Министерство науки и высшего образования РФ
Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт филологии и языковой коммуникации
Кафедра восточных языков

УТВЕРЖДАЮ
Заведующий кафедрой
_____ И.Г. Нагибина

« _____» _____ 2024 г.

БАКАЛАВРСКАЯ РАБОТА

45.03.02 Лингвистика

КОНСТРУИРОВАНИЕ ТЕКСТОВОГО КОРПУСА ДИАЛЕКТА ХАККА

Научный руководитель ст. преп. каф. ВЯ
Ю. Чжан

Выпускник Е.А. Скомороха

Нормоконтролер И.А. Рабцевич

Красноярск 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Истоки корпусной лингвистики лежат в далеком прошлом, но с развитием компьютерных технологий данное научное направление получило наибольшее развитие. Прогресс в области корпусной лингвистики и создание корпусов существенно сказались на работе ученых, позволив ускорить процесс анализа текстов и поиска необходимой информации в разы.

Главной целью корпусной лингвистики является разработка корпусов, в основе которых лежит систематизация и анализ языковых единиц, составляющих текст. Применение таких корпусов на практике значительно сокращает время работы, что позволяет специалистам оперативно справляться с поставленными задачами.

В свою очередь, китайский язык входит в список наиболее изучаемых языков, что привлекает интерес миллионов людей по всему миру. Однако, упоминая китайский, не стоит забывать о множестве диалектов, которые в значительной степени влияют на языковую картину мира носителей языка.

В свою очередь, **актуальность исследования** заключается в стремлении в рамках современного научного направления создать корпус текстов, который будет доступен российским ученым на базе одного из наименее изученных китайских диалектов.

Целью исследования является разработка текстового корпуса китайского языка на основе диалекта Хакка.

В ходе данной работы мы ставим перед собой ряд **задач:**

1. подготовить описательную характеристику теоретической базы корпусной лингвистики;
2. проанализировать проблемы создания корпуса в рамках корпусной лингвистики;
3. описать теоретические основы китайской диалектологии;
4. подобрать материал для создания корпуса китайского языка диалекта Хакка;

5. разработать текстовый корпус китайского языка диалекта Хакка на материале видеороликов.

6. провести сравнительный анализ лингвистических особенностей диалекта Хакка и путунхуа.

Объектом исследования является диалект Хакка китайского языка.

Предметом исследования выступают лингвистические особенности диалекта Хакка, представленные в видеороликах платформы Bilibili.

Материалом данной работы послужили тексты, взятые из видеороликов китайской социальной сети Bilibili. Общий объем составляет 30 видеороликов.

Основной теоретико-методологической базой исследования послужили научные труды отечественных и зарубежных ученых А.Н. Алексахина, О.И. Завьяловой, Л.Л. Касаткина, П. Хао, М.В. Софронова, С. Чжан, Ц. Ян, Ш. Янь, С.Е. Яхонтова в области русского и китайского языкознания. Посвященные исследованиям корпусной лингвистики труды принадлежат ученым С.Ю. Богдановой, В.П. Захарову, А.А. Кибрику, Т. Мак-Энери, В.А. Плунгяну, В.И. Подлесской, В.В. Потапову, М.И. Солнышкиной, С.В. Толдовой, Э. Харди.

Методы исследования – метод сплошной выборки, описательный метод, метод сравнительного анализа, комплексный метод конструирования текстового корпуса.

Практическая значимость исследования заключается в возможности использовать сконструированный корпус в целях изучения диалекта или проведения исследований в области китайской диалектологии. Данные, полученные при анализе корпуса, также могут быть полезны ученым, заинтересованным в диалекте Хакка.

В Главе 1 рассматриваются два научных направления – корпусная лингвистика и диалектология, а также приводится присущий им понятийный аппарат. Глава содержит сведения о методе конструирования корпуса текста, разработанного В.П. Захаровым, который впоследствии используется в

практической части работы. В главе также описывается современная языковая ситуация в Китае.

В Главе 2 данной работы представлены результаты практической части исследования – подробно описан процесс создания корпуса, подготовлен анализ диалекта Хакка в сравнении с путунхуа на материале полученного корпуса.

Данная работа прошла **апробацию** во время XV Международной научно-практической конференции молодых исследователей «Язык, дискурс, (интер)культура в коммуникативном пространстве человека» (Красноярск, 2023) и Онлайн-семинара «Восточный вектор образовательного и научного контента: дискурсивные практики современного Китая». Кроме того, промежуточные результаты исследования были представлены в формате стендового доклада в рамках Международной научно-практической конференции «XXI Березинские чтения: Языковое бытие человека и этноса». Теоретические положения данной работы представлены в статье электронного журнала Litera: Скомороха Е.А., Чжан Ю. Фонетические особенности диалекта Хакка китайского языка уезда Мэй // Litera. 2024. № 5. С. 175–182.

ГЛАВА 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ПОСТРОЕНИЯ КОРПУСА НА ОСНОВЕ ДИАЛЕКТА ХАККА КИТАЙСКОГО ЯЗЫКА

1.1. Корпусная лингвистика как отдельное направление лингвистической науки

Корпусная лингвистика – это раздел прикладной лингвистики, который занимается разработкой общих принципов и методов создания лингвистических корпусов (корпусов текстов) и методов использования корпусных данных в лингвистическом исследовании [Толдова, 2020: 407]. Характеризуя данное направление, уместно использование термина «идеология». Так, М.А. Лаврентьев пишет о том, что «…корпусная лингвистика – это не направление, связанное с определенным ярусом языковой системы, или определенной теорией, или аспектом анализа. Это скорее идеология, согласно которой результаты лингвистического исследования должны опираться прежде всего на анализ текстов (устных или письменных), а не на интуицию исследователя или информанта» [Лаврентьев, 2004: 121]. Следовательно, подразумевается необходимость объективной работы с текстами, для чего и используются корпусы. В.А. Плунгян в рамках публичной лекции говорит о том, что корпусная лингвистика в узком смысле – это всего лишь наука о том, как создавать корпусы и как ими пользоваться, но она претендует на гораздо большее, на роль новой идеологии науки о языке [Плунгян, 2009]. Смысл данного высказывания заключается в актуальности корпусной лингвистики и потребности дальнейшего развития данного научного направления. Оба ученых призывают обратить внимание на достаточно молодую, но делающую огромный вклад в работу лингвистов науку и воспользоваться ее инструментами.

История корпусной лингвистики началась задолго до появления компьютеров. Одним из наиболее известных примеров является работа

доминиканского монаха Г. Сен-Шерского. Ученый во главе группы монахов прихода Святого Джеймса составил первый сборник вариантов прочтения Библии (1230 г.), предоставив для каждого отдельного слова комментарий. Таким образом, слова были расположены в алфавитном порядке, а также было указано предложение, в котором употребляется данное слово. Сделано это было для того, чтобы можно было с легкостью найти стих с этим словом и сравнить варианты его употребления [McEnery, Hardie, 2012]. В истории корпусной лингвистики также закрепилось имя А. Крудена, автора раннего библейского конкорданса, корректора и издателя, а также самозваного корректора национальной морали [Солнышкина, Гатиятуллина, 2020: 133]. Соответственно, эпоха доэлектронных корпусов началась в XIII в. и завершилась к началу 1960-х гг., однако методы, лежащие в основе современной корпусной лингвистики, зародились в период письменных корпусов. Корпусная лингвистика как наука появилась преимущественно на материале английского языка (SEU, Brown Corpus, LOB и Norwegian Computing Centre for the Humanities), но очень быстро начали возникать корпусы на базе и других языков [Козлова, 2013: 79]. В отечественной лингвистике идея создания корпуса русских текстов, или «машинного фонда русского языка» принадлежит академику-информатику А.П. Ершову и была высказана им в 1978 г. В 2003 г. группа лингвистов из Москвы, Санкт-Петербурга и ряда других научных центров приступила к работе над созданием «Национального корпуса русского языка». В настоящее время в корпусе представлены в основном литературные произведения конца XX в. Тексты корпуса снабжены морфологической разметкой, разработанной под руководством В.А. Плунгяна на основе системы «Грамматического словаря» А.А. Зализняка [Лаврентьев, 2004: 132]. Возникновение и развитие данного научного направления напрямую связано с прогрессом в сфере компьютерных технологий. Изначально корпусы текстов использовались специалистами для узконаправленных прикладных задач и не находили отклика среди лингвистов. К 80-м г. прошлого столетия ситуация начала меняться, и корпусная

лингвистика как отдельный раздел языкознания окончательно сформировалась в первой половине 90-х гг. XX в. В это же время начал оформляться и понятийный аппарат научного направления [Плунгян, 2009].

Одним из основных терминов корпусной лингвистики является корпус. В работах разных ученых ему присущи отличающиеся формулировки. Т. Мак-Энери и Э. Вилсон дают следующее определение: корпус — это собрание языковых фрагментов, отобранных в соответствии с четкими языковыми критериями для использования в качестве модели языка [McEnery, Wilson, 2001]. «Корпус текстов, с одной стороны, это исходный речевой материал для корпусной лингвистики и для других лингвистических дисциплин; с другой стороны, результат деятельности корпусной лингвистики», – так в своей работе характеризует корпус Ю.А. Волоснова [Волоснова, 2006: 47]. Н.В. Козлова под корпусом подразумевает «представленный в электронном виде, размеченный для анализа в лингвистических целях, обеспеченный сравнительно простой в использовании поисковой системой репрезентативный массив неотредактированных текстов, представляющих как можно большее количество «вариантов» языка» [Козлова, 2013: 87]. Э. Финеган представляет корпус как репрезентативное собрание текстов, обычно в машиночитаемом формате, включающее информацию о ситуации, в которой текст был произведен. К ней относится информация о говорящем, авторе, адресате или аудитории [Finegan, 2004].

Отдельно выделяют национальный корпус языка. Он является собранием текстов в электронной форме, представляющих данный язык на определенном этапе его существования, отображающим данный язык во всем многообразии жанров, стилей, социальных и территориальных диалектов и т.п. Создание Национального корпуса дает огромные возможности для всех направлений лингвистических исследований. Возможность массовой статистической обработки текстов позволяет математически подтверждать или опровергать гипотезы, составлять грамматики и словари. Британский

Национальный корпус и Национальный корпус русского языка являются яркими примерами корпусов такого рода.

В 1990-е гг. еще одним из активно развивающихся направлений в корпусной лингвистике стало создание корпусов работ учащихся, изучающих язык как иностранный или второй (неродной), и лингвистический анализ этих работ. Такие корпусы получили название learner corpora – корпусы работ учащихся, учебные корпусы или ученические корпусы [Алсуфьева, 2013: 2]. Учебный корпус, как и любой другой текстовый корпус, – это не просто архив работ студентов, а специально организованная система, построенная на определенных принципах и включающая различные виды аннотаций [Киссилев, 2011: 97].

Следовательно, каждому корпусу присущ ряд определенных характеристик. Наиболее значимой особенностью, отличающей корпус, является репрезентативность (англ. representativeness) – способность отражать все свойства проблемной области. Именно репрезентативность превращает обычный набор разнообразных текстов непосредственно в корпус текстов, пригодный для проведения лингвистического исследования. Как отмечает Ю.А. Волоснова, «репрезентативность определяется фонетическими, морфологическими, синтаксическими, стилевыми параметрами» [Волоснова, 2006: 49]. В этом высказывании заложена мысль о необходимости понимания, на кого ориентирован корпус. Отталкиваясь от мысли, для какого круга исследователей он предназначен, автор ставит перед собой определенную цель. Проблема формирования представительной выборки влечет за собой второй важный термин, присущий корпусной лингвистике – сбалансированность корпуса (англ. balance). Соблюдая данную характеристику, автор корпуса отвечает на вопрос, из каких текстов необходимо сформировать тот самый минимально необходимый объем. Подразумевается, что именно репрезентативность и сбалансированность обеспечивают достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов, то есть способность

отражать все свойства языка или подъязыка. Благодаря этим характеристикам полученные на материале корпуса результаты являются достоверными [Захаров, 2020: 22].

Кроме того, совокупность множества текстов, подразумевающих корпус, должна быть представлена в электронном формате (в сети Интернет или на электронном носителе). Такое условие делает корпус общедоступным. Следующая характеристика, требующая внимания, заключается в том, что языковые данные должны быть размечены для анализа в лингвистических целях. Лингвистическая разметка позволяет работать с большими объемами текстов за считанные секунды. Последний пункт – в результате проведенного анализа должна существовать возможность различного распределения полученного языкового материала (по жанровой принадлежности, году создания текста, тематике и т. п.) [Козлова, 2013: 80].

К корпусам текста относится система управления текстовыми и лингвистическими данными, которую называют корпусным менеджером (англ. corpus manager). Это специализированная система поиска, включающая в себя программные средства для поиска запрашиваемых данных в корпусе и предоставления их пользователю в удобной форме, а также для получения статистической информации.

Неотъемлемой частью поиска информации в корпусе является конкорданс – список всех употреблений данного слова в контексте со ссылками на источник [Захаров, 2020: 12]. Конкорданс противопоставляется словарю, так как в данном случае слово предоставляется в контексте, с его словесным окружением. Поворотным моментом в истории развития конкордансов стала разработка методики использования ключевых слов (key words) в системе Keyword out of context (KWOC) или Keyword in title (1856 г.). Современные конкордансы значительно отличаются от первых разработок. Так называемые конкордансы четвертого поколения предлагают больший спектр функций с возможностью составлять свой корпус и

сравнивать полученные результаты с результатами референтных корпусов [Солнышкина, Гатиятуллина, 2020: 154].

