутри тем позволило подтвердить гипотезу о том, что единство стиля и жанра корпусов текстов является одним из факторов, обеспечивающих общность структуры тематических моделей. В ходе лингвистического анализа данных были выявлены сходства и различия между темами подкорпусов ДРП-К, ДРП-Р и ППР, а также установлено, что расширение тематических моделей с помощью автоматического назначения меток тем повышает интерпретируемость моделей.

Результаты тематического моделирования были сопоставлены с данными, полу-

Таблица 7. Сравнение частеречных характеристик корпусов ДРП-Р и ППР с их тематическими моделями
[*Table 7*. Comparing POS tagging results for RWG-C, RWG-R and PAFA to corresponding topic models]

Части речи	**ДРП-Р**		**ППР**	
	Доля в корпусе **(%)**	Доля в тематической модели **(%)**	Доля в корпусе **(%)**	Доля в тематической модели **(%)**
Существительное / **NOUN**	41,60	49,49	32,63	52,48
Прилагательное / **ADJ**	18,02	25,76	14,28	29,21
Глагол / VERB	12,71	19,19	13,34	10,40
Наречие / ADV	3,79	3,03	5,03	2,90
Иные части речи	23,88	2,53	34,72	5,01

ченными в ходе автоматического выделения ключевых выражений. Была проверена гипотеза о частичном пересечении наборов слов-тематизаторов и ключевых выражений. Описание частеречных характеристик слов-тематизаторов в темах доказывает, что тематические модели воспроизводят основные парадигматические и синтагматические отношения в корпусе текстов.

Перспективы развития исследования связаны с расширением корпусных данных и с оценкой перспектив автоматического моделирования тематики корпусов текстов применительно к различным языковым парам.

Литература

Большакова и др. 2011 — Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В. М.: МИЭМ, 2011. 272 с.

Большакова и др. 2017 — Автоматическая обработка текстов на естественном языке и анализ данных : учеб. пособие / Большакова Е. И., Воронцов К. В., Ефремова Н. Э., Клышинский Э. С., Лукашевич Н. В., Сапин А. С. М.: НИУ ВШЭ, 2017. 269 с.

Воронцов 2023 — *Воронцов К. В.* Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM [электронный ресурс] // URL: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf (дата обращения 12.04.2024).

Гусева, Митрофанова 2024 — *Гусева Д. Д., Митрофанова О. А.* Ключевые выражения в русскоязычных научно-популярных текстах: сравнение восприятия устной и письменной речи с результатами автоматического анализа // Terra Linguistica. 2024. Т. 15. № 1. С. 20–35.

Дань 2015 — *Дань Н*а. Русско-китайский параллельный корпус в теории и практике перевода // Университетские чтения – 2015: Материалы научно-методических чтений ПГЛУ, Часть 6. Пятигорск: Пятигорский государственный лингвистический университет, 2015. С. 204–208.

Ерофеева, Митрофанова 2019 — *Ерофеева А. Р., Митрофанова О. А.* Автоматическое назначение меток тем в тематических моделях русскоязычных корпусов текстов // Структурная и прикладная лингвистика. СПб.: , 2019. С. 122–147.

Захаров, Богданова 2020 — *Захаров В. П., Богданова С. Ю.* Корпусная лингвистика. СПб.: СПбГУ, 2020. 234 с.

References

Bolshakova et al. Natural Language Processing and Computational Linguistics. Coursebook. Moscow: Moscow Institute of Electronics and Mathematics, 2011. 272 p. (In Russ.)

Bolshakova et al. Natural Language Processing and Data Analysis. Coursebook. Moscow: HSE University, 2017. 269 p. (In Russ.)

Vorontsov K. V. Probabilistic Topic Modeling: ARTM Regularization Theory and the BigARTM Open Source Library. On: MachineLearning.ru. 2023. Available at: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf (accessed: 12 April 2024). (In Russ.)

Guseva D. D., Mitrofanova O. A. Key phrases in Russian-language popular science texts: Comparison of oral and written speech perception with the results of automatic analysis. *Terra Linguistica*. 2024. Vol. 15. № 1. Pp. 20–35. (In Russ.) DOI: 10.18721/JHSS.15102

Dan Na. Russian-Chinese parallel corpora in the theory and practice of translation. In: PGLU University Readings – 2015. Proceedings. Vol. 6. Pyatigorsk: PGLU University, 2015. Pp. 204–208. (In Russ.)

Erofeeva A.R., Mitrofanova O.A. Automatic assignment of topic labels in topic models for Russian text corpora. In: Nikolaev I. S. (ed.) Structural and Applied Linguistics. Vol. 12. St. Petersburg: St. Petersburg University, 2019. Pp. 122–147. (In Russ.)

