Mudrik, A.V. (2011). Socializaciya cheloveka (pp. 327-339). Moskva.
Plaxtij, I.S. (2017). Razvitie mediaprostranstva v sovremennom obshhestve. *Molodoj uchenij, 17* (151), 204-207. Retrieved from: https://moluch.ru/archive/151/42922/ (accessed 23.05.2024).
Fomin, A.G., & Yakimova, N.S. (2024). *Taktiki i markery verbalnoj agressii v kommunikativnom povedenii rossiyan i amerikancev (po materialam rechesituativnogo issledovaniya).* Retrieved from: https://cyberleninka.ru/article/n/taktiki-i-markery-verbalnoy-agressii-v-kommunikativnom-povedenii-rossiyan-i-amerikantsev-po-materialam-rechesituativnogo-issledovaniya?ysclid=lwjkghcx4k717666080 (accessed 22.05.2024)
Frolova, O. E. (2021). Rechevaya agressiya i reakciya na nee. *Russkaya rech, 4*, 45–60. DOI: 10.31857/S013161170016214-4.
Shabalin, Yu. (2024). *Gejmdev i bezopasnost`: rekomendacii po zashhite ot xakerskix atak*. Retrieved from: https://rb.ru/opinion/gejmdev-i-bezopasnost/?ysclid=lwjjjmfvdr962064110 (accessed: 22.05.2024)

УДК 81´33+811.581
https://doi.org/10.25076/vpl.54.05
Горожанов А.И.
Московский государственный лингвистический университет,
Красикова Е.А.[5]
Московский государственный лингвистический университет

ПОЛУЧЕНИЕ ЗНАЧИМЫХ ДАННЫХ ИЗ НЕПОДГОТОВЛЕННОГО ТЕКСТА ПУТЕМ ЕГО АВТОМАТИЧЕСКОЙ ОБРАБОТКИ АВТОРСКИМИ ЛИНГВИСТИЧЕСКИМИ ИНСТРУМЕНТАМИ (НА МАТЕРИАЛЕ ЭЛЕКТРОННЫХ КИТАЙСКИХ СМИ)

[5] © **Горожанов А.И., Красикова Е.А.** 2024
This work is licensed under a Creative Commons Attribution 4.0 International License https://creativecommons.org/licenses/by/4.0

Статья посвящена рассмотрению возможностей авторского программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер» для нахождения и анализа употребления различных частей речи в текстах электронных китайских СМИ. В ходе исследования были изучены технические параметры анализируемых частей речи, а также описаны некоторых функциональные особенности программного обеспечения. Созданный модуль «китайский язык» позволил произвести сборку сбалансированного лингвистического корпуса объемом 18341 токен и выполнить ряд поисковых запросов к этому корпусу. В частности, были произведены удачные попытки идентификации предложений, содержащих существительные, прилагательные, глаголы, числительные и частицы. Также в ходе корпусного эксперимента, который являлся основным методом исследования наряду с методами профессионально ориентированного программирования, моделирования и анализа, было установлено, что в отличие от индо-европейских языков (русского, английского и немецкого), на которых ранее тестировался программный комплекс, китайский язык вносит особенности в алгоритм наполнения базы данных леммами и токенами, что было оперативно учтено в ходе работы. Полученные в ходе запросов языковые и статистические данные были подвергнуты тщательному анализу, в результате которого было установлено, что погрешность определения заявленных частей речи составляет ок. 7%. В качестве перспектив исследования рассматривается оптимизация поиска данных в рамках модуля «китайский язык», в целом, и составление банков данных по отдельным частям речи и по именам собственным, а также формирование списка «стоп-слов» для уменьшения погрешности, в частности.

Ключевые слова: прикладная лингвистика, отечественное программное обеспечение, корпусный менеджер, части речи, китайский язык, электронные СМИ, обработка естественного языка

UDC 81´33+811.581
https://doi.org/10.25076/vpl.54.05
Gorozhanov A.I.
Moscow State Linguistic University
Krasikova E.A.
Moscow State Linguistic University

OBTAINING MEANINGFUL DATA FROM AN UNPREPARED TEXT BY AUTOMATICALLY PROCESSING WITH AUTHOR'S LINGUISTIC TOOLS (BASED ON THE MATERIAL OF ELECTRONIC CHINESE MEDIA)

The article is devoted to the possibilities of the author's software package "Balanced linguistic corpus generator and corpus manager" for finding and analyzing the use of various parts of speech in the texts of electronic Chinese media. During the research, the technical parameters of the analyzed parts of speech were studied, as well as some functional features of the software were described. The created module "Chinese language" made it possible to assemble a balanced linguistic corpus with a volume of 18341 tokens and perform a number of search queries for this corpus. In particular, successful attempts were made to identify sentences containing nouns, adjectives, verbs, numerals and particles. Also, during the corpus experiment, which was the main research method along with the methods of professionally oriented programming, modeling and analysis, it was found that, unlike Indo-European languages (Russian, English and German), in which the software package was previously tested, the Chinese language introduces features into the algorithm for filling the database with lemmas and tokens, which was promptly taken into account during the work. The linguistic and statistical data obtained during the inquiries were subjected to a thorough analysis, as a result of which it was found that the error in determining the declared parts of speech is about 7%. The prospects of the study are the optimization of data search within the framework of the "Chinese language" module, in general, and the compilation of data banks for individual parts of speech and proper names, as well as the formation of a list of "stop words" to reduce the error, in particular.

Keywords: applied linguistics, domestic software, corpus manager, parts of speech, Chinese language, electronic media, natural language processing

Введение

Цель нашего исследования – определить общий уровень качества работы модуля «китайский язык» программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер»[6].

Поставленная цель достигается путем решения следующих задач:

1. Собрать тестовый корпус актуальных текстов электронных СМИ на китайском языке.
2. Сформировать ряд поисковых запросов к корпусу и проанализировать полученные результаты.
3. Установить уровень точности работы модуля.

Языковым материалом работы является сбалансированный лингвистический корпус текстов китайских электронных СМИ, собранных нами в период с февраля по июнь 2024 года. Объем корпуса составил 724 предложения или 18 341 токен.