На сегодняшний день существуют мнения, что понятийный аппарат корпусной лингвистики в русском языке до сих пор не сформировался полностью. Связано это с тем, что основы научного направления лежат в английском языке, поэтому терминология складывалась и продолжает развиваться в недрах английского языка. Русские термины, как правило, заимствуются из языка-основоположника. Кроме того, нередкими бывают споры по поводу формы употребления того или иного слова. Например, термин «корпус» во множественном числе имеет два варианта употребления: «корпусы» и «корпуса». Ссылаясь на Большой толковый словарь русского языка, для значения «массив», которое имеет место в случае языковых корпусов, именительный падеж множественного числа должен быть «ко́рпусы», и, соответственно, прилагательное «ко́рпусный» должно произноситься с ударением на первом слоге [Большой толковый словарь русского языка, 1998]. Установленная языковая норма по данному вопросу отсутствует, однако исходя из правил, предпочтение должно отдаваться именно такому варианту употребления данного термина [Захаров, 2020: 16].

В данном параграфе была предпринята попытка описать основные теоретические сведения, относящиеся к области корпусной лингвистики. Была рассмотрена историческая составляющая данного направления, выделены и охарактеризованы основные термины, описана проблема до сих пор развивающейся терминологии. Основная цель корпусной лингвистики и разработки корпусов текстов представляется в необходимости создания лингвистически размеченного текстового массива, состоящего из объективно отобранной информации. Представленная цель является главенствующей в данной работе.

1.2. Корпусная лингвистика в китайском языкознании

В Китае создание корпусов китайского языка берет свое начало в 20-х гг. прошлого столетия. Для того времени характерно проведение исследований, направленных на изучение частотности употребления иероглифов при помощи статистических методов с целью в конечном итоге составить список базовых иероглифов китайского языка. Объём таких корпусов сначала был крайне невелик. Существенной предпосылкой для лингвистически обоснованного формирования корпусов послужила их эффективность при изучении частотности – употребимости языковых единиц в речи. Своеобразным прообразом китайских корпусов можно с полным основанием считать собрание китайских текстов для исследования частотности «Сборник текстов для изучения единиц разговорного стиля языка» 语体文应用字汇 / yǔtǐwén yìngyòng zìhuì [Баркович, Ван, 2015: 105]. Более масштабные исследования по разработке машинного фонда текстов на китайском языке начались в 1982 г., когда появился первый в Китае корпус английского языка JDEST, и с тех пор корпусно-ориентированные изыскания активно ведутся научными группами на территории всего Китая и за рубежом. За это время было создано свыше двух десятков корпусов китайского языка, а также целый ряд двуязычных корпусов. Наиболее значительные и существенные разработки были проведены в Уханьском университете «Корпус современной литературы китайского языка», (1979 г.); Пекинском педагогическом университете «Корпус школьных учебников по филологии», (1983 г.); Пекинском институте языков «Корпус частотности слов современного китайского языка», (1983 г.) [Колпачкова, 2015: 2]. Первым собственно китайским лингвистическим корпусом является Размеченный корпус газеты «Жэньминьжибао» «人民日报» 标注语料库 / rénmín rìbào biāozhù yǔliàokù, созданный в 1999 г. Таким образом, первые попытки заложили основу корпусных исследований в Китае и стали в известном смысле прототипами современных корпусов китайского языка.

Основной целью создания ранних корпусов китайского языка было преимущественно определение частотности иероглифа в целях оптимизации преподавания китайского языка. Однако современные корпусы китайского языка проходят очень интенсивный период развития, входят в группу гигакорпусов, предлагают программное обеспечение для быстрого поиска лингвистической информации и широко используются в разных сферах науки [Фэн Юэ, 2020: 170].

Занимаясь вопросом изучения наиболее масштабных современных корпусов китайского языка, стоит уделить внимание следующим примерам. Первым внимания заслуживает Chinese Corpus online 语料库在线 / yǔliàokù zàixiàn, который начал создаваться в 1991 г. по инициативе правительства. Комитет по работе в области языка и письменности Китая видел потребность в создании корпуса для того, чтобы способствовать теоретическим исследованиям китайского языка, оптимизации его преподавания, и с целью предоставить ресурсы для его компьютерной обработки. На данный момент общий объём корпуса составляет 100 млн единиц, языковой материал охватывает период с 1919 по 1997 гг. На сайте данного корпуса Chinese Corpus online представлены программы автоматической сегментации текстов, частеречной разметки слов, подсчета частотности слов и разметки пиньиня. С помощью таких программ обеспечивается принципиальная возможность обработки текстов современного китайского языка в целях корпусной лингвистики. Представленная информация позволяет сделать вывод, что материал данного корпуса многообразен, охватывает длительный период и предоставляет возможность цифровой обработки текстов. Недостатком является отсутствие языкового материала нового столетия.

Второй не менее важный корпус – корпус CCL при Пекинском университете. Он берет свое начало в 2000 г., когда центр исследования китайской лингвистики (Center for Chinese Linguistics Peking University) и исследовательский институт компьютерной лингвистики Пекинского университета (Institute of Computational Linguistics PKU) начали совместно

создавать корпус китайского языка. Первая версия корпуса была принята в работу в 2004 г. Корпус CCL состоит из корпуса современного, древнекитайского языков и параллельного китайско-английского корпуса. Его данные обновляются непрерывно и по сей день. Особенностью корпуса является отсутствие аннотированности, он предоставляет в распоряжение исследователя оригинальные тексты с некоторой информацией об их классификации. В нем отсутствуют также лексическая и синтаксическая разметки, зато предлагается разветвленная поисковая система.

Корпусный центр при Пекинском университете языка и культуры (BLCU Corpus Center – сокр. BCC) является крупнейшим корпусом китайского языка в мире. Его объём составляет 15 млрд иероглифов. В данный момент корпус содержит материал девяти языков, в том числе английский, испанский, французский, немецкий, турецкий. Главным источником англоязычного материала является The Wall Street Journal, объём которого в корпусе достиг 1,2 млрд слов. Кроме того, разработаны параллельные корпусы: английско-китайский, английско-немецкий. Корпус включает неаннотированный материал, материал с сегментацией слов, частеречной разметкой. В настоящее время сделана частеречная разметка для материала современного китайского языка, английского и французского языков, остальной материал является неаннотированным. Корпус соединяет в себе материал и современного, и древнекитайского языков. BCC представляет собой масштабный корпус, всесторонне отражающий состояние современного китайского языка, однако требующий производить самостоятельную выборку подкорпусов в связи с многообразием его материала [Фэн Юэ, 2020: 169].

Национальный корпус русского языка является наиболее удобным инструментом исследователя для работы с текстами, так как он разработан на базе русского языка, но открывает доступ и к работе с другими языками. В целом, данный корпус является представительной коллекцией текстов на русском языке общим объемом более 2 млрд слов, оснащенной лингвистической разметкой и инструментами поиска. Он по праву считается

самым большим открытым параллельным корпусом русского и китайского языков. Временем основания корпуса считается 2003 г., однако работа над пополнением материала ведется до сих пор. На данный момент объем языкового материала корпуса составляет 4,4 млн языковых единиц [Национальный корпус русского языка, 2003].

Появление корпусов китайского языка, размещенных в сети Интернет, долгое время сдерживалось сложностью компьютеризации иероглифической письменности, однако даже после разрешения проблемы ввода иероглифов отсутствие какой-либо разметки в предлагаемых текстовых базах еще долгое время сводило использование корпусов лишь к составлению конкордансов.

Основные принципы организации корпуса китайского языка в целом не расходятся с принятыми в корпусной лингвистике. Специфические проблемы корпусов связаны со структурными особенностями китайского языка как яркого представителя изолирующих языков. Одной из самых дискуссионных является проблема неоднозначных случаев, когда языковая единица может входить в два или три класса слов, и ее актуальность связана именно с необходимостью морфологической разметки корпуса, где важнейшим фактором оказывается не просто определение набора морфологических признаков в целом, но детальная разработка правил присвоения этих признаков единицам текста.

Исходя из вышеперечисленного, сложность связана с невозможностью добавлять морфологическую, синтаксическую или семантическую информацию о той или иной лексической единице автоматически, требуется ручная разметка корпусов, дальнейшая проверка и внесение синтаксических помет лингвистом, что усложняет данную задачу при работе с большими массивами текстов [Колпачкова, 2015: 7].

Еще одним важным отличием китайского текста является то, что в письменной форме китайского языка между иероглифами отсутствуют пробелы, то есть между единицами китайского языка существуют потенциальные границы, которые не указаны на письме. В результате этого

возникают трудности в разбиении текста на слова. В то же время из-за отсутствия показателей категорий числа, падежа и рода в китайском языке нет согласования, следовательно, функция слова в китайском языке становится понятной не на основании формы слова, а исходя из контекста [Лу Исинь, 2016: 23].

Основным содержанием данного параграфа стало описание особенностей китайской корпусной лингвистики. Были выделены временные рамки развития данного направления в Китае, а также наиболее известные и масштабные корпусы китайского языка. В ходе описания была сформулирована цель создания текстовых корпусов на китайском языке, побудившая ученых начать работу в данной области. Кроме того, текст параграфа содержит описание проблем, с которыми могут столкнуться исследователи в процессе создания корпуса на китайском языке.

1.3. Основы построения текстового корпуса

Процесс создания текстового корпуса состоит из нескольких этапов. С развитием технологий и появлением сети Интернет он был модернизирован, некоторые этапы стали автоматизированы, что значительно упростило создание корпуса и сократило время работы над ним. Пользуясь современными корпусами лингвист способен работать с большими объемами текстов в короткие сроки, что позволяет охватить больший объем информации, чем это было в прошлом столетии. Кроме того, ввиду задействованных объемов информации, новые корпусы являются наиболее репрезентативными, если мы рассматриваем их с точки зрения масштаба. Как было сказано выше, сбалансированность также является неотъемлемой характеристикой показательного корпуса. Она отвечает за наполненность корпуса необходимым объемом лексических единиц и отсутствие лишней информации, засоряющей корпус. В целях разработки текстового корпуса, обладающего данными характеристиками и объективно отражающего реальность, но не

субъективное мнение автора, необходимо пройти через несколько этапов на пути его создания.

Общее количество этапов, необходимых для создания текстового корпуса, разнится в зависимости от ученого: в отдельных случаях исследователи выдвигают два этапа: сбор коллекции текстов и ее разметка [Глазкова, 2019: 3]. В данной работе мы рассмотрим более подробный алгоритм создания корпуса, предложенный В.П. Захаровым. Всего он включает в себя девять шагов или этапов, каждый из которых является неотъемлемой частью процесса разработки корпуса.

Первым шагом на пути создания корпуса является определение перечня источников и обеспечение поступления текстов в соответствии с данным перечнем. Подразумевается, что автору необходимо изначально сформировать цель, с которой будет создаваться корпус. Исходя из поставленной цели, автор начинает формировать список литературы или каналов, из которых будет взята информация для будущего корпуса.

После подбора текстов следует их преобразование в машиночитаемый формат. На данном этапе тексты могут быть получены из разных источников, таких как ручной ввод, сканирование, авторские копии, дары и обмен, Интернет, оригинал-макеты, предоставляемые составителям корпусов издательствами. Следовательно, для работы с текстами разного формата необходимо отредактировать их по единому образцу. Стоит отметить, что оцифровка текстов не представляет ни малейшей трудности в случае, если осуществляется с использованием современных технологий оптического ввода информации и распознавания текстовой информации.

На третьем этапе тексты проходят предварительную обработку, следовательно автор корпуса осуществляет филологическую выверку и корректировку текстов. Также осуществляется подготовка библиографического и экстралингвистического описания текста.

После выверки и корректировки текста следует его конвертирование и графематический анализ. В зависимости от текста, он может подвергаться

данному процессу один или несколько раз, в результате которого осуществляются различного рода перекодировка (если требуется), удаление или преобразование нетекстовых элементов (рисунки, таблицы), удаление из текста переносов, обеспечение единообразного написания тире. Как правило, эти операции выполняются в автоматическом режиме. Обычно на этом же этапе осуществляется сегментирование текста на его структурные составляющие.

Разметка текста – следующий и один из наиболее важных этап создания корпуса. Она заключается в приписывании текстам и их компонентам дополнительной информации, то есть метаданных. Метаданные можно поделить на три типа: экстралингвистические, относящиеся ко всему тексту; данные о структуре текста; лингвистические метаданные, описывающие элементы текста. Метаописание текстов корпуса включает как содержательные элементы данных (библиографические данные, признаки, характеризующие жанровые и стилевые особенности текста, сведения об авторе), так и формальные (имя файла, параметры кодирования, версия языка разметки, исполнители этапов работ). Эти данные обычно вводятся вручную. Структурная разметка документа (выделение абзацев, предложений, слов) и собственно лингвистическая разметка обычно осуществляются автоматически.

Как правило, автоматическая разметка не всегда является точной и требует корректировки, поэтому следующий этап заключается в редактировании результатов автоматической разметки. К нему относятся исправление ошибок и снятие неоднозначности (вручную или полуавтоматически).

После успешного завершения внесения правок в результаты автоматической разметки автор корпуса приступает к конвертированию размеченных текстов в структуру специализированной лингвистической информационно-поисковой системы (corpus manager). Оформление корпуса в данный формат обеспечивает быстрый многоаспектный поиск и статистическую обработку как одну из основных функций корпуса.

Когда корпус полностью отредактирован и загружен в специализированную программу, автору необходимо обеспечить доступ к результатам своей работы. Таким образом, созданный корпус будет полезен как самому автору, так и другим исследователям в процессе сбора и анализа материала по определенной теме. Корпус может быть доступен в пределах дисплейного класса, может распространяться на компакт-диске и может быть доступен в режиме глобальной сети. Различным категориям пользователей могут предоставляться разные права и разные возможности.

Последним и завершающим этапом на пути создания текстового корпуса становится создание документационного обеспечения, в котором описываются различные аспекты использования корпуса, в частности, приводятся сведения о разметке, позволяющие искать по метаданным, язык запросов корпус-менеджера и т.д.