Zakharov V. P., Bogdanova S. Y. Corpus Linguistics. St. Petersburg: St. Petersburg University, 2020. 234 p. (In Russ.)

Колпачкова 2015 — *Колпачкова Е. Н.* Корпусы китайского языка: современное состояние и основные проблемы // Труды международной конференции «Корпусная лингвистика – 2015». СПб.: СПбГУ, 2015. С. 278–286.

Kolpachkova E. N. Chinese language corpora: An overview and major problems. In: Corpus Linguistics – 2015. Conference proceedings. St. Petersburg, 2015. Pp. 278–286. (In Russ.)

Кольцов и др. 2014 — *Кольцов С. Н., Кольцова О. Ю., Митрофанова О. А., Шиморина А. С.* Интерпретация семантических связей в текстах русскоязычного сегмента Живого Журнала на основе тематической модели LDA // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Мат-лы XVII Всеросс. объединенной конф. «Интернет и современное общество» IMS–2014, СПб.: СПбГУ, 2014. С. 135–142.

Koltsov S. N., Koltsova O. Ju., Mitrofanova O. A., Shimorina A. S. Interpretation of semantic relations in the texts of the Russian LiveJournal segment based on LDA topic model. In: Information Society Technologies in Science, Education and Culture. Conference proceedings (Internet and Modern Society). St. Petersburg, 2014. Pp. 135–142. (In Russ.)

Ляшевская, Шаров 2009 — *Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка ‘на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. 1087 с.

Lyashevskaya O. N., Sharov S. A. Frequency Dictionary of Modern Russian: [Analyzing] the Russian National Corpus. Moscow: Azbukovnik, 2009. 1087 p. (In Russ.)

Милкова 2019 — *Милкова М. А.* Тематические модели как инструмент «дальнего чтения» // Цифровая экономика. № 1 ‘5). 2019. С. 57–70.

Milkova M. A. Topic models as a tool for “long distance reading”. *Digital Economy*. 2019. No. 1 (5). Pp. 57–70. (In Russ.) DOI: 10.34706/DE-2019-01-06

Митрофанова, Атугодаге 2023 — *Митрофанова О. А., Атугодаге М. М.* Динамическое тематическое моделирование русскоязычного корпуса юридических документов // Terra Linguistica. 2023. Т. 14. № 1. С. 70–87.

Mitrofanova O. A., Athugodage M. M. Dynamic topic modelling of the Russian legal text corpus. *Terra Linguistica*. 2023. Vol. 14. No. 1. Pp. 70–87. (In Russ.) DOI: 10.18721/JHSS.14107

Митрофанова 2016 — *Митрофанова О. А.* Возможности использования параллельных и сопоставимых текстов в построении тематических моделей корпусов // Прикладная лингвистика в науке и образовании: ALPAC Report – полвека после разгрома: труды VIII Междунар. науч. конф. СПб: РГПУ им. А. И. Герцена, 2016. С. 194–199.

Mitrofanova O. A. Possibilities of parallel and comparable texts in building thematic models of corpora. In: Applied Linguistics in Science and Education. ALPAC Report Half a Century after the Destruction. Conference proceedings. St. Petersburg: Herzen University, 2016. Pp. 194–199. (In Russ.)

Мухин, Ян 2016 — *Мухин М. Ю., Ян И.* Проект создания китайско-русского параллельного корпуса официально-деловых текстов с дискурсивно-структурной разметкой // Вестник Южно-Уральского государственного университета. Сер.: Лингвистика. 2016. Т. 13. № 4. С. 23–31.

Mukhin M.Y., Yang Y. Building a Chinese-Russian parallel discourse structure corpus of official texts. *Bulletin of the South Ural State University. Ser. Linguistics*. 2016. Vol. 13. No. 4. Pp. 23–31. (In Russ.) DOI: 10.14529/ling160404

Нокель, Лукашевич 2015 — *Нокель М. А., Лукашевич Н. В.* Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами // Вычислительные методы и программирование. 2015. Т. 16. Вып. 2. С. 215–234.

Nokel M. A., Loukashevich N. V. Topic models: Adding bigrams and taking account of the similarity between unigrams and bigrams. *Numerical Methods and Programming*. 2015. Vol. 16. No. 2. Pp. 215–234. (In Russ.) DOI: 10.26089/NumMet.v16r222

ПиКЛ 2017 — Прикладная и компьютерная лингвистика / Николаев И. С., Митренина О. В., Ландо Т. М. ‘ред.). М.: URSS, 2016. 320 с.