Основным методом исследования стал корпусный эксперимент, хотя на различных этапах применялись также и другие методы. В частности, для решения первой задачи для разработки модуля «китайский язык» мы применили оригинальный метод профессионально ориентированного (лингвистического) программирования (Gorozhanov, Guseynova, 2020), который позволил создать необходимое программное решение. Для решения второй задачи применялся метод моделирования, поскольку любой поисковый запрос к базе данных корпуса представляет собой формальную модель. Наконец, в ходе решения третьей задачи широко применялся анализ корпусных данных, полученных в результате корпусного эксперимента.

---

[6] Свидетельство о государственной регистрации программы для ЭВМ № 2023683209 Российская Федерация. «Генератор сбалансированного лингвистического корпуса и корпусный менеджер» : № 2023682269 : заявл. 25.10.2023 : опубл. 03.11.2023 / А. И. Горожанов ; заявитель федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный лингвистический университет». – EDN JHFXUV.

Китайские СМИ как лингвокультурный феномен: развитие китайских новостных газет в диахроническом аспекте

Книгопечатание (*yinshuashu* 印刷术), наряду с такими изобретениями, как бумага (*zhi* 纸), компас (*zhinan* 指南) и порох (*huoyao* 火药) относится к четырем великим изобретениям Китая (*sidafaming* 四大发明).

Возникновение печатания подвижным шрифтом приписывают Би Шэну 毕昇, который изобрел данный метод во времена Сун Жэньцзуна 宋仁宗 (1010-1063 гг., четвёртый император династии Сун 宋). По мнению китайских авторов, Би Шэн примерно на 400 лет опередил немца Иоганна Гутенберга, создавшего передовую технологию печати подвижным шрифтом (邓广铭, 2005). Однако первые письменные источники передачи информации, которые представляли собой скреплённые бамбуковые дощечки, появились в Древнем Китае задолго до появления бумаги и техники печатного пресса.

Так, возникновение китайской публицистики связывают с династией Хань 汉 (206 до н. э. – 220 н. э.), в период правления которой была зафиксирована не только самая древняя китайская, но и самая ранняя мировая новостная газета Дибао (*Dibao* «邸报», букв. «Отчеты из [официальных] резиденций) (Xu, 1994, p. 169). Данный термин также является общим для обозначения древних китайских газет в целом. Дибао, составленные императорскими учеными, распространялись при феодальном дворе в качестве единственной официальной правительственной газеты. Ранние Дибао были написаны от руки на небумажных носителях информации, однако, начиная с династии Тан 唐 (618-907 гг.), в качестве носителя информации использовалась бумага. Как мы упоминали ранее, хотя многие исследователи определяют Дибао как наиболее раннюю или примитивную форму газеты в Китае и даже во всем мире, но данная газета «соответствует некоторым характеристикам современных газет при описании исторических материалов, таким как скорость передачи информации (относительно упадка технологий распространения информации в древнем обществе), структура *dianduimian* 点对面, согласно которой феодальное центральное правительство передает

информацию для широкого круга чиновников и представителей знати всех уровней), тиражный цикл, а также носители информации (бумага и полиграфия) и т.д. (姬德强 et. al., 2008, p. 1). Таким образом, Дибао является эволюционным примером первоначальной формы современных китайских газет.

Следующим этапом в развитии китайской публицистики становится возникновение цензуры, которое связывают с возникновением популярных среди гражданского общества неофициальных газет *Xiaobao* 小报 (досл. малоформатная газета) во времена династии Сун (960–1279 гг.). Вместе с тем, появляется противоположный официальный и более объемный тип газеты *Changben* 长本. Как отмечает E. Mokros, «короткие издания» (小报), были доступны раньше, чем официальные «длинные издания» (长本), которые, как правило, считались более авторитетными (Mokros, 2023, p. 168).

В качестве одной их самых ранних новостных газет исследователи также выделяют *Jingbao* 京报 (The Peking Gazette). Например, J.Lane называет данное издание, как «patriarch of periodicals» и причисляет его к старейшим мировым газетам. По замечанию J.Lane, вышеуказанное издание нельзя назвать газетой в современном понимании, поскольку в нем «не публиковались редакционные мнения, не создавался уникальный контент и не давались социальные и культурные комментарии к событиям дня, но оно действительно содержало "новостную ценность" в той же мере, в какой любая традиционная правительственная газета публикует современные документы, относящиеся к событиям дня и повседневной работе государства» (Lane, 2018, p. 2). Более подробное описание и развитие древних китайских газет также исследуется в монографии Хуан Чжомина «中国古代报纸探源» (Исследование древних газет Китая) (黄卓明, 1983), в которой прослеживается происхождение китайских газет до династии Хань и подробно описываются исторические изменения в области печати, распространения, названий, содержания газет от династии Тан до династии Цин.

Современные газеты и журналы начали проникать в Китай вместе с западными религиозными миссиями в начале девятнадцатого века. Когда европейские миссионеры и купцы

впервые начали в большом количестве прибывать к границам империи Цин, они искали источники новостей, которые помогли бы им понять, что происходит в крупнейшей империи мира.

Появление западных версий газет и периодических изданий связывают с публикацией первой португальской газеты "*A Abelha da China"* (蜜蜂画报) в 1822 году, которая была выпущена в связи с «новым регулированием деятельности прессы под влиянием Конституции Португалии, принятой в 1822 году» (Tang, 2022, p. 44). Затем, согласно данным китайской онлайн-энциклопедии 中華百科全書[7], в 1883 году был опубликован первый китайскоязычный журнал "*East Western Monthly Magazine*" (東西洋考每月統紀傳), издававшийся прусским протестантским миссионером Карлом Гютцлаффом в Гуанчжоу. После опиумной войны иностранные издатели газет постепенно расширили свою издательскую базу и круг читателей с прибрежных районов на восточные, центральные и северные районы страны. В течение 100 лет после опиумной войны в газетах и периодических изданиях на иностранных языках преобладал английский, за ним следовали японский, русский, французский, португальский и немецкий языки.