Рассмотренный алгоритм, разработанный В.П. Захаровым, включает наибольшее количество этапов для создания корпуса, однако оставляет возможным внесение изменений в данную последовательность [Захаров, 2020]. Каждый из шагов создает условия для работы на следующем этапе, что позволяет успешно завершить работу над корпусом. Наиболее ответственными моментами в процессе разработки являются разметка и конвертирование готового корпуса в специальный формат для дальнейшей работы с ним. Правильно подобранные тексты позволят с большей точностью проанализировать то или иное явление и сделать соответствующие выводы.

1.4. Китайская диалектология как объект лингвистических исследований

Диалектология представляет собой науку, занимающуюся изучением территориальных разновидностей языка, то есть диалектов, в их синхронном состоянии и историческом развитии. Термин «диалектология» восходит к

греческим корням: dialektos – говор, диалект; logos – учение, знание [Касаткин, 2005: 1].

Согласно классификации Л.Л. Касаткина, основными единицами диалектологии являются:

1. Говор – наименьшая единица диалектного членения. Представляет собой язык одного или нескольких соседних населенных пунктов, равных в языковом отношении.

2. Диалектный язык – совокупность говоров, где каждый из них может рассматриваться отдельно.

Как правило, диалекты противопоставляются литературному языку, который является языком культуры: политики, искусства, науки. Диалекты же обычно служат языками сельского населения, на их основе строится народное творчество [Касаткин, 2005: 5].

В китайском языке тоже существует понятие литературного языка и целый ряд противопоставленных ему диалектов. Целью китайской диалектологии, как пишет О.И. Завьялова, является «сбор всеобъемлющей информации о существующих территориальных разновидностях китайского языка самого разного уровня» [Завьялова, 2012: 134].

В работах российских лингвистов диалекты китайского языка обладают рядом особых характеристик. Основанием утверждения фонетических различий, характерных для диалектов китайского языка, является работа А.А. Рыбникар. Разным диалектам присуща разная лексика, различается грамматика. Несмотря на это, основы словарного состава языка и грамматики остаются едины, на них и базируются диалекты. Литературный язык опирается на северные диалекты, а фонетической нормой является пекинское произношение [Рыбникар, 2013: 3]. Согласно М.В. Софронову, «диалект является средством общения внутри исторически сложившейся области устойчивого общения, которая формируется комплексом языковых и этнографических границ» [Софронов, 2007: 241]. А.Н. Алексахин в своей книге «Диалект Хакка» пишет о том, что представители разных диалектных

групп с большим трудом взаимодействуют друг с другом, либо же их устная коммуникация полностью невозможна [Алексахин, 1987: 13].

В китайском языке существует специальный термин для обозначения диалекта – 方言 / fāngyán. История данного термина, который дословно переводится как «местные слова», уходит корнями глубоко в древность. Первые его упоминания датируются эпохой Хань: он использовался в названиях словарей и противопоставлялся термину 正音 / zhèngyīn «правильное произношение». В значении «диалект», каким мы привыкли его обозначать, он стал применяться только в начале XX века в работах западных ученых.

На данный момент существуют мнения, что термин 方言 / fāngyán следует употреблять в значении «местный говор», так как он не выражает тех качеств, которые заложены в понимании европейского диалекта [Баров, 2019: 153]. Кроме того, как говорит С.А. Баров, «диалект – это не просто доступный с детства и понятный инструмент общения, познания и взаимодействия с внешним миром». В него еще входят и культурные особенности определенных регионов – специфическая лексика, характерная для отдельной местности [Баров, 2019: 156].

Китайская цивилизация является одной из древнейших в мире, по утверждениям ученых, ее возраст может насчитывать около пяти тысяч лет. Одни только письменные источники охватывают период не менее трех тысяч лет. Следовательно, совершенно естественно, что за столь долгий срок китайская культура накопила огромное количество языковых особенностей. Диалекты китайского языка являются ярким примером данного явления.

На протяжении многих веков население Китая говорило на многочисленных диалектах. В силу неграмотности простого народа, диалекты были исключительно устными – люди попросту не умели пользоваться письменными принадлежностями и не видели в этом необходимости. Ссылаясь на работы ученых, А.Н. Алексахин говорит о том, что причиной

значительного углубления диалектной раздробленности является длительное господство феодалов в Китае [Алексахин, 1987: 11]. Таким образом, в стране преобладал крестьянский уклад жизни, большая часть населения трудилась на земле и не собиралась заниматься образованием, а для ежедневного общения было достаточно знания устного языка.

Согласно словам Е.Б. Астрахан, природные ресурсы также сказывались на формировании диалектов, в частности водораздельные хребты, которые служили границами поселений. Реки, наоборот, не выступали границами, представляли собой наиболее благоприятные места для поселения [Астрахан, 1985: 14].

Следующим немаловажным фактором образования диалектов являются исторические события, связанные с демографическими и этническими процессами, выпавшими на судьбу китайского народа. Под этим подразумеваются многочисленные миграции, войны и изменения границ территорий, которые оказали значительное влияние на развитие языка и его диалектов.

В средневековом Китае отсутствовал общегосударственный устный язык. Преимущественно на севере единым литературным языком считался вэньянь 文言 / wényán – для его понимания достаточно было знать лексику и грамматику. Он до сих пор считается языком классических произведений китайской литературы, а рассвета достиг в эпоху Тан (VII — начало X в.). Позднее его вытеснил байхуа 白话 / báihuà, став новым литературным языком на основе разговорного. На юге ситуация развивалась другим образом – существовало несколько изолированных областей устойчивого общения. Там и зародились наиболее известные на данный момент южные диалектные группы [Там же, 1985: 15]. Исходя из вышесказанного, история до XX в. характеризуется разрозненностью двух частей Китая: создание байхуа на севере и укрепление диалектов на юге.

Изучение исторических записей показывает, что правительство Китая в XX в. стремилось унифицировать диалекты и распространить единый национальный язык – путунхуа [Там же, 1985: 28]. Однако, как мы можем наблюдать, полной замены диалектов не произошло, они до сих пор продолжают существовать как в устной, так и в письменной форме. В последние годы все чаще возникают вопросы о сохранении особенностей того или иного диалекта.

В настоящее время в Китае существует несколько групп наиболее распространенных диалектов. В соответствии с классификацией О.И. Завьяловой к ним относятся:

1. Северная группа диалектов во главе с пекинским диалектом.
2. Кантонская группа с основным диалектом этого региона кантонским – (гуанчжоуским).
3. Диалекты У (группа восточных диалектов – провинции Цзянсу и Чжэцзян), где основными являются сучжоуский и шанхайский.
4. Миннаньская группа диалектов во главе с сямынским диалектом города Сямынь, близким тайваньскому. Другое его название – «амойский диалект».
5. Хунаньская группа, в основе которой лежит диалект города Чанша – столицы провинции Хунань.
6. Группа диалектов Хакка, в которой главенствующую роль занимает диалект города Мейчжоу.
7. Группа диалектов провинции Цзянси и основной в этой группе – диалект города Наньчан.

Каждая из перечисленных выше групп включает некоторое количество диалектов, более или менее распространенных, обладающих географическими границами и культурным значением.

Тем не менее диалектная ситуация в стране на данный момент неоднозначна. Опираясь на мнение российских ученых, становится ясно, что с возникновением политики, направленной на распространение путунхуа,

явно прослеживается идея пренебрежения интересами отдельных людей ради «общего блага» [Баров, 2019: 152]. В частности, численность владеющих диалектом Хакка как неродным изменилась за десятилетия – наблюдается тенденция к снижению с 3,6% до 3,5% [Гутин, 2018]. И хотя в конституции КНР существует статья, подтверждающая право населения «пользоваться своими языками и развивать их письменность», в стране активно протекает процесс вытеснения диалектов национальным языком [Баглаева, 2021: 238]. В последние годы становится актуальным лозунг 说普通话、用规范字、做文明人 / shuō pǔtōnghuà, yòng guīfàn zì, zuò wénmíng rén, что в переводе означает «говорите на путунхуа, пользуйтесь стандартными иероглифами и будете культурным человеком». Подразумевается, что каждый, кто не пользуется путунхуа повсеместно, становится человеком низкого социального или образовательного уровня [Баров, 2019: 156].

В то же время, исследуя работы китайских авторов, мы наблюдаем противоположную ситуацию: диалекты не только не вытесняются, наоборот, им уделяется особое внимание. На сайте Министерства Образования КНР закреплен документ, подтверждающий эту позицию. Посыл данного документа следующий: 方言与普通话互补共荣 / fāngyán yǔ pǔtōnghuà hùbǔ gòng róng, что переводится как «диалекты и путунхуа взаимно дополняют друг друга и вместе процветают». В китайских школах и университетах проходят «уроки диалектов», направленные на распространение и приобщение молодого поколения к пользованию диалектной лексикой. Кроме того, в китайском интернет-пространстве встречаются такие лексемы, как 方言梗 / fāngyángěng «шутка, основанная на различиях в диалектах», 方言保护 / fāngyán bǎohù «народная политика поддержки диалектов», 方言卫士 / fāngyán wèishì «прозвище для защитника диалектов» и др. Все это свидетельствует о положительной тенденции поддержания развития диалектов.

Диалекты содержат множество устойчивых выражений или лексем, присущих определенному региону. При утрате языка теряется и часть

культуры, которая так важна при описании многоликости китайской цивилизации. Согласно С.А. Барову, исчезновение части культуры станет серьезной проблемой для страны, так как в Китае отсутствует объединяющее религиозное начало, «именно культура для китайцев становится неким общим источником их исторической уникальности» [Баров, 2019: 156]. К числу достижений в «битве» за сохранение диалектов можно отнести кантонскую оперу или музыкальные представления на южнофуцзяньских диалектах, включенные в список нематериального культурного наследия КНР и ЮНЕСКО. Однако всем неравнодушным к судьбе диалектов Китая предстоит еще много работы.

ВЫВОДЫ ПО ГЛАВЕ 1

В данной главе была предпринята попытка описать основы корпусной лингвистики и диалектологии китайского языка. Были установлены временные рамки зарождения двух научных направлений, отмечены основные переломные моменты, такие как изобретение компьютеров в сфере корпусной лингвистики или переход от 文言/ wényán веньяня к 白话 / báihuà байхуа в китайском языке. В рамках корпусной лингвистики был установлен понятийный аппарат данного научного направления, а также была рассмотрена проблема его неокончательной сформированности. Были упомянутые основные корпусы, такие как Национальный корпус русского языка. Кроме того, в работе были представлены особенности китайской корпусной лингвистики, в частности, были приведены примеры наиболее известных и масштабных корпусов китайского языка. Проблемой, долгие годы сдерживающей развитие данного направления в Китае, стало отсутствие иероглифической письменности в компьютерной среде, однако сейчас создание корпусов на базе китайского языка стало возможным, что вызвало стремительный рост популярности среди исследователей. Пользуясь разработками западных ученых, китайское научное сообщество адаптирует их под себя и создает свои инструменты для развития данного научного направления.

Помимо вышеперечисленного, в исследовании был описан алгоритм создания корпуса, разработанный В.П. Захаровым, который включает в себя основные шаги на пути создания корпуса текста. Всего их насчитывается девять, начиная с подбора источников и поиска текстов, заканчивая выгрузкой готового корпуса на электронный носитель. Данный алгоритм был использован в практической части нашей работы для конструирования диалекта Хакка китайского языка.

Помимо вопросов корпусной лингвистики, в работе были рассмотрены особенности китайской диалектологии, так как именно на основании одного

из китайских диалектов строится наш корпус. Было дано понятие диалекта, приведена классификация диалектов в зависимости от региона страны и представлена диалектная ситуация в современном Китае. Наряду с защитниками культурного наследия страны, в том числе диалектов, существует большое количество сторонников путунхуа, которые выступают против сохранения диалектов. Такое влияние распространяется на разные сферы жизни, в том числе школу или работу. Данное явление свидетельствует о неоднозначности диалектной ситуации в стране. Тем не менее насчитывается ряд относительно недавних работ, целью которых является фиксация диалектных особенностей на разного рода носителях. Это позволяет сделать вывод, что ситуация скорее направлена в положительное русло, в котором официальный язык гармонично сосуществует с диалектами разных провинций.

Таким образом, оба научных направления, описанные в теоретической части работы, являются актуальными, так как в их рамках существует ряд неразрешенных проблем и присутствуют перспективы для дальнейшего развития.

ГЛАВА 2. КОНСТРУИРОВАНИЕ И АНАЛИЗ ТЕКСТОВОГО КОРПУСА ДИАЛЕКТА ХАККА КИТАЙСКОГО ЯЗЫКА

2.1. Процесс создания текстового корпуса диалекта Хакка китайского языка

Текстовый корпус китайского языка в значительной степени отличается от корпусов на русском языке. При создании данного корпуса перед исследователем возникает множество вопросов, на которые необходимо дать ответ. Например, как правильно определить часть речи при разметке китайского текста? В процессе сегментации, ввиду отсутствия пробелов, каким образом определить потенциальные границы между иероглифами? Стоит ли анализировать каждый отдельно взятый иероглиф или необходимо исследовать его значение в словосочетании [Довнар, 2011: 203]. В ходе данной работы мы постарались ответить на поставленные вопросы и предоставить научное обоснование. В ходе построения корпуса мы придерживались алгоритма, который был предложен В.П. Захаровым и описан выше. Мы находим его наиболее развернутым и подробно описанным.

2.1.1. Определение перечня источников текстового корпуса диалекта Хакка

Целью нашего исследования является разработка текстового корпуса диалекта Хакка и дальнейший анализ собранных данных. С учетом данной цели, в перечень источников текста работы были включены видеоролики с китайского видеохостинга Bilibili. Отбор материала проходил в соответствии со следующими критериями:

1. Видеоролик содержит текст на диалекте Хакка и на путунхуа (официальный язык Китая). Данное требование выдвинуто в целях упрощения понимания приведенного в видеоролике текста.