Седова, Митрофанова 2017 — *Седова А. Г., Митрофанова О. А.* Тематическое моделирование русскоязычных текстов с опорой на леммы и лексические конструкции // Компьютерная лингвистика и вычислительные онтологии: Труды XX Междунар. Объединенной науч. конф. «Интернет и современное общество». СПб.: ИТМО, 2017. С. 132−143.

Тао, Захаров 2015 — *Тао Юань., Захаров В. П.* Разработка и использование параллельного корпуса русского и китайского языков // Научно-техническая информация Сер. 2: Информационные процессы и системы. 2015. № 4. С. 18–29.

Чжу, Захаров 2024 — *Чжу Хуэй, Захаров В. П.* Корпусное сравнение языка китайских и российских политических текстов // Политическая лингвистика. 2024. № 1 ‘103). С. 115–128.

Чэнь, Кукушкина 2018 — *Чэнь Сяохуэй, Кукушкина О. В.* О параллельных корпусах русских и китайских текстов // Вестник Московского университета. Сер. 9: Филология. 2018. №2. С. 170–197.

Campos et al. 2020 — *Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A*. YAKE! Keyword Extraction from Single Documents using Multiple Local Features // Information Sciences. № 509. Pp. 257–289. DOI: 10.1016/j.ins.2019.09.013

Cao 2020 — *Cao S. Y.* How does discourse affect Spanish-Chinese Translation? A case study based on a Spanish-Chinese parallel corpus. In Proceedings of the First Workshop on Computational Approaches to Discourse, Pp. 1–10, Online. Association for Computational Linguistics [электронный ресурс] // URL: https://aclanthology.org/2020.codi-1.1 (дата обращения: 15.06.2024).

Cui, Zhang 2014 — *Cui W., Zhang L*. Research on Parallel Corpus of Russian-Chinese Translation and Its Application // Journal of PLA University of Foreign Languages. 2014. № 1. Pp. 81–87. (In Chin.)

Dalianis et al. 2010 — *Dalianis H., Xing Haochun., Zhang X.* Creating a Reusable English-Chinese Parallel Corpus for Bilingual Dictionary Construction // Proceedings of the Seventh International Conference on Language Resources and Evaluation ‘LREC’10), Valletta, Malta. European Language Resources Association ‘ELRA). 2010. Pp. 1700–1705.

Nikolaev I. S., Mitrenina O. V., Lando T. M. (eds.) Applied and Computational Linguistics. Moscow: URSS, 2016. 320 p. (In Russ.)

Sedova A. G., Mitrofanova O. A. Topic modelling of Russian texts based on lemmata and lexical constructions. In: Computational Linguistics and Ontology. Conference proceedings (Internet and Modern Society). St. Petersburg: ITMO University, 2017. Pp. 132–143. (In Russ.) DOI: 10.17586/2541-9781-2017-1-132-144

Tao Yuan, Zakharov V. P. Creation and use of a parallel Russian-Chinese corpus. *Nauchno-tekhnicheskaya informatsiya Ser. 2: Informatsionnye protsessy i sistemy*. 2015. No. 4. Pp. 18–29. (In Russ.)

Zhu Hui, Zakharov V. P. A corpus-based linguistic comparison of Chinese and Russian political texts. *Political Linguistics*. 2024. No. 1 (103). Pp. 115–128. (In Russ.)

Chen Xiaohui, Kukushkina O. V. The parallel corpora of Russian and Chinese texts. *Lomonosov Philology Journal*. 2018. No. 2. Pp. 170–197. (In Russ.)

Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*. 2019. No. 509. Pp. 257–289. (In Eng.) DOI: 10.1016/j.ins.2019.09.013

Cao S. Y. How does discourse affect Spanish-Chinese translation? A case study based on a Spanish-Chinese parallel corpus. In: First Workshop on Computational Approaches to Discourse. Proceedings [online edition]. 2020. Pp. 1–10. Available at: https://aclanthology.org/2020.codi-1.1 (accessed: 15 June 2024). (In Eng.)

Cui W., Zhang L. Research on parallel corpus of Russian-Chinese translation and its application. *Journal of PLA University of Foreign Languages*. 2014. No. 1. Pp. 81–87. (In Chin.)

Dalianis H., Xing H.-Ch., Zhang X. Creating a reusable English-Chinese parallel corpus for bilingual dictionary construction. In: Seventh International Conference on Language Resources and Evaluation (LREC’10). Proceedings. Valletta, Malta: European Language Resources Association (ELRA), 2010. Pp. 1700–1705. (In Eng.)

Daud et al. 2010 — *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // Proceedings of Frontiers of Computer Science in China, 2010. Pp. 280−301.

Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: A survey. *Frontiers of Computer Science in China*. 2010. Vol. 4. No. 2. Pp. 280−301. (In Eng.)

Huang et al. 2015 — *Huang X. L., Li X., Liu T. L., Chiu D., Zhu T. S., Zhang L.* Topic Model for Identifying Suicidal Ideation in Chinese Microblog. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. Shanghai, China, 2015. Pp. 553–562.

Huang X. L., Li X., Liu T.L., David Chiu, Zhu T. S., Zhang L. Topic model for identifying suicidal ideation in Chinese microblog. In: 29th Pacific Asia Conference on Language, Information and Computation. Proceedings. Shanghai, 2015. Pp. 553–562. (In Eng.)

Li, Hu 2017 — *Li X.Q., Hu K.B.* Keywords and Their Collocations in the English Translations of Chinese Government Work Reports // Foreign Language in China. 2017. № 6. Pp. 81–89. (In Chin.)

Li X. Q., Hu K. B. Keywords and their collocations in the English translations of Chinese government work reports. *Foreign Language in China*. 2017. No. 6. Pp. 81–89. (In Chin.)

Liu et al. 2021 — *Liu P. F., Yuan W. Z., Fu J. L., Jiang Z.B., Hayashi H., Neubig G.* Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing [электронный ресурс] // URL: https://arxiv.org/abs/2107.13586 (дата обращения: 15.06.2024).

Liu P. F., Yuan W. Z., Fu J. L., Jiang Z. B., Hayashi H., Neubig G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. On: ArXiv.org. Available at: https://arxiv.org/abs/2107.13586 (accessed: 15 June 2024). (In Eng.)

Liu, Shao 2016 — *Liu M., Shao Q*. The Creation of a Corpus of Russian-Chinese Literary Translations--Design and Construction of a Parallel Corpus Based on Chekhov's Novels // Foreign Language Research. 2016. № 1. Pp. 154–158. (In Chin.)

Liu M., Shao Q. The creation of a corpus of Russian-Chinese literary translations: Design and construction of a parallel corpus based on Chekhov's novels. *Foreign Language Research*. 2016. No. 1. Pp. 154–158. (In Chin.)

Mamaev, Mitrofanova 2020 — *Mamaev I. D, Mitrofanova O. A.* Automatic Detection of Hidden Communities in the Texts of Russian Social Network Corpus // Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, Proceedings. Communications in Computer and Information Science / A. Filchenkov, J. Kauttonen, L. Pivovarova 'eds.). Vol. 1292. Springer, 2020. Pp. 17–33.

Mamaev I. D, Mitrofanova O. A. Automatic detection of hidden communities in the texts of Russian social network corpus. In: Filchenkov A., Kauttonen J., Pivovarova L. (eds.) Artificial Intelligence and Natural Language (AINL 2020). Conference proceedings (Communications in Computer and Information Science 1292). [Helsinki]: Springer, 2020. Pp. 17–33. (In Eng.)

Manning, Schütze 2000 — *Manning Ch., Schütze H*. Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA, 2000. 680 p.

Manning Ch., Schütze H. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press, 2000. 680 p. (In Eng.)

Mimno et al. 2009 — *Mimno D., Wallach H. M., Naradowsky J., Smith D. A., McCallum A*. Polylingual Topic Models // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, 6–7 August 2009. Pp. 880–889.

Mimno D., Wallach H. M., Naradowsky J., Smith D. A., McCallum A. Polylingual topic models. In: 2009 Conference on Empirical Methods in Natural Language Processing. Proceedings. Singapore, 2009. Pp. 880– 889. (In Eng.)

Wu, Dredze 2020 — *Wu S. J., Dredze M*. Are All Languages Created Equal in Multilingual BERT? In the Proceedings of the 5th Workshop on Representation Learning for NLP. 2020. Pp. 120–130. DOI: 10.18653/v1/2020.repl4nlp-1.16

Wu S. J., Dredze M. Are all languages created equal in multilingual BERT? In: 5th Workshop on Representation Learning for NLP. Proceedings. 2020. Pp. 120–130. (In Eng.) DOI: 10.18653/v1/2020.repl4nlp-1.16

Zhai et al. 2020 — *Zhai Y. M., Liu L. F., Zhong X. Y., Illouz G., Vilnat A*. Building an English-Chinese Parallel Corpus Annotated with Sub-sentential Translation Techniques. In Proceedings of the Twelfth Language Resources and Evaluation Conference. Marseille, France. European Language Resources Association, 2020. Pp. 4024–4033.