Следующий важный этап в развитии китайского новостного дискурса связан с революционными движениями, возникшими в Китае в начале 20-го века. Значительный рост публикационной активности произошел, когда лидеры реформаторского движения начали использовать средства массовой информации для сплочения людей и распространения новых идей. Как и сторонники реформ, революционеры во главе с Сунь Ятсеном также обратились к газетам и периодическим изданиям для мобилизации масс. До революции 1911 года издавались такие газеты, как *Zhongguoribao*（中国日报） Гонконг; *Guominbao* (国民报) Токио; *Dagongbao* (大公报) Тяньцзинь и др.

В 1919 году группа молодых интеллектуалов издавала газеты «Новая молодежь» (新青年), «Борьба» (奋斗) и другие, чтобы агитировать за научные и демократические идеи против режимов

[7] https://web.archive.org/web/20140517214326/http://ap6.pccu.edu.tw/Encyclopedia/data.asp?id=635

военачальников и империализма, что привело к возникновению исторически влиятельного «Движения 4 мая» (五四运动). Отдельного внимания заслуживает правительственная газета Гоминьдана *Zhongyangribao* (中央日报), которая была основана в Шанхае в 1928 году и впоследствии стала издаваться на Тайване, куда в 1949 году было вытеснено националистическое правительство Гоминьдана. На материковом Китае Коммунистическая партия Китая (КПК) учредила главные общенациональные и городские газеты, как только освободила города и взяла в свои руки государственную власть от Гоминьдана. Например, *Xinhuaribao* (新华日报) в Нанкине, *Jiefangribao* (解放日报) в Шанхае, *Changjiangribao* (长江日报) в Ухане, *Qunzhongribao* (群众日报) в Сиане, *Tianjingribao* 天津日报 в Тяньцзине и т.д. Газета *Renminribao* (人民日报) в Северном Китае была провозглашена официальной газетой КПК.

Все вышеизложенное указывает на то, что китайская газетная индустрия организована вокруг КПК. Правительство определяет круг читателей и распределяет новостную информацию в зависимости от рядов и связей КПК. Механизм заключается в том, что КПК редактирует свои официальные газеты и использует их в качестве примера для всех других специализированных или непартийных газет. Начиная с *Renminribao* и заканчивая ежедневными изданиями провинциальных, муниципальных, префектурных и уездных комитетов под различными названиями, существует около 426 официальных газет КПК со стабильным тиражом в 28 миллионов экземпляров. В подтверждение приведем статистические данные. К 1950 году в стране насчитывалась 281 газета, из которых 116 находились в ведении государства, 58 - народных организаций, 55 - частных издательств, 33 - армии и 19 - других. В период культурной революции, с 1968 по 1970 год, количество названий газет в стране было сокращено до 42 (Xu, 1994, p.179). Во второй половине 20-го века на развитие китайского новостного дискурса оказали экономические и политические предпосылки. Экономические реформы Дэн Сяопина вызвали настоящий газетный бум. С 1980 по 1985 год каждые полтора дня выходила одна новая газета. В результате к 1985 году по всей стране выходила 2191 газета. Из них 227 были ежедневными газетами, что

составляет 10,36 процента от общего числа. Общий тираж составил 207,22 миллиона экземпляров, что составляет примерно один экземпляр на пять жителей страны (Xu, 1994, p.175). Хорошо известная газета (参考消息), которая публикует новости из первых рук в международных СМИ, первоначально была внутренней газетой КПК. В настоящее время она стала «одной из самых популярных газет, доступных широкой публике» (Там же). Вокруг официальных газет КПК сосредоточены газеты для различных областей и специальностей. Существуют утренние и вечерние газеты, ориентированные на городскую жизнь, множество различных газет для рабочих, крестьян, молодежи, женщин, пожилых людей, посвященные зарубежным делам Китая, экономическим вопросам, юридической практике, образованию, спорту, науке и технике, культуре и искусству, народонаселению, здравоохранение и медицина, охрана окружающей среды, общественная жизнь, радио и телевидение, книгоиздание, бизнес, предприятия, университеты и военные. Среди них экономические, научные и деловые газеты занимают самые высокие места как по количеству названий, так и по тиражам (Malyuga & Akopova, 2021).

На сегодняшний день двумя крупными изданиями, которые наиболее тесно связаны с китайским правительством являются газета *Жэньминьжибао* и информационное агентство *Синьхуа* (Kaplan, 2018, p.13). Основным языком, на котором выходят газеты, является китайский. В 1989 году 95,2% газет издавались на китайском языке. Также существует несколько газет и периодических изданий на иностранных языках, которые выходят международными тиражами. Наиболее известные из них China Today, People's China, Beijing Review, China daily (Xu, 1994, p.175). Данные новостные ресурсы также оформлены и в электронном формате, что является отдельным направлением для анализа. Таким образом, для настоящего исследования в качестве основного источника новостной информации было выбрано информационное агенство *Синьхуа* (URL: http://m.news.cn/) как одно из наиболее авторитетных на материковом Китае.

Ход исследования и его результаты

Решение первой задачи исследования потребовало большой технической подготовки.

Базой в плане технического (программного) решения для нас послужил авторский программный комплекс «Генератор сбалансированного лингвистического корпуса и корпусный менеджер», который успешно был апробирован ранее на материале русского, английского и немецкого языков (Gorozhanov et al., 2024; Человек…, 2024; Бондарчук, 2024; Степанова, 2023).

Программа является универсальной оболочкой, однако «универсальность» выступает здесь скорее как переменная, а не константа, поэтому каждый новый язык, вводимый в систему, требует масштабных испытаний на текстовом материале. Итак, мы добавили в генератор китайский язык как готовый пакет от spaCy, внеся соответствующий пункт в графический интерфейс пользователя, и инициировали процесс сборки базы данных, которая была успешно изготовлена. Изучая графическое представление таблиц баз данных (Gorozhanov et al., 2024, p. 200) нами была обнаружена особенность, которая не была характерна для индо-европейских языков. Поскольку для китайского языка нерелевантно понятие словоформы, генератор заполнил ячейки «текст токена», оставив ячейки «лемма» пустыми.