2. Для видеоролика подготовлены субтитры на китайском языке. Такое требование позволяет наглядно оценить использование иероглифики, в том числе проследить различия с официальным китайским языком. На результатах анализа субтитров частично строится вывод о связи диалекта Хакка с древнекитайским языком.

3. Текст видеоролика озвучен носителем языка. Голосовое сопровождение текста позволяет выявить фонетические различия в двух языковых системах и составить представление о фонетических особенностях диалекта Хакка.

4. Героями видеоролика являются представители китайской нации. Данный критерий выдвинут с целью отобрать материал на диалекте Хакка, который наиболее приближен к настоящему диалекту.

Рисунок 1. Образец видеоролика на платформе Bilibili

В результате отбора видеороликов в соответствии с выдвинутыми критериями в перечень источников было включено 30 видео. Все они содержат текст на диалекте Хакка, обладают звуковым сопровождением, к ним

подготовлены субтитры и персонажами видео являются носители китайского языка. Такой отбор позволяет наиболее объективно осуществить выборку видео на диалекте Хакка.

2.1.2. Процесс обработки текста: оцифровка и корректировка

Обработка текста состоит из нескольких этапов. Первый из них – это оцифровка текстов, то есть их преобразование в компьютерную форму. На данном этапе мы воспользовались программой ELAN, которая была создана для аннотирования звуковых и видеофайлов.

После завершения установки данной программы, мы постепенно загружали в нее отобранные видеоролики и преобразовывали их в текст. На рисунке ниже представлена панель управления программы ELAN. Программа предоставляет пользователю возможность совершать различные операции, например: регулировка скорости и громкости видеофайла; автоматический распознаватель речи; функция добавления таблиц, создания субтитров, оставления комментариев.

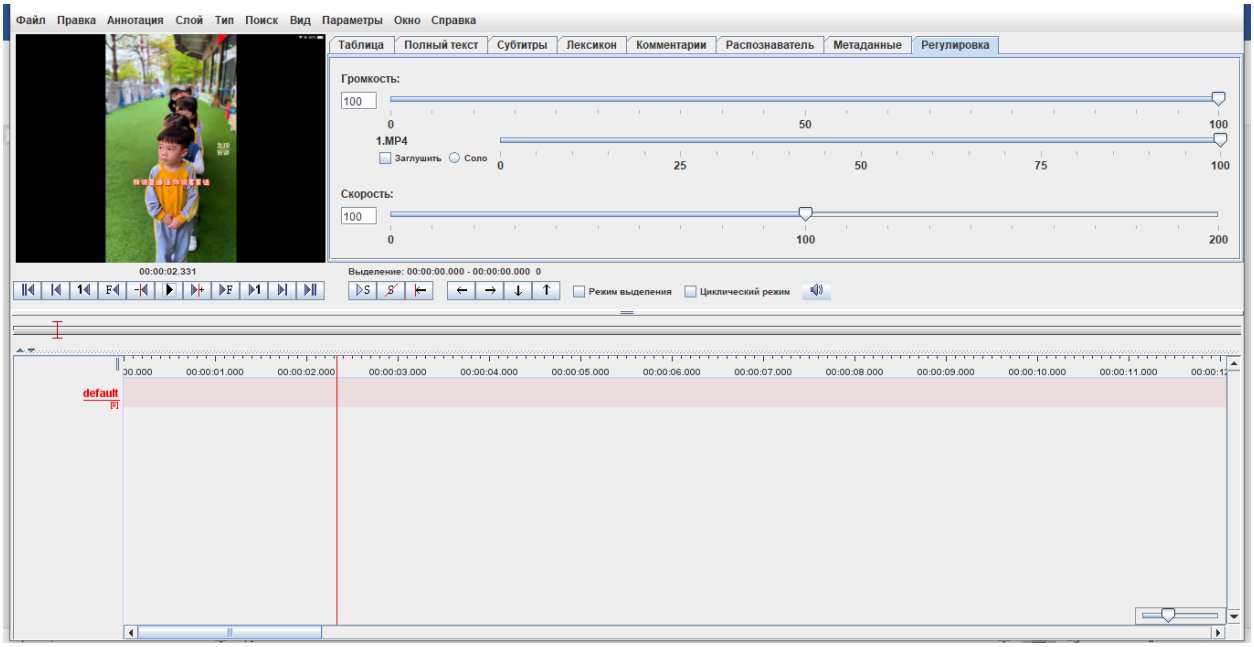


Рисунок 2. Панель управления программы ELAN

После загрузки видео пользователю доступна возможность установить количество спикеров, участвующих в видео. Для этого необходимо в строке «Тип» создать новый тип, в нашем случае мы назвали его «Диалог», так как в видеоролике идет диалог между воспитателем и детьми. Создав новый тип, мы перешли к добавлению участников диалога: для этого в разделе «Новый слой» мы создали отдельную строку для детей и воспитателя.

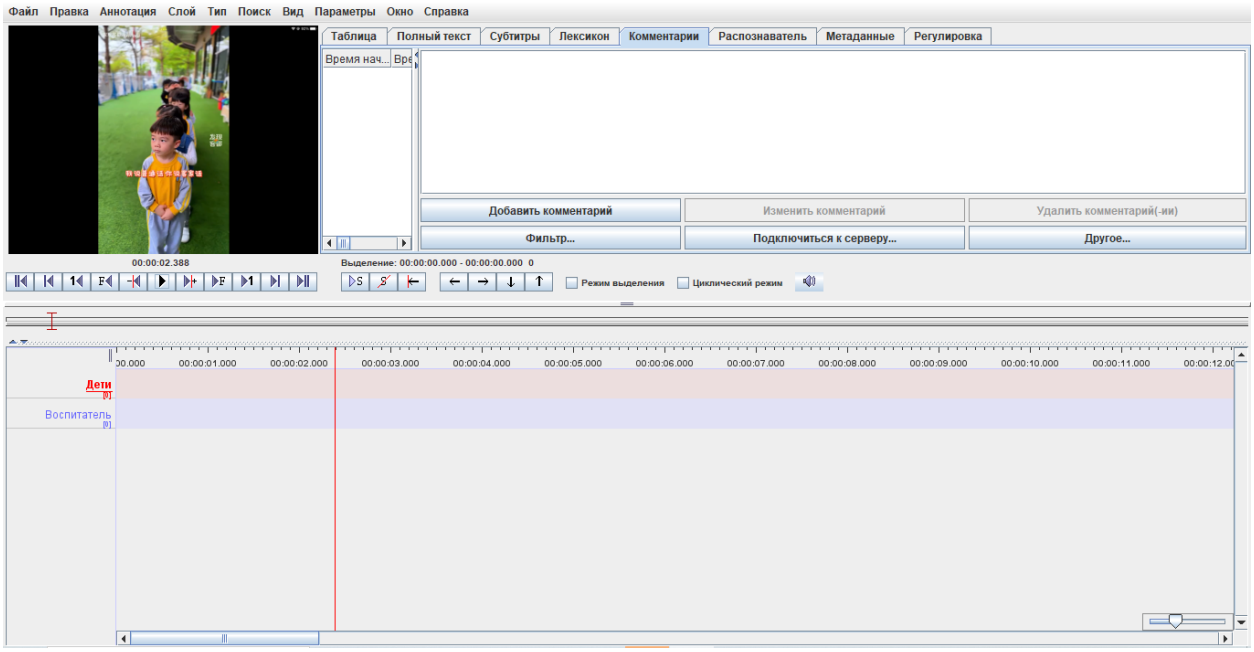


Рисунок 3. Добавление новых спикеров

После того как для каждого из участников видеофрагмента была создана отдельная строка, мы приступали к процессу аннотирования документа, то есть добавлению реплик говорящих. Для это было необходимо перейти в режим сегментации документа и вручную добавить комментарий для каждой из реплик. Завершив процесс добавления сегментных окон для каждого высказывания, мы переходили в режим транскрипции, где вносили правки в созданные комментарии. В конечном итоге, для каждого спикера мы подготовили ряд комментариев, в которых были прописаны готовые реплики каждого.

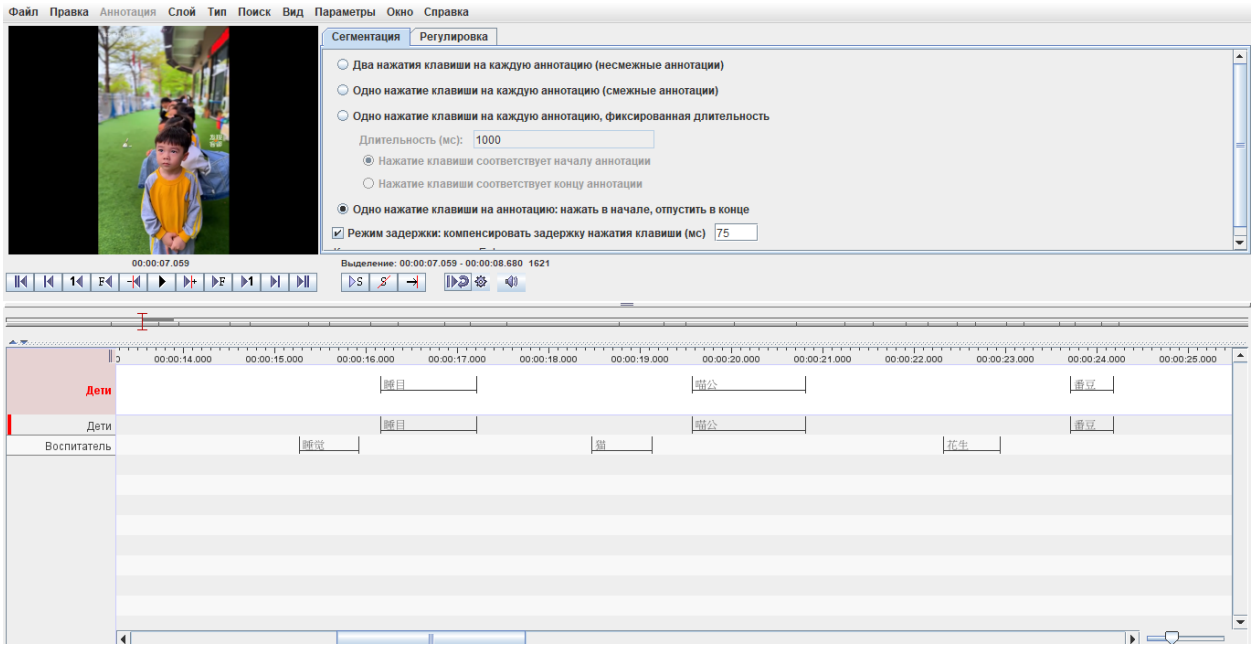


Рисунок 4. Процесс сегментации

Закончив работу над делением текста на сегменты, мы получили готовый текст, поделенный на отдельные реплики с указанием точного времени, когда высказывание было произнесено. Итоговый файл по правилам программы сохраняется в формате TXT, после чего его необходимо преобразовать в привычный формат WORD. В конце концов, мы получили такой скрипт.

```
Воспитатель 螃蟹
TC    00:00:11.271 - 00:00:11.828
SD    (0.29)


Дети  捞嗨
TC    00:00:12.119 - 00:00:13.050
SD    (2.22)


Воспитатель 睡觉
TC    00:00:15.270 - 00:00:15.929
SD    (0.23)

Дети  睡目
TC    00:00:16.168 - 00:00:17.210
SD    (1.26)


Воспитатель 猫
TC    00:00:18.477 - 00:00:19.133
SD    (0.43)
```

Рисунок 5. Образец готового документа после обработки

2.1.3. Этап разметки текста

Следующим этапом после перекодировки текста в необходимый формат и последующей его редактуры предстоит немаловажный процесс разметки. На данном этапе текст уже разделен на отдельные сегменты, которыми в нашем случае являются слова, словосочетания или предложения в зависимости от исходных данных, представленных в видеоролике.

В нашем корпусе была осуществлена частеречная и фонетическая разметка как целого слова и отдельного иероглифа. На данном этапе мы столкнулись с проблемой определения наименьшей единицы текста, то есть токена данного корпуса. Связано это с описанной выше особенностью, характерной для китайского языка – одни и те же иероглифы могут употребляться как самостоятельно, так и составе сложных слов, в зависимости от контекста возможно изменение и части речи конкретного иероглифа. На данном этапе конструирования корпуса выделять все возможные варианты чтения отдельного иероглифа и указывать целый ряд частей речи, в виде которых он может выступать, представляет для нас трудность. Тем не менее это может стать отличной перспективой развития данного корпуса.