Zhai Y. M., Liu L. F., Zhong X. Y., Illouz G., Vilnat A. Building an English-Chinese parallel corpus annotated with sub-sentential translation techniques. In: Twelfth Language Resources and Evaluation Conference. Proceedings. Marseille: European Language Resources Association, 2020. Pp. 4024–4033. (In Eng.)

Zhang et al. 2020 — *Zhang B. L., Nagesh A., Knight K.* Parallel Corpus Filtering via Pre-trained Language Models // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020. Pp. 8545–8554.

Zhang B. L., Nagesh A., Knight K. Parallel corpus filtering via pre-trained language models. In: 58th Annual Meeting of the Association for Computational Linguistics. Proceedings [online edition]. Association for Computational Linguistics, 2020. Pp. 8545–8554. (In Eng.)

Mitrofanova et al. 2021 — *Mitrofanova O., Sampetova V., Mamaev I., Moskvina A., Sukharev K*. Topic modelling of the Russian corpus of Pikabu posts: Author-topic distribution and topic labelling // CEUR Workshop Proceedings, 2813. 2021. Pp. 101–116.

Mitrofanova O., Sampetova V., Mamaev I., Moskvina A., Sukharev K. Topic modelling of the Russian corpus of Pikabu posts: Author-topic distribution and topic labelling. In: Internet and Modern Society 2020. CEUR Workshop Proceedings. St. Petersburg, 2021. Pp. 101–116. (In Eng.)

Newman et al. 2009 — *Newman D., Asuncion A., Smyth P., Welling M.* Distributed Algorithms for Topic Models // Journal of Machine Learning Research. Vol. 10. 2009. Pp. 1801-1828.

Newman D., Asuncion A., Smyth P., Welling M. Distributed algorithms for topic models. *Journal of Machine Learning Research*. 2009. Vol. 10. Pp. 1801–1828. (In Eng.)

Mitrofanova et al. 2021 — *Mitrofanova O., Kriukova A., Shulginov V., & Shulginov V.* E-hypertext Media Topic Model with Automatic Label Assignment // Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Supplementary Proceedings. Communications in Computer and Information Science, vol. 1357. Springer, 2021. Pp. 102−114.

Mitrofanova O., Kriukova A., Shulginov V., Shulginov V. E-hypertext media topic model with automatic label assignment. In: Recent Trends in Analysis of Images, Social Networks and Texts (AIST 2020). Conference proceedings (Communications in Computer and Information Science 1357). Springer, 2021. Pp. 102−114. (In Eng.)

Sherstinova et al. 2020 — *Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M.* Topic modelling with NMF vs. expert topic annotation: the case study of Russian fiction // Advances in Computational Intelligence. 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Proceedings / L. Martínez-Villaseñor, H. Ponce, O. Herrera-Alcántara, F.A. Castro-Espinoza ‘eds.). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 12469. Springer, 2020. Pp. 134–151.

Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. Topic modelling with NMF vs. expert topic annotation: The case study of Russian fiction. In: Martínez-Villaseñor L., Ponce H., Herrera-Alcántara O., Castro-Espinoza F. A. (eds.). Advances in Computational Intelligence (MICAI 2020). Conference proceedings (Lecture Notes in Computer Science 12469). Springer, 2020. Pp. 134–151. (In Eng.)

Sun et al. 2010 — *Sun J. S., Wang T. M., Li L., Wu X.* Person Name Disambiguation based on Topic Model // CIPS-SIGHAN Joint Conference on Chinese Language Processing. 2010. Pp. 1–8.

Sun J. S., Wang T. M., Li L., Wu X. Person name disambiguation based on topic model. In: CIPS-SIGHAN Joint Conference on Chinese Language Processing. Proceedings [online edition]. 2010. Pp. 1–8. (In Eng.)

Tian et al. 2014 — *Tian L., Derek F. Wong., Lidia S. Chao., Paulo Quaresma., Francisco Oliveira., Lu Y., Li S., Wang Y.M., Wang L. Y.* UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland. European Language Resources Association 'ELRA), 2014. Pp. 1837–1842.

Tian L., Wong D. F., Chao L. S., Quaresma P., Oliveira F., Lu Y., Li S., Wang Y. M., Wang L. Y. UM-Corpus: A large English-Chinese parallel corpus for statistical machine translation. In: Ninth International Conference on Language Resources and Evaluation (LREC'14). Proceedings. Reykjavik: European Language Resources Association (ELRA), 2014. Pp. 1837–1842. (In Eng.)

Vulić, Moens 2012 — *Vulić I. Moens M.-F.* Detecting Highly Confident Word Translations from Comparable Corpora without Any Prior Knowledge // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France, April 23–27, 2012. Pp. 449–459.