Проанализировав, каким образом эта особенность может оказать влияние на работу корпусного менеджера, мы пришли к выводу о том, что, в первую очередь, это затрагивает программные функции построения частотных списков (для всего текста и по отдельным частям речи).

Очевидными казались два решения проблемы: модифицировать генератор, так чтобы он заполнял ячейки «лемма» параллельно с ячейками «текст токена» или внести изменения в функции частотных списков. Первое решение увеличило бы объем базы данных, хотя и сняло бы впоследствии все прочие вопросы, связанные с работой корпусного менеджера. Тем не менее, не отказываясь окончательно от мысли о модификации генератора, нами было принято решение в качестве эксперимента внести в функции частотного списка дополнительные проверки. Таким образом, модуль «китайский язык» можно охарактеризовать как совокупность программных решений, позволяющих работать с китайским языком, максимально используя возможности корпусного менеджера.

Проверим, насколько правильно генератор определил части речи в корпусе. Для этого сформируем ряд запросов к существительным (NOUN), глаголам (VERB), прилагательным (ADJ), числительным (NUM) и частицам (PART).

Определение существительных

Запрос на вывод существительных формируется путем активации флажка «ВКЛ ЧР» и выбора «NOUN» в ниспадающем меню. В результате проведенного запроса наиболее многочисленной группой токенов оказались существительные. В общей сложности было выведено 5 602 токена в 718 предложениях, что в относительном исчислении составляет 5 602 / 18 341 x 100 % = **30,54 %** от числа всех токенов корпуса. Данные статистические показатели являются характерными для публицистического стиля китайских новостных СМИ, поскольку наряду с книжными языковыми единицами и нейтральной лексикой, широко употребляются обобщающие слова, абстрактные существительные, профессионализмы и терминологическая лексика. Например:

1. 武器 (книж.): 美联社、路透社等媒体援引多名美国官员的话报道，拜登同意乌克兰使用美制武器打击俄罗斯境内俄军目标，但仅限于靠近乌克兰第二大城市哈尔科夫并对这座城市发动攻击或准备发动攻击的俄军。Associated Press, Reuters и другие СМИ со ссылкой на ряд официальных лиц США сообщили, что Байден договорился с Украиной об использовании ***оружия*** американского производства для нанесения ударов по российским военным объектам в России, но только вблизи второго по величине города Украины Харькова, а также о начале атаки на город или подготовке к атаке на центр города русской армией.
2. 目标 (обобщ.): 北约秘书长延斯•斯托尔滕贝格、法国总统埃马纽埃尔•马克龙、英国外交大臣戴维•卡梅伦以及瑞典、挪威、芬兰等国家官员都支持乌方使用西方援助的武器打击俄境内目标。Генеральный секретарь НАТО Йенс Столтенберг, президент Франции Эммануэль Макрон, министр иностранных дел Великобритании Дэвид Кэмерон, а также официальные лица Швеции, Норвегии, Финляндии и других стран поддерживают использование Украиной оружия, полученного с помощью Запада, для нанесения ударов по ***целям*** в России.

3. 5月 (нейтр.): 斯托尔滕贝格 5 月 31 日试图淡化普京发出的警告：“这并不新鲜。 Столтенберг попытался преуменьшить значение предупреждения Путина от 31 ***мая***: «Это не ново».
4. 多样性 (терм.): 双方发表关于中东局势、人工智能和全球治理、生物多样性与海洋、农业交流与合作 4 份联合声明，签署绿色发展、航空、农业食品、商务、人文等领域近 20 项双边合作文件。Стороны выступили с четырьмя совместными заявлениями по ситуации на Ближнем Востоке, искусственному интеллекту и глобальному управлению, ***биоразнообразию*** и океанам, сельскохозяйственным обменам и сотрудничеству, а также подписали около 20 двусторонних документов о сотрудничестве в области экологически чистого развития, авиации, агропродовольственной промышленности, торговли и гуманитарных наук.
5. 压力 (абстр.): 美联储大规模降息，不仅推动通胀飙升，更通过超发美元进口商品、投资他国等方式输出资本，收割全球财富；激进加息又导致全球流动性快速收紧、多种货币大幅贬值，以美元计价借贷的国家清偿债务压力骤增。Масштабное снижение процентных ставок ФРС не только привело к резкому росту инфляции, но и к экспорту капитала за счет чрезмерной эмиссии долларов США для импорта товаров и инвестирования в другие страны с целью получения мирового богатства; агрессивное повышение процентных ставок также привело к быстрому сокращению глобальной ликвидности, резкой девальвации нескольких валют и резкому усилению ***давления*** на страны, которые берут кредиты в долларах США для погашения своих долгов.

Из вышеуказанных примеров следует, что корпусный менеджер идентифицировал не только многосложные существительные, состоящие из двух или трех иероглифических знаков, но и существительные, состоящие из числительного и иероглифического знака, которые не характерны для пользователя русского языка. В примере 3 существительное ***май*** в китайском языке формируется путем сочетания числительного 5 и иероглифического знака 月 *месяц* (досл. пятый месяц). Однако поданному запросу также была зафиксирована небольшая величина погрешности. Например, по запросу «NOUN» корпусный менеджер

выявил имена собственные, прилагательные, а также атрибутивные словосочетания. Например:

1. 习近平 (Им. Собств.) 新时代以来，在习近平强军思想引领下，人民海军听党指挥，在中国特色强军之路上迈出了坚实步伐，正以崭新姿态加速向全面建成世界一流海军迈进。Начиная с новой эры, под руководством ***Си Цзиньпина***, придерживающегося идеологии сильной армии, Народный военно-морской флот подчинился приказу партии и предпринял решительные шаги на пути к созданию сильной армии с китайской спецификой, а также ускоряется к созданию военно-морского флота мирового класса с всесторонним развитием. совершенно новое отношение.
2. 历史 (сущ.) 岁月流逝，人民海军组建之初的 13 名官兵之一，也是最后一名历史见证者的黄胜天将军于 2023 年 12 月离世，没能见证人民海军成立 75 周年。Шли годы, и генерал Хуан Шэнтянь, один из 13 офицеров и солдат Народного военно-морского флота в начале его формирования и последний ***исторический*** свидетель, скончался в декабре 2023 г. Ему не удалось стать свидетелем празднования 75-й годовщины основания Народного военно-морского флота.
3. 去年 (атрибут. словосоч.): 他们同样来自田间地头、来自工厂车间、来自科研院所、来自边关军营…… 去年，习近平总书记首次来到他所在的十四届全国人大江苏代表团参加审议，面对扎根农村、应用数字化技术来种大田的“80 后”，一句由衷的点赞“像魏巧这样的同志到农村去，很好！”，给多少“新农人”以莫大鼓舞。Они также приезжают с полей, из фабричных цехов, из научно-исследовательских институтов, из приграничных военных городков... В **прошлом году** Генеральный секретарь Си Цзиньпин впервые приехал в составе своей делегации из провинции Цзянсу на 14-е Всекитайское собрание народных представителей, чтобы принять участие в обсуждениях. Столкнувшись с “пост-80-ми”, укоренившимися работая в сельской местности и применяя цифровые технологии для выращивания больших полей, он искренне похвалил: “Такие товарищи, как Вэй Цяо, ездят в

сельскую местность, это очень хорошо!", это большое поощрение для многих "новых фермеров".

Поскольку для китайского языка характерен процесс конверсии, некоторые прилагательные были распознаны системой как существительные. Пример 2 демонстрирует типичный для китайского языка случай, когда одно и то же слово может интерпретироваться и как существительное и как прилагательное. В данном примере порядок слов в предложении указывает на то, что слово 历史 *история*, *исторический* является прилагательным, поскольку предшествует существительному 见证 *свидетель*, сочетаясь вместе они формируют атрибутивное словосочетание 历史见证 *исторический свидетель*.

Определение глаголов

Второй по частотности частью речи оказались глаголы. Запрос на вывод глаголов формируется путем активации флажка «ВКЛ ЧР» и выбора «VERB» в ниспадающем меню. В результате было выведено 3342 токена в 716 предложениях, что в относительном исчислении составляет 3342 / 18 341 x 100 % = **18,22 %** от числа всех токенов корпуса. Например:

1. 表示 (глаг.): 这些官员表示，美方仍然不允许乌方使用美制远程导弹打击俄境内纵深目标。Эти официальные лица ***заявили***, что Соединенные Штаты по-прежнему не разрешают Украине использовать ракеты дальнего радиуса действия американского производства для нанесения ударов по глубинным целям в России.
2. 使用 (глаг.) 匈牙利总理欧尔班 5 月 31 日在匈牙利国家电台说，允许乌方使用西方武器打击俄境内目标和向乌克兰派遣西方军事人员的想法让北约"离战争越来越近"。Премьер-министр Венгрии Орбан заявил в эфире венгерского национального радио 31 мая, что идея разрешить Украине ***использовать*** западное оружие для нанесения ударов по целям в России и направить западных военнослужащих на Украину "все ближе и ближе подводит НАТО к войне".
3. 维持 (глаг.): 日本央行在 4 月 26 日举行的货币政策会议上决定，维持现行货币政策不变，并没有如外界预期的那样实施量化紧缩。На своем заседании по денежно-кредитной политике,

состоявшемся 26 апреля, Банк Японии принял решение ***сохранить*** текущую денежно-кредитную политику без изменений и не стал проводить количественное ужесточение, как ожидалось.

4. 分析 (глаг): 不过，彭博社一篇文章分析，大多数亚洲国家如今具备外汇储备更稳固等有利条件，能够避免类似上世纪 90 年代末亚洲金融危机的动荡重演，因此人们几乎不必担忧亚洲再次发生金融危机。 Однако в статье Bloomberg ***проанализировано***, что большинство азиатских стран сейчас имеют благоприятные условия, такие как более стабильные валютные резервы, что позволяет избежать повторения потрясений, подобных азиатскому финансовому кризису конца 1990-х годов. Таким образом, людям практически не нужно беспокоиться об очередном финансовом кризисе в Азии.
5. 住 (глаг.) 34 岁的拉法居民穆罕默德•纳赛尔是 3 个孩子的父亲，一家目前住在加沙地带中部城市代尔拜拉赫一处避难所中。Мохаммед Нассер, 34-летний житель Рафаха, является отцом троих детей. В настоящее время семья ***проживает*** в приюте в городе Дейр-эль-Балах в центральной части сектора Газа.
6. 看到 (глаг.+модификатор) “只要看到我们是在往前走着，就要保持定力。” - Пока ты ***видишь***, что мы идем вперед, ты должен сохранять свою силу.”

Запрос на вывод глаголов показал, что наряду с многосложными глаголами, состоящими из нескольких иероглифов, корпусным менеджером были успешно выявлены односложные глаголы, состоящие из одного иероглифического знака (пример 5). Интересно отметить, что системе удалось распознать специфический для китайского языка класс результативных глаголов, которые состоят из двух частей: смысловой основы и результативной морфемы (пример 6).

При запросе «VERB» также был установлен небольшой процент погрешности. К глаголам системой были отнесены сочетание указательного местоимения и глагола-связки *быть*, а также китайский фразеологический оборот с фиксированной структурой, состоящей из четырех иероглифических знаков (чэнъюй):

1. 这是 (указат. мест. + глагол-связка): 这是当今世界上最典型的霸道霸凌！***Это*** самое типичное деспотичное хулиганство в современном мире!
2. 一国两制 (именной фразеологический оборот.)毛宁表示，美方蓄意攻击“一国两制”，抹黑香港国安法，妄议香港民主自由状况，干预香港特区司法，滥施签证限制。Мао Нин заявил, что Соединенные Штаты намеренно атаковали политику “***Одна страна - две системы***”, дискредитировали закон Гонконга о национальной безопасности, ложно обсуждали статус демократии и свободы в Гонконге, вмешивались в отправление правосудия в Специальном административном районе Гонконг и злоупотребляли визовыми ограничениями.