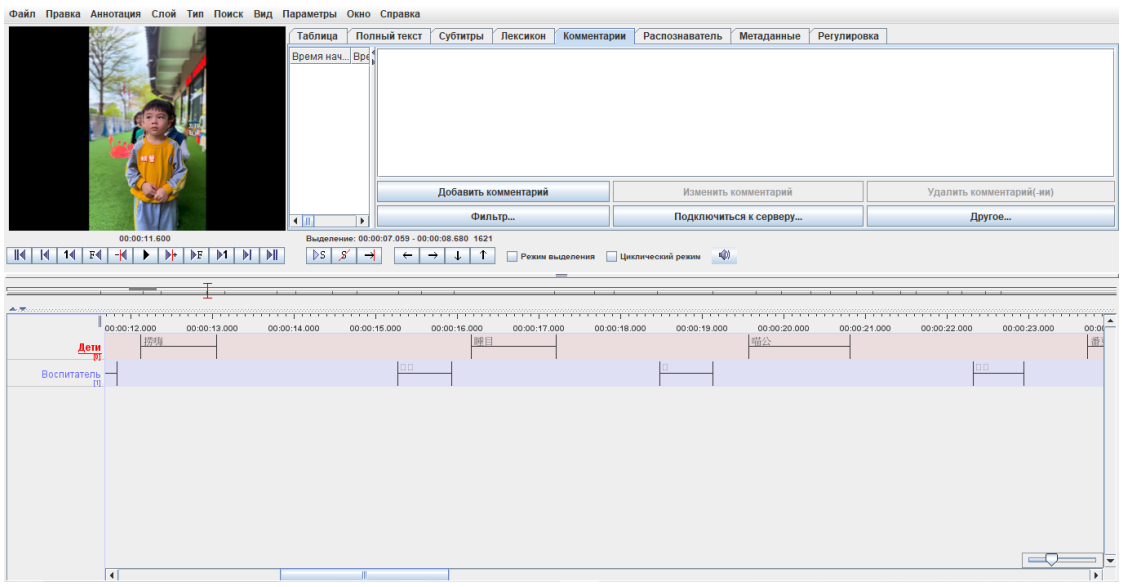
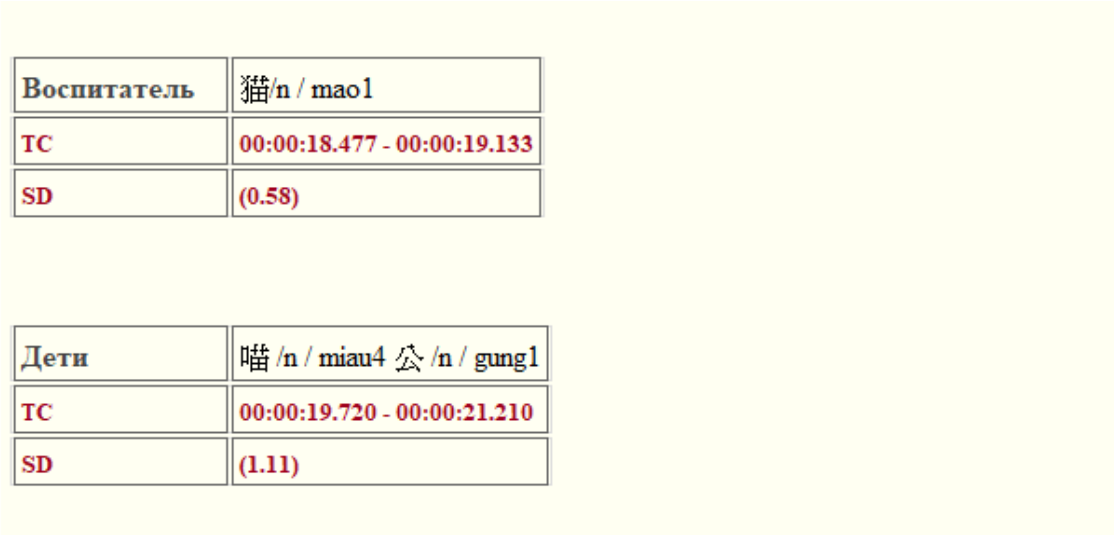


Рисунок 6. Режим разметки текста

Для осуществления разметки в программе ELAN необходимо непосредственно перейти в режим разметки документа. Далее, к ранее созданным сегментам, необходимо оставить комментарий, таким образом осуществить разметку. К иероглифам, представленным в каждом отдельном сегменте, мы добавили транскрипцию и указали их часть речи. Данные о фонетических особенностях слова и конкретные примеры употребления лексической единицы в контексте предложения были взяты из онлайн-словаря диалекта Хакка Syndict.com [Syndict, 2008]. В результате для каждого слова получилась подобная разметка.



Воспитатель	猫/n / mao1
TC	00:00:18.477 - 00:00:19.133
SD	(0.58)

Дети	喵 /n / miau4 公 /n / gung1
TC	00:00:19.720 - 00:00:21.210
SD	(1.11)

Рисунок 7. Пример готовой разметки текста

На примере иероглифа 猫 / māo «кот» мы можем наблюдать, как изменилось аналогичное ему слова в диалекте Хакка: лексическая единица, состоящая из одного иероглифа, стала двусложной, кроме того значительно изменилось чтение. В данном случае мы наблюдаем сочетания инициалей и финалей, нехарактерных для путунхуа. Помимо этого, написание самого иероглифа 猫 претерпело изменения: ключ «рука» был заменен на ключ «рот». Подобные отличия будут описаны ниже как выделенные нами особенности, однако на этапе разметки мы определили только часть речи и отметили вариант произнесения иероглифа.

2.1.4. Оформление размеченных данных в корпус

Этапом, следующим после разметки текста, следует выгрузка готовых данных в специализированную программу для формирования полноценного корпуса. Наиболее важным на данной стадии, на наш взгляд, является правильно подготовленная разметка, так как в противном случае программа не сможет верно сконструировать корпус.

Для создания корпуса мы использовали программу Sketch Engine. Данная программа позволяет создать корпус и является платформой, на которой он будет храниться. Посредством Sketch Engine у исследователей появляется возможность изучить различные языковые явления и особенности, воспользоваться трудами других ученых и поделиться своими. После создания корпуса, в рамках этой программы можно проследить, в каком контексте чаще встречается слово, а также определить ключевые слова текста.

Первым шагом на пути конструирования нашего корпуса стала необходимость определения языка. Кроме того, в появившемся окне нам предлагалось выбрать тип нашего корпуса. В связи с тем, что наш корпус базируется на диалекте Хакка, было принято решение создать параллельный корпус в целях удобства при анализе информации.

Рисунок 8. Регистрация нового корпуса

После того как мы определились с типом корпуса, нам была предложена возможность загрузить тексты. На выбор были предоставлены две опции:

загрузить обработанные тексты с готовой разметкой или отправить на обработку неподготовленный материал. Несмотря на то, что текст нашего корпуса был предварительно обработан, данные в формате TXT возможно было загрузить только как неподготовленные.

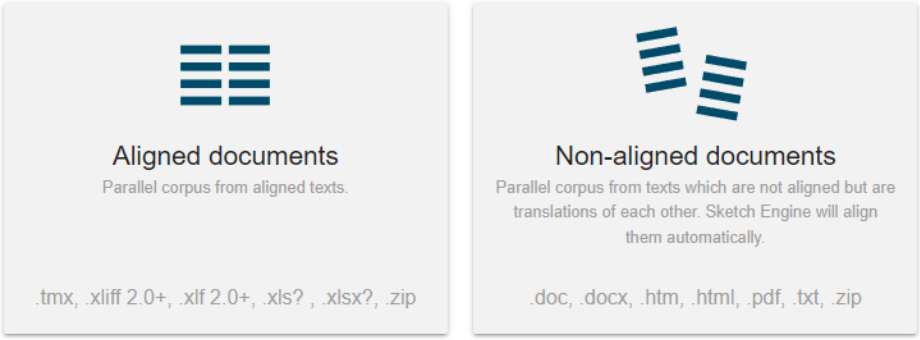


Рисунок 9. Загрузка текстов

После этого нам необходимо было разделить наши тексты на путунхуа и Хакка по отдельным документам. Это связано с тем, что корпус является параллельным, в связи с чем возникает необходимость загружать два отдельных документа. Таким образом, мы отделили диалект от официального языка и загрузили в программу. После того, как тексты были загружены и обработаны, у нас появилась возможность воспользоваться различными функциями Sketch Engine. Например, мы смогли осуществить параллельный поиск по корпусу. В данном случае мы решили посмотреть использование иероглифа 好 / hǎo «хорошо» в Хакка и путунхуа. Нам были представлены варианты использования данного иероглифа в составе других слов.

doc#0	<s> 你 / n/ni3 好 /a/hao3 ! </s>	<s> 你 / n/n 3 好 / a / hau4 ! </s>
doc#0	<s> 你 好 e/ni3hao 3 ! </s>	<s> 你 好 e/n 3 hau4 ! </s>
doc#0	<s> 大/a/da4 家 /n/jia1 好 /n/hao 3 ! </s>	<s> 大/a/tai4 家 /n/ga 1 好 / n / hau3 ! </s>
doc#0	<s> 大家 好 e/da4jia1 hao 3 ! </s>	<s> 大家 好 e/tai4 ga1hau3 ! </s>

Рисунок 10. Параллельный поиск по корпусу

Кроме того, вышеупомянутая программа позволяет отследить частотность употребления того или иного знака. В настройках можно

установить определенные параметры. Например, выбирать слова только указанной части речи или вести подсчет только тех лексических единиц, которые начинаются с определенной буквы. На практике оказалось, что в нашем корпусе программа засчитывала не только иероглифы, но и пунктуационные знаки, а также чтение иероглифов.

Помимо вышеперечисленных функций, нам была предоставлена возможность определять ключевые слова в корпусе. Благодаря данной функции становится возможным установить, какие слова являются наиболее частотными. Данная опция похожа на предыдущую, связанную с частотностью слов, однако она не показывает конкретные числа.

Таким образом, после загрузки размеченных документов, программа самостоятельно произвела конструирование корпуса и позволила воспользоваться различными функциями для анализа текстов. На данном этапе корпус считается практически готовым, так как по нему уже возможно осуществлять поиск и проводить исследования. Тем не менее, для завершения конструирования корпуса, необходимо сделать его общедоступным.

2.1.5. Предоставление доступа к созданному корпусу

Завершающим этапом в создании корпуса является его выгрузка на цифровой носитель, доступным общественности. В нашем случае с решением этой задачи не возникло проблем, так как сама программа способствует решению данного вопроса. В Sketch Engine в разделе инструментов есть функция «Поделиться», которая предоставляет возможность сделать корпус открытым для других пользователей. Ниже представлен рисунок, демонстрирующий панель управления программы.

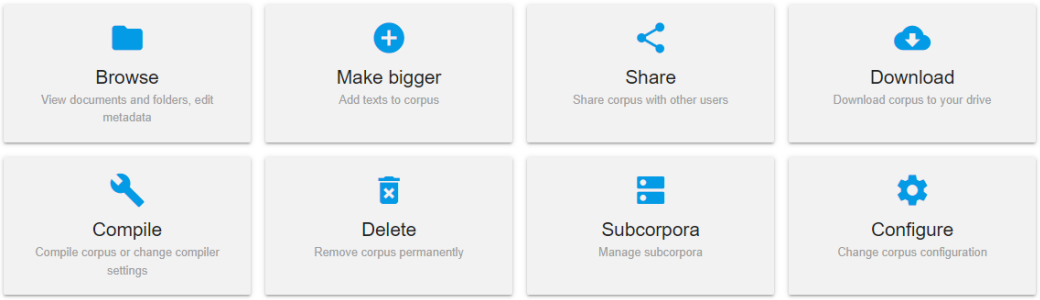


Рисунок 11. Панель управления корпусом

Для того чтобы поделиться собственными наработками, достаточно ввести личный email-адрес человека, группового аккаунта или института. Впоследствии пользователи, обладающие доступом к данной почте, смогут воспользоваться разработанным корпусом.

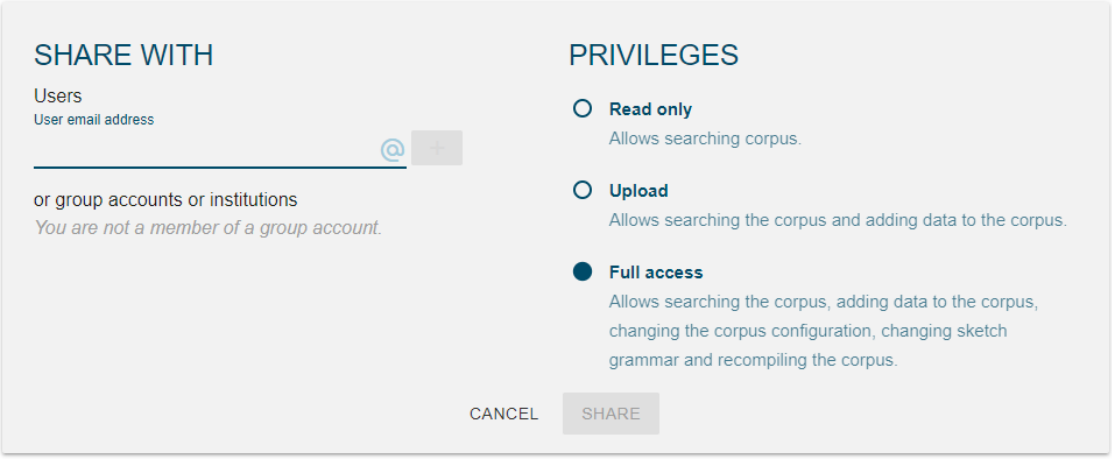


Рисунок 12. Открытие доступа к корпусу

Подводя итоги создания нашего корпуса, стоит отметить, что этапы, предложенные в алгоритме В.П. Захарова, не всегда были раздельно описаны в данной работе. Это связано с тем, что работа над созданием корпуса является процессом, стадии которого тесно связаны друг с другом. Поэтому в данном исследовании некоторые этапы были представлены вместе, без четкого разграничения. Тем не менее корпус диалекта Хакка на основе видеороликов сети BiliBili был создан на основе программы Sketch Engine. Результаты анализа данного корпуса представлены в следующем параграфе.

2.2. Анализ лингвистических особенностей диалекта Хакка на материале созданного корпуса

Одним из названий диалекта Хакка до сих пор является 唐语 / tángyǔ «язык эпохи Тан». Такое наименование отражается в фонетической составляющей диалекта, в том числе в особенностях чтения, произношения отдельных звуков, в отличиях тональной системы. Диалектная лексика также отличается от стандартизированной путунхуа и уходит корнями в древность. В данной работе мы рассматриваем вышеперечисленные отличительные черты, основываясь на материале видеороликов платформы Bilibili.

2.2.1. Особенности речевого потока Хакка

Анализируя фонетический состав диалекта Хакка, прежде всего стоит обратиться к понятию речевого потока. Под термином «речевой поток» подразумевается «звучащая речь в ее линейной последовательности, представленная звуками речи, употребляемыми в составе слога, слова и фразы» [Жеребило, 2010: 309]. Рассматривая речь носителей диалекта, представленную в коротких видеороликах, прежде всего следует обратить внимание на систему тонов. Основным отличием мы выделили расхождения в тональном значении, они представлены в таблице ниже.