Vulić I. Moens M.-F. Detecting highly confident word translations from comparable corpora without any prior knowledge. In: 13th Conference of the European Chapter of the Association for Computational Linguistics. Proceedings. Avignon, 2012. Pp. 449–459. (In Eng.)

Wang, Qin 2009 — *Wang K. F., Qin H. W.* A Parallel Corpus-based Study of General Features of Translated Chinese // Foreign Language Research. 2009. № 1. Pp. 102–105. (In Chin.)

Wang K. F., Qin H. W. A parallel corpus-based study of general features of translated Chinese. *Foreign Language Research*. 2009. No. 1. Pp. 102–105. (In Chin.)

Приложение 1.

Таблица 1. Результаты тематического моделирования корпуса ДРП-К
[*Table 1*. RWG-C Corpus. Topic modeling results]

№	Темы	Метки YandexGPT
1	经济 'экономика', 推进 'продвигать / продвижение', 政策 'политика', 建设 'строить / строительство', 社会 'общество', 加快 'ускорять', 实施 'претворять', 服务 'услуга', 创新 'инновация', 提高 'повышать', 支持 'поддерживать / поддержка', 推动 'способствовать', 就业 'трудоустройство', 扩大 'расширять', 国家 'государство', 新 'новый', 继续 'продолжать', 坚持 'настаивать', 产业 'производство', 稳定 'стабильность / стабилизировать'	Экономическое развитие. Инновационная экономика. Меры государственной поддержки.
2	物价 'товарные цены', 控制 'контроль / контролировать', 分配 'распределять / распределение', 关系 'отношение', 预算 'бюджет', 小型 'малый', 快 'быстрый', 危机 'кризис', 节能 'энергосбережение', 理顺 'упорядочивать / отрегулировать', 对外 'внешний', 债务 'долг', 累计 'итог/суммировать', 平稳 'стабильный / устойчивый', 社会保障 'социальное обеспечение', 协调性 'координация', 微型 'микро-', 基本 'основной / основа'	Социальное обеспечение. Экономическая стабильность. Меры социальной поддержки.
3	发展 'развитие', 改革 'реформа', 加强 'укреплять / укрепление', 企业 'предприятие', 促进 'способствовать / стимулировать', 全面 'всесторонний', 完善 'совершенствовать', 中国 'Китай', 工作 'работа', 增长 'повышать/повышение', 元 'юань', 财政 'финансы', 地区 'регион', 深入 'углублять', 保障 'обеспечивать / обеспечение', 新 'новый', 地方 'местность', 领域 'сфера', 合作 'сотрудничество', 我国 'наша страна'	Реформы в Китае. Региональное развитие. Экономические преобразования.