Определение прилагательных

Запрос на вывод прилагательных формируется путем активации флажка «ВКЛ ЧР» и выбора «ADJ» в ниспадающем меню. В результате было выведено 566 токена в 360 предложениях, что в относительном исчислении составляет 566 / 18 341 x 100 % = **3,08 %** от числа всех токенов корпуса. Вышеуказанные количественные данные не типичны для китайского языка, поскольку для китайского предложения характерны предложения с распространенными определениями, которые оформляются при помощи постановки структурной частицы 的. Данное служебное слово маркирует определение, которое может быть выражено существительным или прилагательным, связывая его с определяемым словом. С одной стороны, корпусный менеджер не выявил ни одного сочетания существительного с частицей 的, однако без погрешности выявил случаи употребления прилагательных. Например:

1. 光荣 (прилаг.): 34 岁的闵江涛在海军军士队伍中是个“小字辈”，却有着光荣的履历：34-летний Мин Цзянтао - “мелкая сошка” в звании сержанта военно-морского флота, но у него ***великолепное*** резюме.
2. 严重 (прилаг.): 2022 年 3 月又“急转弯”，开始激进加息以应对通胀，给世界经济带来严重负面外溢效应，多种非美货币经历了多轮大幅贬值。В марте 2022 года произошел “резкий поворот” и началось агрессивное повышение процентных

ставок для борьбы с инфляцией, что привело к ***серьезным*** негативным побочным эффектам для мировой экономики, а различные валюты, не относящиеся к США, неоднократно подвергались резкому обесцениванию.

3. 新 (прилаг.): 在全球经济复苏的关键时刻，国际社会应当正告美方，不要再给世界制造新的麻烦了。В критический момент восстановления мировой экономики международное сообщество должно призвать Соединенные Штаты прекратить создавать ***новые*** проблемы для всего мира.
4. 全球性 (прилаг.) 俄罗斯拒绝任何排他性主张，将尽一切努力防止全球性冲突。Россия отвергает любые притязания на исключительность и приложит все усилия для предотвращения ***глобальных*** конфликтов.
5. 巨大 (прилаг.) 旧金山会晤以来，中方言出必行，中美禁毒合作取得了进展，中方也为此作出了巨大努力。После встречи в Сан-Франциско Китай заявил, что он должен делать то, что говорит, и китайско-американское сотрудничество в области борьбы с наркотиками достигло прогресса, и Китай также приложил ***огромные*** усилия в этом направлении.

Таким образом, в данном случае сложно говорить о проценте погрешности ввиду недостаточного количества полученных данных. В качестве рекомендации может быть предложен ручной специальный запрос (РСЗ), призван находить последовательности токенов в корпусе, по заданным параметрам, которые имеются в базе данных. по схеме «сущ.+ частица 的». Например, для поиска прилагательных, которые в китайском языке строятся по схеме «сущ. + частица 的» или «сущ. №1, сущ. №2, …+ частица 的» валиден тип запроса «РЗС» (ручной запрос специальный). Данный алгоритм поиска позволит обнаруживать многокомпонентные прилагательные, состоящие из двух и более компонентов, оформленные структурной частицей 的.

Определение числительных

Запрос на вывод прилагательных формируется путем активации флажка «ВКЛ ЧР» и выбора «NUM» в ниспадающем меню. В результате было выведено 813 токенов в 355 предложениях, что в относительном исчислении составляет 813 / 18 341 x 100 % = **4,43 %** от числа всех токенов корпуса. Важно отметить, что

используемое программное обеспечение в качестве числительных распознало порядковые и количественные целые числительные записанные как арабскими цифрами, так и иероглифическими знаками, в том числе успешно были идентифицированы проценты и дробные числа. Например:

1. 六 (колич. числ. 6) 为应对通货膨胀，美联储 2022 年开始激进加息，2023 年 7 月最后一次加息 25 个基点，此后已连续六次在货币政策会议中决定维持利率不变。В ответ на инфляцию ФРС начала агрессивно повышать процентные ставки в 2022 году и в последний раз повысила их на 25 базисных пунктов в июле 2023 года. С тех пор она приняла решение сохранить процентные ставки неизменными на ***шести*** последовательных заседаниях по денежно-кредитной политике.
2. 第 2758 (порядк. числ.) 董军引用了《开罗宣言》、《波茨坦公告》和联合国大会第 2758 号决议等国际文件，明确指出台湾作为中国一部分的国际法基础。Дун Чжун сослался на международные документы, такие как Каирская декларация, Потсдамское воззвание и резолюция ***2758-й*** Генеральной Ассамблеи ООН, чтобы четко указать на основы международного права для Тайваня как части Китая.
3. 三分之二 (дроб. числ.): 《纽约时报》4 月底刊登报道说，在彭博社追踪的约 150 种货币中，有三分之二的货币对美元走弱。В конце апреля газета New York Times сообщила, что ***две трети*** из примерно 150 валют, отслеживаемых агентством Bloomberg, ослабли по отношению к доллару США.
4. 5,25% (проц.): 这是美联储连续第六次将联邦基金利率目标区间维持在 5.25%至 5.5%之间。Это шестой раз подряд, когда ФРС сохраняет целевой диапазон ставки по федеральным фондам между ***5,25%*** и 5,5%.

К числительным система ошибочно отнесла счетные слова, которым предшествуют числительные, например, универсальное счетное слово для существительных 个 (на русский язык не переводится) или некоторые словосочетания, в состав которых входит иероглифический знак 一 (единица):

1. 个 (сч. сл.) 据路透社报道，本次民调 17 日至 20 日展开，误差范围为 3 个百分点。По данным агентства Reuters, опрос

проводился с 17-го по 20-е число с погрешностью в 3 процентных пункта.

2. 一点 (нареч.) 他指出，台湾自古以来就是中国的一部分，从历史上、法律上和现实中，这一点都不容置疑。Он отметил, что Тайвань был частью Китая с древних времен, и в этом нет ***ни капли*** сомнения исторически, юридически и реально.