Таблица 1. Различия в тональном значении диалекта Хакка

Тон	Тональное значение
Ровный тон (陰平, 阴平)	45
Восходящий тон (陽平, 阳平)	213
Нисходяще-восходящий тон (上聲, 上声)	31
Нисходящий тон (去聲, 去声)	53
Верхний входящий тон (陰入, 阴入)	2
Нижний входящий тон (陽入, 阳入)	5

Входящий тон диалекта Хакка представляет особый интерес, так как значительно отличается от остальных. Произносится он очень быстро, по форме значительно короче. Его особенностью является построение слога: инициалью является гласный звук (a, o, i, e), на месте финали располагается согласный звук (b, d, g). Таким образом, мы получаем следующие сочетания звуков: ab, ad, ag, od, og, ib, id, eb, ed. При чтении согласный звук не произносится, однако, произнося начальный гласный звук, следует вытянуть губы так, будто бы говорящий собирается произносить согласный звук. Следовательно, мы слышим слог, в котором гласный звук ограничен попыткой произнести согласный звук.

Говоря о произношении, диалект Хакка богат звуками с придыханием. Данное явление осуществляется независимо от тона и обусловлено древними аффрикатами. К числу аспиратов относятся звуки p, t, k, которые встают на место привычных нам в путунхуа b, d, g: 白色 / pag5sed1, 棉被 / mien11pi44, 办事 / pan52sii52, 大家 / tai52ga44, 代表 / toi52biau31, 口袋 / keu31toi52, 共同 / kiung52tung11, 跪拜 / kui52bai52, 书柜 / su44kui52. Помимо вышеперечисленных аспиратов, придыханием обладают звуки q, c. Чаще всего они располагаются на месте тех согласных, где в путунхуа ставятся j, z: 尽量 / qin52liong52, 成就 / siin11qiu52, 逐渐 / cug5qiam52, 座位 / co52vi52, 木凿 / mug1cau52, 罪过 / cui52guo52.

Помимо аспирации, для фонетического строя диалекта Хакка характерно частое употребление зубных согласных звуков: в путунхуа распространен фрикативный звук h, в Хакка он заменяется лабиодентальным звуком f: 呼唤 / fu44fon52, 胡来 / fu11loi11, 蝴蝶 / fu11tiab5. Такая особенность восходит ко временам династии Сун и связана с древним произношением: в процессе трансформации некоторые звуки сливались, образуя новый. Например, звук g соединился со звуком i, в результате чего получилось сочетание qian. Подобная ситуация произошла с гласным звуком u, что и

послужило толчком к введению в фонетическую систему звука f, который до сих пор используется в соответствии с древними правилами.

Инициали zh, ch, sh пришли в современный язык из древности посредством преобразования переднеязычных инициалей d, t. В трактате китайского лингвиста Цянь Дасиня, посвященного различным сферам жизни, в том числе китайской фонологии (十驾斋养新录 / Shí jià zhāi yǎng xīn lù), отмечается следующее: 古无舌上音 / Gǔ wú shéshàngyīn «В древнекитайском отсутствуют верхнеязычные согласные» [Чжэн, 2016: 26]. В то же время для диалекта Хакка характерно некоторое отличие: при чтении они произносятся как z, c, s: 中间 / zung44gan44, 知道 / zi44tau52, 选择 / sien31ced5. Однако здесь тоже встречаются свои исключения. Таковыми являются следующие слова: 中心 / dung52sim44 «центр», литературное чтение морфемы 知道 / di44tau52 «знать» также отличается от просторечной, транскрипция словосочетания 择菜 / tog5coi52 «выбирать овощи» разнится с вышеуказанным чтением. Это связано с тем, что в разговорной речи до сих пор встречаются случаи употребления звуков d, t, что доказывает родство современной фонетической системы с древним произношением. Данные инициали могут употребляться с финалью i, а также с финалями, начинающимися на звук i: 七 / cít, 四 / sì.

Рассматривая подробнее вопрос инициалей диалекта, стоит отметить губнозубной звук f (轻唇音 / qīngchúnyīn, легкий губный согласный), который отсутствует в фонетической системе говорящих на Хакка. Это связано с тем, что фонетическое устройство древнекитайского языка не включало в себя использование данных звуков. Об этом свидетельствует высказывание Цянь Дасиня 古无轻唇音/ Gǔ wú qīngchúnyīn «В древнекитайском отсутствуют легкие губные согласные звуки» [Чжэн, 2016: 26]. Примерами выступают звуки, представленные в следующих словах: 非 / fēi, 敷 / fū, 奉 / fèng, 微 / wēi. Его альтернативой является губногубный звук f (重唇音 / zhòngchúnyīn, билабиальный звук): 飞机 / fi44gi44, 飞行 / fi44hang11, 飞翔 / fi44ciong11. В

Хакка его заменяет билабиальный звук b: 鸟粪 / diau44bun52, 斧头 / bu31teu11, 沸水 / bi52sui31, а также звук p: 辅导 / pu52tau52, 甫志高 / pu31zii52gau44, 肥料 / pi11liau52, 护身符 / fu52siin44pu11, 狗吠 / geu31poi52.

Другой фонетической особенностью диалекта являются финали, характерные для входящего тона: t (一 / yít, 七 / cít), m (三 / sām), k (六 / liùk), p (十 / sèp). Финаль m особенно отличается, так как является уникальной в Хакка, не характерна для путунхуа.

Анализируя особенности фонетической системы носителей диалекта Хакка, мы пришли к следующему выводу: большинство звуков претерпели изменения еще в период становления средневекового китайского языка и с тех времен закрепились в данном диалекте. Расхождение в количестве тонов с путунхуа обусловлено тем же фактором. Некоторые ученые утверждают, что в схожести с древнекитайским языком Хакка уступает только кантонскому диалекту [Чжан, Ян, 2022: 12]. Кроме того, в Хакка встречаются особенности, схожие с отличительными чертами других южных диалектов, что свидетельствует о современном влиянии на фонетику Хакка соседних диалектов, например диалекта города Фуцзянь. В то же время, такого же влияния общепринятого языка на диалект Хакка мы не отмечаем.

2.2.2. Лексические особенности диалекта Хакка

Фонетическая составляющая, будучи главным отличием диалекта от путунхуа, не исключает того факта, что различия присутствуют и в других областях языка. Рассматривая вопрос лексической составляющей и грамматического строя диалекта Хакка, мы отметили ряд характеристик: использование традиционной письменности, преобладание односложных иероглифов над многосложными, использование книжной лексики, активное употребление служебных частей речи.

Частым явлением в устной речи носителей диалекта Хакка является употребление устаревших иероглифов, давно вышедших из оборота в

путунхуа. Произношение в данном случае так же кардинально отличается от представленного в общепринятом языке. В качестве примера рассмотрим следующую пару слов: 镬 / vog5 (huò в путунхуа) – глубокая сковорода, которая обычно используется для приготовления пищи. Аналогом в путунхуа является иероглиф 锅 / guō. Таких примеров в Хакка встречается огромное множество: 樵 / ciau11 – хворост – в Хакка противопоставляется 柴/ chái в путунхуа: 倒樵 / dau31ciau11 – 砍柴 / kǎnchái; 烧樵 / sau44ciau11 – 烧柴 / shāochái. Кроме того, многие из этих иероглифов являются односложными. В качестве примеров приведем следующие иероглифы: 生气 / shēngqì, 嘴巴 / zuǐba, 尾巴 / wěiba, 脖子 / bózi, 翅膀 / chìbǎng в путунхуа равноценны 阿 / ad1, 喙 / fi52, 尾 / mi44, 颈 / giang31, 翼 / id5 в Хакка соответственно. Важно отметить тот факт, что одни и те же слова на Хакка могут различаться в зависимости от возраста человека, использующего то или иное слово в речи. Отмечается, что некоторые лексические единицы чаще употребляются молодым поколением, как например слово 电吹风 / diànchuīfēng «фен», пришедшее из путунхуа. Представители старого поколения привыкли употреблять слово 风筒 / fung44tung52 «фен». В качестве подобных примеров можно привести следующие слова: 口水 / heu31shui31 и 口澜渣 / keu31lon52za44 «слюна», 面帕 / mien53pa53 и 毛巾 / mau44giun44 «полотенце», 白味 / pag5mi52 и 酱油 / ziong53-iu11 «соевый соус». В то же время существуют слова, которые представители разных поколений с одинаковой частотностью используют при общении: 当昼头 / dong44zhiu53teu11 «полдень», 丈人老 / chong44njin11lou31 «тесть», 拖格 / to44gag1 «выдвижной ящик» [Тан, 2019: 16]. Таким образом, употребление некоторых лексических единиц зависит от возраста носителя диалекта: молодое поколение частично встраивает в свою речь слова, пришедшие из путунхуа, в то время как пожилые люди остаются верны традициям.

Классификации по частям предшествует разделение слов на знаменательные и служебные. Обращаясь к разделению по частям речи, для диалекта Хакка, как и любого другого, присущи оба класса. В данной работе мы рассматриваем некоторые из них в обеих категориях. Обращаясь к самостоятельным частям речи, следует обратить внимание на существительные, так как их количество превалирует над остальными частями речи.

Таблица 2. Существительные диалекта Хакка

№	Путунхуа	Хакка	Комментарий
1	苹果	苹果	Написание полностью совпадает, разница в чтении практически незаметна
	[píngguǒ] яблоко 这个苹果怎么卖?	[pin11guo31] яблоко 吩个苹果样欸卖?	
2	事	事	Написание полностью совпадает, разница в чтении
	[shì] дело, случай 你有什么事找我?	[sii52] дело, случай 汝有脉个事寻涯?	
3	箱子	箱欸	Замена словообразовательного суффикса, разница в чтении
	[xiāngzi] чемодан 小李，你把那个箱子搬过来。	[siong44oi31] чемодан 小李，汝把个个箱欸搬过来。	
4	傻瓜	傻货	Замена словообразовательного суффикса, разница в чтении
	[shǎguā] глупец 这个傻瓜，被人骗了几十块都不知道!	[so31fo52] глупец 吩只傻货，分人撮撇几十元都唔知!	
5	块	元	Полная замена иероглифа, разница в чтении
	[kuài] юань 这个傻瓜，被人骗了几十块都不知道!	[ian11] юань 吩只傻货，分人撮撇几十元都唔知!	
6	昨天	秋哺日	Полная замена иероглифов, разница в чтении
	[zuótiān] вчера 你昨天为什么不来上班?	[ciu44bu44ngid1] вчера 汝秋哺日做脉个唔来上班?	

Примеры, представленные в таблице выше, характеризуют три типа существительных, которые могут встречаться в Хакка: они либо полностью совпадают с путунхуа, либо заменяется часть слова. В некоторых случаях возможна полная замена иероглифов, а где-то и присоединение дополнительного иероглифа. Поэтому слова, наиболее схожие в произношении и написании с общепринятым языком, легче воспринимаются носителями путунхуа. Иероглифы, кардинально отличающиеся от привычных не носителям Хакка, являются особенностью диалекта.

Что касается глаголов, они представляют больший интерес, так как среди них чаще встречаются слова, нехарактерные для общепринятого языка.

Таблица 3. Глаголы диалекта Хакка

№	Путунхуа	Хакка	Комментарий
1	帮	同	Иероглиф 同 помимо функции союза (和) и выражения схожести (相同) используется в значении глагола
	[bāng] помогать 书包放在这里，我帮你看着。	[tung11] помогать 书包放啊吟欸，涯同汝看等。	
2	找	寻	Глагол 寻 является односложным, равнозначен глаголу 找, однако в путунхуа чаще они используется в связке 寻找
	[zhǎo] спрашивать 妹妹在那儿，你自己过去找她。	[cim11] спрашивать 老妹在个欸，汝自家过去寻佢。	
3	要	爱	Значение глагола 爱 совпадает с привычным в путунхуа «любить», однако часто встречается и со значением глагола 要
	[yào] хотеть, желать 你要哪个?	[oi52] хотеть, желать 汝爱奈只?	
4	站	企	

	[zhàn] остановиться 你站在车站哪边？	[ki44] остановиться 汝企啊车站奈片？	Значение иероглифа 企 отличается в Хакка, используется как однослог в значении глагола 站
5	吃	食	Глаголу 吃 в путунхуа соответствует глагол 食
	[chī] есть 你吃了饭吗？	[siid5] есть 汝食欸饭么？	
6	听	听	Иероглиф не изменяется, разница в чтении
	[tīng] слушать 他听不懂客家话。	[tang52] слушать 佢听唔识客家话。	

В пяти из шести случае глаголы полностью отличаются как в написании, так и в чтении: используются иероглифы, схожие по смыслу, разница в чтении так же обусловлена особенностями фонетической системы. Единственным совпадением является глагол 听 / tīng «слушать». Следовательно, глаголы диалекта Хакка представляют особую сложность для понимания, так как корнями уходят в древность.

Стоит обратить внимание на прилагательные, которые представлены в таблице ниже.

Таблица 4. Прилагательные диалекта Хакка

№	Путунхуа	Хакка	Комментарий
1	不错	唔错	Иероглиф 错 остается без изменений, различие – отрицательная частица
	[bùcuò] неплохой 我们关系不错。	[m11co52] неплохой 涯等人关系唔错。	
2	快	快	Полностью совпадает написание и чтение
	[kuài] быстрый 这边排队的人少，快点过来!	[kuai52] спрашивать 吩片排队个人少，快滴过来!	
3	忙	唔得闲	

			Сочетание 唔得闲, совпадающее с кантонским, полностью отличается от иероглифа 忙
	[máng] занятой 这几天我都忙，那样，你下周一过来。	[m11ded1han11] занятой 吟几日涯都唔得闲个咬样，汝下周一过来。	

Исходя из представленных примеров, важно отметить схожесть прилагательных путунхуа и Хакка. В то же время отмечаем влияние кантонского диалекта на Хакка в примере с лексемой 忙 / máng «занятой».