4	稳 ‘стабильность’, 新 ‘новый’, 创新 ‘инновация’, 创业 ‘предпринимательство’, 治理 ‘управлять / управление’, 建设 ‘строить / строительство’, 微 ‘микро-’, 培育 ‘воспитывать / воспитание’, 加大 ‘наращивать / наращивание’, 互联网 ‘Интернет’, 化 ‘превращать’, 防治 ‘профилактика’, 更多 ‘многочисленный’, 准 ‘точный’, 就业 ‘трудоустройство’, 提升 ‘продвигать / продвижение’, 科技 ‘наука и техника’, 区间 ‘интервал’, 坚决 ‘решительно’, 强化 ‘усиливать / усиление’	Научно-технические инновации. Развитие науки и техники. Технологические преобразования.
5	疫情 ‘эпидемия’, 防 ‘профилактика’, 控 ‘контроль’, 脱贫 ‘ликвидация бедности’, 支持 ‘поддерживать / поддержка’, 贷款 ‘кредит’, 强化 ‘усиливать / усиление’, 保 ‘защищать / защита’, 民生 ‘народное благосостояние’, 链 ‘цепочка’, 恢复 ‘восстановить / восстановление’, 企 ‘предприятие’, 加强 及时 ‘вовремя’, ‘укреплять / укрепление’, 能力 ‘способность’, 供应 ‘снабжать / снабжение’, 碳 ‘углерод’, 微 ‘микро-’, 失业 ‘безработица’, 阶段性 ‘этапность’	Меры по борьбе с эпидемией. Социально-экономическая поддержка. Укрепление благосостояния народа.
6	政府 ‘правительство’, 基本 ‘основной / основа’, 发展 ‘развивать / развитие’, 安全 ‘безопасность’, 建设 ‘строить / строительство’, 目标 ‘цель’, 市场 ‘рынок’, 主体 ‘субъект / основа’, 增加 ‘повышать / повышение’, 重大 ‘значительный’, 做好 ‘хорошо выполнять’, 坚决 ‘решительно’, 投资 ‘инвестиция’, 合理 ‘рациональный / умеренный’, 贫困 ‘бедность’, 构建 ‘создавать/создание’, 批 ‘рецензировать / партия’, 主要 ‘ключевой’	Инвестиционная деятельность. Развитие рынка. Экономическая политика.
7	负 ‘негативность / поражение’, 纾 ‘облегчать / облегчение’, 时间 ‘время’, 下调 ‘снизить’, 碧水 ‘изумрудные воды’, 贫 ‘бедность’, 门诊 ‘поликлиника’, 力戒 ‘всемерно избегать’, 共商 ‘совместное обсуждение’, 伙伴关系 ‘партнерство’, 外部 ‘внешний’, 铸 ‘сформировать’, 立 ‘определить’, 博览会 ‘выставка’, 名义 ‘номинальный’, 潜能 ‘потенциал’, 冬奥会 ‘зимняя Олимпиада’, 产权 ‘имущественное право’, 予 ‘оказать’, 从严 ‘со всей строгостью’	Международное сотрудничество. Партнёрские отношения. Совместная работа.
8	政府 ‘правительство’, 管理 ‘управлять / управление’, 制度 ‘режим’, 增 ‘повышать / повышение’, 长 ‘длительный’, 积极 ‘активный / активно’, 投资 ‘инвестиция’, 建立 改革 ‘реформа’, ‘строить / строительство’, 发展 ‘развивать/развитие’, 社会 ‘общество’, 结构 ‘структура’, 居民 ‘население’, 万 ‘десять тысяч’, 经济 ‘экономика’, 农村 ‘деревня’, 调整 ‘реконструкция’, 建设 ‘строительство’, 安全 ‘безопасность’, 文化 ‘культура’, 体制 ‘система’	Реформы в сельском хозяйстве. Развитие сельских территорий. Культурная и экономическая жизнь села.
9	推行 ‘внедрять’, 严肃 ‘строгий’, 效率 ‘эффективность’, 众 ‘масса’, 雾 ‘туман’, 棚户区 ‘район с ветхими домами’, 燃煤 ‘энергетический уголь’, 库存 ‘запасы’, 办事 ‘заниматься’, 办法 ‘приём’, 路 ‘дорога’, 强 ‘сильный’, 施政 ‘проводить административные мероприятия’, 打造 ‘вырабатывать’, 利 ‘польза’, 国 ‘государство’, 措施 ‘меры’, 走 ‘идти’, 结构性 ‘структурированность’, 媒体 ‘СМИ’	Реформы в жилищной сфере. Меры правительства по улучшению жилищных условий. Жилищная политика государства.

10	脱贫 ‘ликвидация бедности’, 动能 ‘драйвер’, 侧 ‘предложение’, 变革 ‘переворот’, 杠杆 ‘рычаг’, 企 ‘предприятие’, 一半 ‘половина’, 产权 ‘имущественное право’, 公里 ‘километр’, 随机 ‘вероятный’, 缴费 ‘уплачивать взносы’, 允许 ‘разрешать’, 责 ‘ответственность’, 升级 ‘обновление/усовершенствование’, 税率 ‘тариф’, 人民群众 ‘народы’, 收费 ‘получение денег’, 高校 ‘высшее учебное заведение’, 新旧 ‘старый и новый)	Социально-экономические преобразования. Реформы в сфере образования. Борьба с бедностью.

Таблица 2. Результаты тематического моделирования корпуса ДРП-Р
[*Table 2*. RWG-R Corpus. Topic modeling results]

№	Темы	Метки YandexGPT
1	необходимый, содействовать, развивать, способствовать, расширять, интенсифицировать, базовый, усилие, всесторонне, активизировать, инновационный, направить, предоставление, придерживаться, медицинский, высококачественный, услуга, защита, последовательно, интенсивно	Инновационное развитие. Высококачественные услуги. Последовательная интенсификация.
2	новый, механизм, область, реализация, технический, распределение, год, система, регион, продвигать, повышение, стратегический, производство, деятельность, усиливать, вид, единый, увеличение, достигнуть	Научно-техническое развитие. Стратегические инновации. Повышение технологического уровня.
3	следовать, обеспечение, мера, высокий, реформа, сфера, необходимый. обслуживание, должный, качество экономики, цена, предприятий, стимулировать, отношение, услуга, борьба, политический, народный, повысить	Экономические реформы. Меры стимулирования. Повышение качества.
4	финансовый, политика, результат, субъект, стабилизация, время, текущий, занятость, экономика, КПК, рыночный, снижение, малый, риск, развиваться, сложный, позволить, экономика, микропредприятие	Экономическая стабилизация. Развитие микропредприятий**.** Сложная экономическая ситуация.
5	всё, развитие, местный, год, процент, главный, экономика, система, трансформация, страна, экономический, путь, обеспечение, регулирование, международный, процент, хороший, население, расти, ряд	Экономическое развитие. Трансформация системы. Международный путь.