Определение частиц

Запрос на вывод прилагательных формируется путем активации флажка «ВКЛ ЧР» и выбора «PART» в ниспадающем меню. В результате было выведено 1074 токена в 513 предложениях, что в относительном исчислении составляет 1074/ 18 341 x 100 % = **5,85 %** от числа всех токенов корпуса. Например:

1. 之 (част.) 习近平在贺信中指出，30 年来，在党的坚强领导下，中国工程院团结凝聚院士和广大工程科技工作者，大力推动工程科技发展，不断攻克科技难关，建设大国工程，铸造国**之**重器，为推动我国工程科技创新进步、促进经济社会高质量发展作出了重要贡献。В своем поздравительном письме Си Цзиньпин отметил, что за последние 30 лет под сильным партийным руководством Китайская инженерная академия объединила академиков и большинство работников инженерной науки и техники, чтобы энергично содействовать развитию инженерной науки и техники, постоянно преодолевать научные и технологические трудности, строить это внесло важный вклад в продвижение инноваций и прогресса в области инженерной науки и техники в нашей стране, а также в содействие качественному экономическому и социальному развитию.
2. 的 (част.) 习近平强调，工程科技是推动人类社会发展**的**重要引擎。Си Цзиньпин подчеркнул, что инженерная наука и технологии являются важным двигателем развития человеческого общества.
3. 了 (суфф. пр. вр) 奥比昂接受采访时，记者向他展示**了**这所小学的近照。Когда Обианг давал интервью, репортер показал ему недавнюю фотографию начальной школы.
4. 地 (част.) 纳赛尔两周前离开拉法，生怕以色列出其不意**地**进攻拉法，让人无法逃脱。Насер покинул Рафах две недели

назад, опасаясь, что Израиль нападет на Рафах врасплох, что сделает побег невозможным.

5. 得(част.) 不过，英国“舆观调查公司”的民意调查显示，目前斯塔默的支持率比苏纳克高**得**多。Однако опрос, проведенный британской “Компанией по исследованию общественного мнения”, показывает, что рейтинг одобрения Стармера в настоящее время намного выше, чем у Сунака.

Исходя из данных примеров видно, что корпусному менеджеру удалось успешно выявить разные типы частиц в китайском языке, например, 之, 的, 地, 得, 了 и т.д. Также были выявлены единичные случаи погрешности, например:

1. 年前 (сущ./сч.сл.+предлог) 75 年前，3 辆缴获的美式吉普车就装得下人民海军的全部家当。75 ***лет назад*** в трех захваченных американских джипах можно было разместить все имущество Народного военно-морского флота.

Общий статистический анализ полученных в ходе запросов данных говорит о том, что средняя погрешность обнаружения указанных частей не выходит за пределы 7 %, что можно назвать приемлемым результатом в рамках текущего периода работы над модулем «китайский язык».

Выводы

Заключим, что цель нашего исследования (определить общий уровень качества работы модуля «китайский язык» программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер») была достигнута.

При этом последовательно были решены все три поставленные задачи. Во-первых, был собран тестовый корпус актуальных текстов электронных СМИ на китайском языке, причем программа-генератор не выявила ошибок в процессе составления корпуса. Во-вторых, был сформирован ряд поисковых запросов к корпусу, в который мы включили базовые запросы к идентификации существительных, глаголов, прилагательных, числительных и частиц. Прежде всего, сами запросы были выполнены без технических сбоев, что уже является хорошим результатов. Далее, анализ полученных образцов корпуса (пронумерованных предложений) показал, что заявленные части речи были получены в большинстве образцов, однако наблюдалась и некоторая

погрешность, устранить которую возможно с помощью других функций корпусного менеджера (например, «РЗС»). В-третьих, мы установили, что точность проведенных запросов находится в пределах 7 %.

В практическом отношении результаты исследования могут быть полезны при интерпретации текста, а также при анализе идиостиля автора. В качестве перспектив исследования можно рассматривать оптимизацию поиска данных в рамках модуля «китайский язык», в целом, и составление банков данных по отдельным частям речи и по именам собственным на основе частотного списка, а также формирование списка «стоп-слов» для уменьшения погрешности, в частности.

Разработка собственных программных средств лингвистического назначения вносит вклад в дело создания передовых отечественных образцов программного обеспечения в различных сферах экономики Российской Федерации.

Литература

1. Бондарчук Г. Г. Семиотические функции английских наименований одежды в публицистическом тексте (корпусное исследование) // Вестник Московского государственного лингвистического университета. Гуманитарные науки. – 2024. – Вып. 4(885). – С. 23–29.
2. Степанова Д. В. Программный комплекс для генерации динамического корпуса текстов СМИ / Д. В. Степанова // Вестник Минского государственного лингвистического университета. Серия 1: Филология. – 2023. – № 6(127). – С. 123-130. – EDN FMBTKO.
3. Человек - язык - компьютер. Исследователи будущего: Материалы Научно-практической (заочной) конференции с международным участием. Москва, 25 декабря 2023 года. – Москва: Московский государственный лингвистический университет, 2024. – 172 с. – ISBN 978-5-00120-500-5. – EDN ISDFHS.
4. Malyuga E. N., Akopova A. S. Precedence-setting tokens: Issues of classification and functional attribution // Training, Language and Culture. – 2021. – № 5(4). – Pp. 65-76. https://doi.org/10.22363/2521-442X-2021-5-4-65-76