В работе мы также рассматриваем местоимения, так как процент их употребления в речи соотносится с частотой использования существительных, следовательно они заслуживают внимания. После анализа примеров, приведенных в видеороликах, местоимения были систематизированы в зависимости от разряда.

Таблица 5. Личные местоимения диалекта Хакка

№	Единственное число			Множественное число		
Лицо	Первое	Второе	Третье	Первое	Второе	Третье
Хакка	涯 [ngai2]	汝 [n2]	佢 [gi2]	涯等人 [ngai2 den1 ngin2]	汝等人 [n2 den1 ngin2]	佢等人 [gi2 den1 ngin2]
Путунхуа	我 [wǒ]	你 [nǐ]	她 / 他 / 它 [tā]	我们 [wǒmen]	你们 [nǐmen]	他们 / 她们 / 它们 [tāmen]
Перевод	Я	Ты	Он / Она / Оно	Мы	Вы	Они

Местоимения, представленные в таблице, относятся к разряду личных местоимений. Отдельно мы выделяем местоимения 大齐家 / tai4 ce2 ge1 «все», которому в путунхуа соответствует местоимение 大家 и местоимение 自家 / cii52ga44 «сам» , аналогичное местоимению 自己 / zìjǐ «сам». Основное различие первого местоимения заключается в иероглифе 齐 / ce2, который с

Хакка переводится как «все». Следовательно, использование данного иероглифа в центре слова вполне уместно. Кроме того, данные примеры демонстрируют, насколько продуктивен суффикс « 家 » в создании местоимений. Как упоминалось ранее, в диалекте Хакка насчитывается большое количество заимствований из языка периода эпохи Тан. Учитывая источник возникновения данных местоимений, в частности местоимения второго лица 汝 / nji11 «ты», это утверждение является верным. Что еще стоит отметить, местоимение 涯 / «я» первого лица и местоимение 汝 / nji11 «ты» второго лица прошли через некоторые трансформации, прежде чем войти в повседневное пользование в таком виде: ключ 氵 / shuǐ «вода» вытеснил изначально использовавшийся ключ 亻 / rén «человек». Другая отличительная особенность местоимений – это суффикс множественного числа 等人. Он состоит из двух иероглифов, которые в путунхуа трактуются как 等 / děng «ждать» и 人 / rén «человек». В путунхуа такое сочетание имеет значение «и другие» при перечислении лиц, что объясняет логику использования данного двуслога в значении суффикса множественного числа. Отличительной его чертой является использование лишь с одушевленными лицами.

Указательные местоимения в Хакка тоже обладают рядом особенностей.

Таблица 6. Указательные местоимения диалекта Хакка

№	Путунхуа	Хакка	Комментарий
1	这个	吟个	Иероглиф 这 заменяется иероглифом 吟
	[zhège] этот, это	[ge3ge4] этот, это	
2	那个	个个	Иероглиф 那 заменяется иероглифом 个
	[nàge] тот, то	[ge4ge4] тот, то	
3	这些	吟兜	Иероглиф 兜 равноценен по значению иероглифу 些, в некоторых случаях может выступать как классификатор для небольшого количества (兜人)
	[zhèxiē] эти	[ge3deu1] эти	

4	那些	个兜	Иероглиф 兜 равноценен по значению иероглифу 些, в некоторых случаях может выступать как классификатор для небольшого количества (兜人)
	[nàxiē] те	[ge4deu1] те	
5	这里	吟欸	Иероглиф 里 заменяется иероглифом 欸
	[zhèlǐ] тут	[ge3e2] тут	
6	那里、那儿	个欸	Иероглиф 里 заменяется иероглифом 欸
	[nàli, nàr] там	[ge4e2] там	
7	这只	吟只	Преимущественно относится к людям
	[zhèzhǐ] этот	[ge3zag5] этот	
8	那只	个只	Преимущественно относится к людям
	[nàzhǐ] тот	[ge4zag5] тот	
9	这会儿	吟下	Краткость выражается иероглифом 下
	[zhèhuìr] теперь	[ge3ha4] теперь	
10	那会儿	个下	Краткость выражается иероглифом 下
	[nàhuìr] тогда	[ge4ha4] тогда	
11	这边	吟片	Иероглиф 边 заменяется иероглифом 片
	[zhèbiān] на этой стороне	[ge3pien3] на этой стороне	
12	那边	个片	Иероглиф 边 заменяется иероглифом 片
	[nàbiān] на той стороне	[ge4pien3] на той стороне	
13	这么	咬欸	Полная замена иероглифов
	[zhème] настолько	[an3e2] настолько	
14	那么	个咬欸	Полная замена иероглифов, добавление иероглифа 个
	[nàme] в такой степени	[ge4an3e2] в такой степени	
15	这样	咬样	Иероглиф 这 заменяется иероглифом 咬
	[zhèyàng] такой; так	[an3ngiong4] такой; так	

16	那样	个咬样	Иероглиф 那 заменяется иероглифом 咬, добавление иероглифа 个
	[nàyàng] такой; так	[ge4an3ngiong4] такой; так	

Приведенные в таблице примеры указательных местоимений значительно отличаются от общепринятых в путунхуа по ряду особенностей: чтение, написание, количество иероглифов в рамках одной лексемы. Характерной чертой является замена иероглифов из путунхуа на привычные в Хакка. В некоторые случаях такая замена основана на схожести звучания, как в случае с 边 / biān и 片 / pien3. Также есть примеры, где изменения обоснованы смысловой составляющей иероглифа: 会儿 / huìr и 下 / ha4 – обе лексемы обозначают короткий период времени. В некоторых случаях для разграничения двух местоимений в Хакка используется дополнительный иероглиф, как в случае с местоимением «такой»: 这样 / zhèyàng и 咬样 / an3ngiong4, а также 那样 / nàyàng и 个咬样 / ge4an3ngiong4 соответственно. Следовательно, отмечается большое количество различий при сравнении путунхуа и Хакка.

Вопросительные местоимения – еще один многочисленный раздел местоимений, присущий диалекту Хакка.

Таблица 7. Вопросительные местоимения диалекта Хакка

№	Путунхуа	Хакка	Комментарий
1	哪些	奈兜	Иероглиф 哪 заменяется иероглифом 奈, иероглиф 兜 равноценен по значению иероглифу 些, в некоторых случаях может выступать как классификатор для небольшого количества (兜人)
	[nǎxiē] какие?	[nai4deu1] какие?	
2	哪里	奈欸	Иероглиф 哪 заменяется иероглифом 奈
	[nǎlǐ] где, куда?	[nai4e2] где, куда?	

3	哪个	奈只	Полная замена иероглифов
	[nǎge] который?	[nai4zag5] который?	
4	哪边	奈片	Полная замена иероглифов
	[nǎbiān] где?	[nai4pien3] где?	
5	怎么	样欸	Синонимом выступает слово 样般 / iong52ban44 «таким образом»
	[zěnme] как? каким образом?	[ngiong4e2] как? каким образом?	
6	什么	脉个	Полная замена иероглифов
	[shénme] что?	[mag5ge4] что?	
7	为什么，干吗	做脉个	Полная замена иероглифов
	[wèishénme, gànmá] почему? зачем?	[zo4mag5ge4] почему? зачем?	
8	谁	满人	К вопросительному иероглифу 满 добавляется лексема 人
	[shuí, shéi] кто?	[man1ngin2] кто?	
9	多少	多少	Написание совпадает, различие в чтении
	[duōshǎo] сколько?	[do1sau3] сколько?	
10	几	几	Написание совпадает, различие в чтении
	[jī] сколько?	[gi3] сколько?	

Среди данных примеров есть такие, которые кардинально отличаются как написанием, так и чтением, однако иероглиф 几 / jī «сколько» одинаково используется в обеих языковых системах. Снова отмечается замена некоторых иероглифов на привычные в Хакка, соответственно, наблюдается и разница в чтении. Особый интерес представляет вопросительное местоимение «кто», которое в Хакка восходит еще к древности и на письме выражается как 满人 / man1ngin2. Особенным является также местоимение 做脉个 / zo4mag5ge4, которое выполняет функцию сразу двух вопросительных слов «почему» и «зачем», которые в путунхуа принято разделять.

Таким образом, мы отмечаем значительные расхождения чтении и написании местоимений Хакка. Многие местоимения в составе содержат

иероглифы, которые не используются в путунхуа, однако мы не выявили значительных сходств с системой местоимений древнекитайского языка. Большинство местоимений являются отличительной чертой диалекта.

Обобщая рассмотренные части речи, мы пришли к выводу, что по ряду характеристик наиболее отличаются глаголы и местоимения: им присуще использование иероглифов, отличных от путунхуа, в отдельных случаях отмечается расхождение в количестве иероглифов у некоторых слов, значительна разница в чтении. В вопросе остальных частей речи, написание многих иероглифов совпадает, однако особенности фонетической системы Хакка осложняют понимание отдельных слов на слух.

2.2.3. Грамматический строй диалекта Хакка

Грамматика китайского языка характеризуется фиксированным порядком членов предложения. В сравнении с русским языком она довольно простая, так как в ней отсутствуют наклонения, спряжения, слова не изменяются в зависимости от времени. В то же время, в ней есть свои особенности, которые мы рассматриваем в следующих примерах. Анализируя предложения, приведенные в видеороликах, мы сопоставили наиболее распространенные конструкции в путунхуа и в Хакка.

Первым и самым базовым, присущим обоим языковым системам, является глагол-связка 是 / «быть, являться, находиться».

Таблица 8. Примеры использования глагола-связки 系

№	Пример на путунхуа	Пример на Хакка
1	我是客家人。	涯系客家人。
2	他们是一伙的。	佢等人系一伙个。
3	那些人都是什么人?	个兜人都系脉个人?
4	老板，这些水果都是进口的吗?	老板，吩兜水果都系进口个么?

В Хакка данный глагол заменяется на иероглиф 系 / he52, выполняющий все те же функции, что и 是 / shì в путунхуа, а именно используется в различных грамматических конструкциях, обозначает как одушевленные, так и неодушевленные предметы.

Следующим в списке рассмотренных является предлог 在 / zài «в», выполняющий функцию глагола «быть в».

Таблица 9. Примеры использования глагола 在 / 啊

№	Пример на путунхуа	Пример на Хакка
1	书包放在这里，我帮你看着。	书包放啊吟欸，涯同汝看等。
2	妹妹在那儿，你自己过去找她。	老妹在个欸，汝自家过去寻佢。
3	你在哪里?	汝在奈欸?
4	你站在车站哪边?	汝企啊车站奈片?

Позиция этого предлога в диалекте Хакка неоднозначна. Чаще он используется без изменений, сохраняя за собой написание и значение, однако зафиксированы случаи замены на частицу 啊 / a44, выполняющую те же функции.

Отдельного внимания заслуживает притяжательная частица 的, которая используется для указания на принадлежность.

Таблица 10. Примеры использования притяжательной частицы 个

№	Пример на путунхуа	Пример на Хакка
1	哪些是你的行李?	奈兜系汝个行李?
2	这边排队的人少，快点过来!	吟片排队个人少，快滴过来!

В Хакка она заменяется иероглифом 个 / ge52, однако функции частицы совпадают с вариантом, представленным в путунхуа.

Конструкция 是。。。的, строящаяся из двух предыдущих элементов, соответственно заменяется на 系。。。个.

Таблица 11. Примеры использования конструкции 系。。。个

№	Пример на путунхуа	Пример на Хакка
1	他们是一伙的。	佢等人系一伙个。
2	老板，这些水果都是进口的吗?	老板，吩兜水果都系进口个么?
3	这边排队的人少，快点过来!	吩片排队个人少，快滴过来!

Данная конструкция используется для акцентирования внимания на определенной информации. Она строится по тем же правилам, что и общепринятом языке.

Еще одна грамматическая конструкция, представленная в отсмотренных материалах – это пассивный залог с частицей 被 / bèi – служебное слово, обозначающее воздействие.

Таблица 12. Пример использования конструкции пассивного залога

№	Пример на путунхуа	Пример на Хакка
1	这个傻瓜，被人骗了几十块都不知道!	吩只傻货，分人撮撇几十元都唔知!

Сопоставляя пример на путунхуа с диалектом Хакка, мы выяснили, что порядок слов и смысл предложения остается фиксированным, однако происходит замена иероглифа на 分 / bun44.

Образование вопросительных предложений в Хакка возможно не только благодаря вопросительным местоимениям. Как и в путунхуа, для этого используется вопросительная частица.

Таблица 13. Примеры использования вопросительной частицы 么

№	Пример на путунхуа	Пример на Хакка
1	你吃了饭吗?	汝食欸饭么?
2	老板，这些水果都是进口的吗?	老板，吩兜水果都系进口个么?
3	你们去过北京吗?	汝等人去过北京么?

Однако мы отметили различие в выборе иероглифа: привычная частица 吗 / ma – конечная частица вопросительных предложений, заменяется на

частицу 么 / ma44, входящую в состав вопросительного местоимения 什么 / «что».

В процессе анализа грамматических особенностей диалекта Хакка, значительных отличий мы не выявили: порядок слов в грамматических конструкциях, использование глаголов-связок, употребление пассивного залога и другие структуры сохраняются в Хакка. В то же время, главной отличительной чертой является замена иероглифов путунхуа на иероглифы диалекта Хакка, в связи с чем меняется и чтение. Нельзя сказать, что в данной ситуации легко понять смысл предложений на Хакка, однако обладая знаниями о базовых конструкциях в путунхуа и при условии владения базовым набором лексических единиц диалекта, понимание некоторых грамматических структур значительно упрощается.