6	денежный, предприятие, объем, ввести, прежний, сфера, инновационный, налоговый, важный, нагрузка, рациональный, ставка, развитие, долг, вырасти, направить, регистрация, вид, развернуть, избыточный	Рациональное развитие. Инновационная сфера. Налоговое регулирование.
7	сельский, уровень, усиливать, совершенствовать, страна, поддержка, государственный, поддерживать, жизнь, улучшать, строительство, общий, вид, контроль, претворять, китайский, рынок, образование, условие, научный	Государственный контроль. Развитие страны. Улучшение жизни.
8	новый, контроль, правительство, продвигать, ширить, структура, создание, порядок, программа, рабочий, улучшение, полностью, поддержка, составить, юань, народный, поддержание, стабильный, китайский, общественный	Правительство и программа. Создание и улучшение. Поддержание стабильности**.**
9	продолжать, экономический, рост, финансовый, миллион, новый, ещё, административный, отношение, миллиард, центральный, активный, реформа, важно, структурный, стратегия, число, общественный, развитие, проект	Экономический рост. Реформа и стратегия. Общественное развитие.
10	развитие, работа, система, управление, предстоять, стимулировать, обеспечивать, основный, реформа, политика, интерес, углублять, городской, нужно, повышать, сторона, производство, частность, создавать, человек	Развитие и реформа. Управление и работа. Производство и система.

Таблица 3. Результаты тематического моделирования корпуса ППР
[*Table 3*. PAFA Corpus. Topic modeling results]

№	Темы	Метки YandexGPT
1	международный, сила, борьба, сотрудничество, народ, армия, ответственность, страна, попытка, дальнейший, военный, уважение, можно, угроза, суверенитет, показать, оружие, исторический	Международные отношения. Военная политика. Угроза безопасности.
2	медицинский, образование, культура, общественный, русский, учреждение, пора, профессиональный, история, некоммерческий, народ, школа, общество, НКО, нуждаться, работа, самоуправление, кадровый, привлечь	Развитие общественной сферы. Некоммерческие организации. Социальные проекты.
3	ребёнок, семья, выплата, доход, детский, рождение, дети, смочь, дом, демографический, материнский, размер, минимум, прожиточный, зависеть, налоговый, рубль	Социальное обеспечение. Поддержка семей с детьми. Меры социальной поддержки.

4	должен, регион, просить, школа, правительство, бюджетный, внимание, орган, власть, ребёнок, ответственность, срок, бюджет, человек, нельзя, национальный, экологический, обращать, происходить, образовательный	Социально-экономическое развитие региона. Финансирование бюджетных организаций. Обеспечение прав и свобод граждан.
5	научный, центр, технологический, исследовательский, передовой, запустить, подготовка, инфраструктура, средний, школьник, нацелить, цифровой, технология, рабочий, рост, поколение, банковский, бизнес, компания, искусственный	Научно-технологическое развитие. Подготовка кадров. Цифровая трансформация.
6	деловой, рост, рынок, налоговый, процент, бизнес, экспорт, просить, промышленный, иностранный, фонд, миллиард, свобода бизнеса, доход, производство, орган, производство, хозяйство	Экономическое развитие. Бизнес-процессы. Инвестиционная деятельность.
7	год, должен, нужно, всё, уже, Россия, новый, развитие, уважаемый, страна, ещё, человек, российский, система, работа, необходимый, сделать, хотеть, сегодня	Развитие страны. Необходимые изменения. Планы на будущее.
8	медицинский, миллион, город, здравоохранение, населить, строительство, миллиард, пункт, жилищный, доступность, цифровой, здравоохранение, рост, продолжительность, рубль, демографический, предстоящий, стоимость, комплексный	Здравоохранение и социальная сфера. Инфраструктура города. Социально-экономическое развитие.
9	стратегический, ядерный, ракета, система, США, новый, боевой, оружие, договор, военный, дальность, комплекс, ракетный, перспективный, вооружённый, подводный, создание, испытание, американский	Ядерное оружие. Военная техника. Оборонительные системы.
10	политический, государственный, общество, право, президент, закон, партия, дума, уголовный, общественный, законодательный, местный, орган, организация, лицо, считать, законный, суд, участие, судебный	Общественно-политическая сфера. Государственные институты. Законодательная и судебная власть.