5. Mokros E. Chinese Gazettes on the Margins of Book History: Movable Type, Wax Stereotypes, and Vernacular Techniques in Late Imperial China. – Book History. – Volume 26, – Issue 1, – 2023, pp. 164-202.
6. Gorozhanov A. I. Programming for specific purposes in linguistics: A new challenge for the humanitarian curricula / A. I. Gorozhanov, I. A. Guseynova // Training, Language and Culture. – 2020. – Vol. 4, No. 4. – Pp. 23-38. – DOI 10.22363/2521-442X-2020-4-4-23-38. – EDN ENGZZF.
7. Gorozhanov A. I. Natural Language Processing and Fiction Text: Basis for Corpus Research / A. I. Gorozhanov, I. A. Guseynova, D. V. Stepanova // RUDN Journal of Language Studies, Semiotics and Semantics. – 2024. – Vol. 15, No. 1. – P. 195-210. – DOI 10.22363/2313-2299-2024-15-1-195-210. – EDN FKVAOI.
8. Kaplan M. Media Gatekeeping with Chinese Characteristics: An Analysis of the Chinese Government's Role in English-Language State-Owned News Organizations [Электронный ресурс] // Centre for East and South-East Asian Studies, Lund University. – 2018. – Режим доступа: https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8967295&fileOId=8967296 (дата обращения: 10.04.2024)
9. Lane J. Harris. The Peking Gazette: A Reader in Nineteenth-Century Chinese History. – Leiden; Boston: Brill. – 2018. – P. 374.
10. Tang Io Weng. The Research History of A Abelha da China, China's First Foreign Newspaper. Revista De Cultura. A Abelha Da China: 200 years of foreign-language press in Macao. – 2022. – Vol. 70. – Pp. – 42-53.
11. Xinyi Xu. The Chinese Mass Media // Handbook on Chinese Popular Culture / edited by D. Wu and P. Murphy. – Westport, CT: Greenwood Press. –1994. – Pp. 169-195.
12. 邓广铭. 邓广铭全集第六卷. 河北: 河北教育出版社，2005 年，671 页。[Дэн Гуанмин. Шестой том полного собрания сочинений Дэн Гуанмина. Хэбэй: Издательство Hebei Education Press, 2005, 671 с.].
13. 姬德强 陈旭静 石艳 张磊 一份古代媒介的"传者图像"——邸报的传播者研究. 北京: 媒介研究 2008 年第 2 期], 1-12 页。[Цзи Дэцян, Чэнь Сюцзин, Ши Янь, Чжан Лэй. Образ

коммуникатора древних китайских СМИ - исследование о Дибао. Пекин: Медиа-исследования. Выпуск 2. – 2008. – С.1-12].

14. 黄卓明. 中国古代报纸探源. 河北: 人民日报出版社， 1983 年，184 页 。 [Хуан Чжо Мин. Исследования источников возникновения древних китайских газет. – Хэбэй: Издательство Жэньминь жибао, 1983. – С. 184.].

References

Bondarchuk, G. G. (2024). Semiotic functions of English clothing names in a journalistic text (corpus-based study). *Vestnik of Moscow State Linguistic University*. *Humanities, 4*(885), 23–29.

Stepanova, D. V. (2023). Programmnyj kompleks dlya generacii dinamicheskogo korpusa tekstov SMI. *Bulletin of the Minsk State Linguistic University. Series 1: Philology, 6*(127), 123-130. – EDN FMBTKO.

Chelovek - yazyk - komp'yuter. Issledovateli budushchego: Materialy Nauchno-prakticheskoj (zaochnoj) konferencii s mezhdunarodnym uchastiem. Moskva. – Moscow: Moscow State Linguistic University, 2024. – 172 p. – ISBN 978-5-00120-500-5. – EDN ISDFHS.

Malyuga, E. N., & Akopova, A. S. (2021). Precedence-setting tokens: Issues of classification and functional attribution. *Training, Language and Culture, 5*(4), 65-76. https://doi.org/10.22363/2521-442X-2021-5-4-65-76

Mokros, E. (2023). Chinese Gazettes on the Margins of Book History: Movable Type, Wax Stereotypes, and Vernacular Techniques in Late Imperial China. *Book History*, 26(1), 164-202.

Gorozhanov, A. I. & Guseynova, I. A. (2020). Programming for specific purposes in linguistics: A new challenge for the humanitarian curricula. *Training, Language and Culture, 4*(4), 23-38. DOI 10.22363/2521-442X-2020-4-4-23-38. – EDN ENGZZF.

Gorozhanov, A. I., Guseynova, I. A., Stepanova, D. V. (2024). Natural Language Processing and Fiction Text: Basis for Corpus Research. *RUDN Journal of Language Studies, Semiotics and Semantics, 15*(1), 195-210. DOI 10.22363/2313-2299-2024-15-1-195-210. – EDN FKVAOI.

Kaplan, M. (2018). *Media Gatekeeping with Chinese Characteristics: An Analysis of the Chinese Government's Role in English-Language State-Owned News Organizations* [Electronic resource]. Centre for East and South-East Asian Studies, Lund University. Retrieved from: https://lup.lub.lu.se/luur/download?func=DownloadFile&recordOId=8967295&fileOId=8967296 (accessed: 04.10.2024)

Harris, L. J. (2018). *The Peking Gazette: A Reader in Nineteenth-Century Chinese History.* Leiden; Boston: Brill.

Tang Io Weng. (2022). The Research History of A Abelha da China, China's First Foreign Newspaper. *Revista De Cultura. A Abelha Da China: 200 years of foreign-language press in Macao, 70,* 42 -53.

Xinyi, Xu. (1994). The Chinese Mass Media. In D. Wu and P. Murphy (Eds.) *Handbook on Chinese Popular Culture* (pp. 169-195). Westport, CT: Greenwood Press.

邓广铭. (2005). 邓广铭全集第六卷. 河北: 河北教育出版社， 671 页。[Deng Guangming. The sixth volume of the complete works of Deng Guangming. Hebei: Hebei Education Press].

姬德强 陈旭静 石艳 张磊 (2008) 一份古代媒介的"传者图像"——邸报的传播者研究. : 媒介研究 2008 年第 2 期], 1-12 页。[Ji Deqiang, Chen Xiujing, Shi Yan, Zhang Lei. The image of the communicator of the ancient Chinese media is a study on Dibao. Beijing: Media Studies, issue 2, pp.1-12].

黄卓明. (1983). 中国古代报纸探源. [Huang Zhuomin. Research on the sources of ancient Chinese newspapers. Hebei: People's Daily Publishing House, 1983, p. 184.].