ВЫВОДЫ ПО ГЛАВЕ 2

В практической части выпускной квалификационной работы был описан процесс создания корпуса на материале диалекта Хакка, представленного в текстах видеороликов, а также был проведен анализ данных, структурированных в формате корпуса. Кроме того, трудности, с которыми нам пришлось столкнуться во время создания корпуса, также описаны в работе. К ним относятся особенность деления на части речи иероглифов китайского языка, а также множество вариантов произнесения одного и того же слова. Способы решения данных проблем, основывающиеся на нашей точке зрения и подкрепленные теоретическими данными, присутствуют в тексте работы.

За основу плана создания корпуса диалекта Хакка был взят алгоритм В.П. Захарова, состоящий из девяти этапов. В нашем корпусе некоторые этапы были объединены, в результате чего процесс немного отличался. На начальном этапе нами были отобраны видеоматериалы, после чего мы перевели их в текстовый формат посредством программы ELAN. В этой же программе мы провели сегментацию и разметку текста, в частности, определили части речи каждого слова и отдельно взятого иероглифа, а также указали фонетическую транскрипцию. Затем, готовые файлы в формате TXT были загружены в специальную программу для создания корпуса Sketch Engine. Нами было принято решение разделить тексты на путунхуа и диалекте Хакка по отдельным файлам и создать параллельный корпус. Это позволило нам в дальнейшем проводить сравнительный анализ тех или иных иероглифов. В конечном итоге, у нас получился параллельный корпус на официальном языке и диалекте Хакка, который в дальнейшем был проанализирован на наличие сходств и различий двух языковых систем.

В результате анализа корпуса был выявлен ряд особенностей, присущих диалекту Хакка. Первой и наиболее очевидной отличительной чертой является фонетическая система диалекта, которая насчитывает шесть тонов, включает в себя сочетания слогов, отсутствующих в путунхуа, а также характеризуется

употреблением одиночных инициалей, таких как звук m. Что касается лексической составляющей, она содержит определенный процент слов, отличающихся от путунхуа в плане иероглифики, однако мы не можем утверждать, что данное явление обусловлено исключительно связями с древнекитайским и среднекитайским, мы также отмечаем влияние южных диалектов на Хакка, кантонского в частности. Грамматический строй диалекта не представляет особой трудности для понимания, так как в грамматике Хакка соблюдается большинство правил, присущих путунхуа. Единственным различием, которое мы выделили, является расхождение в иероглифике, однако это больше вопрос лексического наполнения языка. Таким образом, мы делаем вывод, что диалект Хакка в значительной степени сложен для слухового и визуального восприятия, однако обладая некоторым количеством знаний в обеих областях, нетрудно будет освоить грамматический строй данного диалекта, так как он имеет много схожих моментов с путунхуа.

ЗАКЛЮЧЕНИЕ

Корпусная лингвистика берет свои истоки с давних времен, когда у людей появилось понимание о необходимости структурировать готовые тексты в целях удобства работы с ними. Однако настоящий расцвет данного направления связан с изобретением компьютера и его дальнейшим развитием.

Первые цифровые корпусы открыли научному сообществу новый способ анализа больших объемов данных и дали возможность ускорить данный процесс. Несмотря на то, что первые корпусы появились в 60-ые гг. XX в. в США, их развитие по всему миру не заставило себя долго ждать. Тем не менее китайские ученые долгое время были лишены возможности приобщиться к новой тенденции ввиду особенностей китайского языка. Лишь с течением времени, после добавления иероглифики в компьютерные системы письменности, китайский язык получил возможность стать основой для корпуса.

Наиболее известными и масштабными корпусами китайского языка являются Chinese Corpus online, корпус CCL при Пекинском университете, а также Корпусный центр при Пекинском университете языка и культуры (BLCU Corpus Center). Несмотря на то, что вышеперечисленные базы данных являются главенствующими среди корпусов китайского языка, большинство текстов в них являются новостными, историческими или же деловыми. Некоторые включают в себя литературные произведения, однако среди них нет корпусов, посвященных диалектам.

В китайском научном пространстве существует достаточно работ, посвященных диалектным особенностям, однако нами не было обнаружено подобных корпусов. В связи с данной ситуацией, нами было принято решение заполнить существующую лакуну и создать собственный корпус диалекта Хакка, тем самым объединив некоторый объем работ, посвященных диалекту. В китайской поисковой системе Baidu мы смогли найти онлайн-словарь

диалекта Хакка Syndict.com, который был взят в работу как источник данных о диалекте.

Алгоритм конструирования корпуса, разработанный В.П. Захаровым, лег в основу нашего исследования и стал главным ориентиром. Несмотря на детальное описание каждого этапа, мы столкнулись с рядом трудностей, которые были обусловлены как особенностями китайского языка, так и необходимостью осваивать новую сферу деятельности, изучать специализированные программы.

На этапе подбора материалов глобальных проблем не возникло, так как мы заранее определились с форматом текста и прописали критерии отбора. Нами были отобраны 30 видеороликов с китайской интернет-платформы Bilibili, посвященные диалекту Хакка. Нам было важно, чтобы в видеоролике присутствовал перевод на путунхуа, были подготовлены субтитры, имелось звуковое сопровождение говорящего. Последним критерием для нас стало происхождение автора видео. Таким образом, мы отсеивали видео, созданные иностранцами, для того, чтобы максимально приблизиться к настоящей диалектной речи.

Следующим шагом после отбора материала следовала его перекодировка в текстовый формат. Для данной цели мы использовали автоматизированную систему аннотирования ELAN. Помимо перевода в письменный формат, приложение предоставило функцию сегментации текста и лингвистической разметки, что позволило осуществить сразу три этапа в одной программе.

После того как материал видеороликов был приведен в текстовый формат и отредактирован в соответствии с требованиями, следующим шагом для нас стала загрузка готовых текстов в другую специализированную программу для конструирования корпуса. На данном этапе возникла проблема с форматом текстов, так как подготовленные нами материалы были оформлены в кодировке TXT, а программа требовала другие виды кодировки. В противном случае, тексты распознавались как необработанные. Несмотря на

эту трудность, тексты удалось загрузить и они прошли процесс обработки, в результате которого мы получили параллельный корпус на путунхуа и диалекте Хакка. Таким образом, у нас получилось создать небольшой корпус, который охватывает наиболее часто употребляемые лексические единицы диалекта с указанием их частей речи и фонетической составляющей иероглифов. Удобство данного корпуса в том, что он получился параллельным, соответственно, при поиске какого-либо слова можно осуществлять поиск по диалекту Хакка и получать его аналог на путунхуа.

После того как корпус был создан и отредактирован, посредством его мы проанализировали, насколько велик процент сходства между диалектом Хакка и путунхуа, а также выделили список некоторых особенностей диалекта. В первую очередь, стоит отметить фонетическую систему диалекта, которая в значительной степени отличается от официального языка: сочетание инициалей и финалей, нехарактерное для путунхуа; два тона, отсутствующих в официальном языке; удвоенные гласные в слоге. Данные отличия обуславливаются рядом факторов, например, влияние южных диалектов или тесная связь Хакка с китайским языком предыдущих эпох. Кроме фонетической составляющей, мы также выделили отличия в лексической системе диалекта: в некоторых словах присутствуют традиционные иероглифы, встречается изменение ключей в иероглифах. Важно, что последняя особенность не всегда меняет смысл слова. Кроме того, одно и то же слово не всегда совпадает по количеству знаков с тем же словом в путунхуа, часто используются однослоги, что является результатом влияния древнекитайского языка. Грамматический строй диалекта представляет наименьший интерес, так как большинство правил путунхуа остаются неизменными и для Хакка. Таким образом, ряд выделенных отличий демонстрирует особенность диалекта Хакка и позволяет проследить его связь с соседними диалектами, а также с языком времен эпохи Тан.

Сконструированный нами корпус на данный момент находится на ранней стадии развития, так как объем представленных в нем данных

сравнительно невелик. Тем не менее уже сейчас его можно использовать для проведения небольших исследований и изучения устройства диалекта Хакка. Внесение новых данных, обновление разметки и добавление сведений об источниках информации могут стать дальнейшими перспективами развития данного корпуса.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Алексахин. А.Н. Диалект Хакка, М.: Наука, 1987 г. 88 с.
2. Астрахан Е.Б., Завьялова О.И., Софронов М.В. Диалекты и национальный язык в Китае. М., Наука, 1985. 366 с.
3. Баглаева А.С. Проблема сохранения диалектов в Китае как результат современной политики // Иностранные языки в современном мире: сб. ст. по матер. междунар. науч.-практ. студ. конф. Ростов-на-Дону: Издательско-полиграфический комплекс Рост. гос. экон. ун-та. 2021. С. 237–240.
4. Баркович А.А., Ван Ц. Лингвистические корпусы китайского языка: функциональный аспект // Вестник МГЛУ. Сер.: Филология. 2015. Т.5. Вып. 78. С. 105–113.
5. Баров С.А., Егорова М.А. Кантонский диалект в современном Китае: проблема сохранения // Вестник Российского университета дружбы народов. Сер.: Теория языка. Семиотика. Семантика, 2019. Т. 10. № 1. С. 152–166.
6. Большой китайско-русский словарь [Электронный ресурс]. 2007. URL: https://bkrs.info/ (дата обращения: 22.05.2024).
7. Введение в науку о языке / А.Е. Кибрик и др.; под ред. О.В. Федоровой и С.Г. Татевосова. М.: Буки Веди. 2019. 672 с.
8. Волоснова Ю.А. Корпусная лингвистика: проблемы и перспективы // Лесной вестник. М.: Изд-во МГУЛ, 2006. 7, С. 43–49.
9. Гутин И.Ю. Языковая ситуация в специальном административном районе Гонконг КНР и политика властей в сфере языка // Международный научно-исследовательский журнал. 2018, № 2. С. 79–83.
10. Довнар П.Ю., Воронцов А.В. Лингвистический процессор китайского языка. Особенности разработки // Международный конгресс по информатике: информационные системы и технологии: материалы международного научного конгресса. 2011 г. Минск: БГУ, 2011. С. 202–207.

11. Жеребило Т.В. Словарь лингвистических терминов. Изд. 5-е, испр. и доп. Назрань: Издательство Пилигрим, 2010. 486 с.

12. Завьялова О.И. Путунхуа и диалекты: Новые реалии китайского мира // Журнал РАН. Сер.: Культура. 2012. Вып. 6. С. 130–138.

13. Завьялова О.И. Большой мир китайского языка. 2-е изд. М.: Восточная книга, 2014. 317 с.

14. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник. 3-е изд., перераб. СПб.: Изд-во С.-Петерб. ун-та, 2020. 234 с.

15. Касаткин Л.Л. Русская диалектология: Учебник для студ. филол. фак. высш. учеб. заведений, 2005. 280 с.

16. Киссилев О.В., Яценко А.А. Разработка лингвистического корпуса на основе письменных работ студентов, изучающих русский язык как иностранный // Профессиональное лингвообразование: материалы Пятой международной научно-практической конференции. Нижний Новгород. 2011. С. 96–108.

17. Козлова Н.В. Лингвистические корпуса: определение основных понятий и типология // Вестник НГУ. Сер.: Лингвистика и межкультурная коммуникация. 2013. Т. 11 Вып. 1. С. 79–88.

18. Колпачкова Е.Н. Корпусы китайского языка: современное состояние и основные проблемы // Труды международной конференции «Корпусная лингвистика – 2015». СПБ: Издательство Санкт-Петербургского университета, 2015. С. 278–286.

19. Лаврентьев А.М. Корпусная лингвистика: идеология, методы, технологии // Сибирский филологический журнал. 2004. Т. 3. Вып. 4. С. 121–134.

20. Лу И. Принципы создания корпусов китайского языка // Известия РГПУ им. А. И. Герцена. 2016. С. 22–29.

21. Лу И. Методы создания китайского корпуса текстов лингводидактики // Вестник Нижегородского университета им. Н.И. Лобачевского. Сер.: Филология. 2018. С. 195–200.

22. Национальный корпус русского языка. [Электронный ресурс]. 2003. URL: http://www.ruscorpora.ru/ (дата обращения: 13.04.2024).

23. Потапов В.В., Матвеева А.Е. Развитие корпусной лингвистики как науки и ее влияние на общую теорию языка. (Обзор) // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 6: Языкознание. Реферативный журнал. Москва: ИНИОН. 2022. 4. С. 54–61.

24. Рыбникар А.А. Исследование диалектов Китая российскими учёными // В мире науки и искусства: вопросы филологии, искусствоведения и культурологии: сб. ст. по матер. XXVIII междунар. науч.-практ. конф. № 9 (28). Новосибирск: СибАК, 2013. С. 1–7.

25. Самарцева Е.А. Лексические особенности диалекта хакка // Наука в мегаполисе. Раздел: Литературоведение и языкознание, 2020. Выпуск № 10 (26). С. 1–3.

26. Скорлыгина Н.В. Китайцы-хакка в Южном Китае и Юго-Восточной Азии: XVIII-XIX века: автореферат дис ... канд. ист. наук: 07.00.03. Москва, 2005. 38 с.

27. Солнышкина М.И., Гатиятуллина Г.М. История развития корпусной лингвистики (на примере англоязычных корпусов) // Вестник ТГУ. Сер.: Филология. 2020. С. 132–160.

28. Софронов М.В. Китайский язык и китайская письменность. Курс лекций. М.: АСТ: Восток-Запад, 2007. С. 241–243.

29. Юэ Ф. Специфика корпусных исследований в современном китайском языкознании // Вестник МГЛУ. Сер.: Гуманитарные науки. 2020. Т. 832. Вып. 3. С. 159–172.

30. Яхонтов С.Е. Классификация диалектов китайского языка // Проблемы китайского и общего языкознания. К 90-летию С.Е. Яхонтова / отв. ред. Е.Н. Колпачкова. СПб.: Изд-во «Студия «НП-Принт», 2016. С. 117–125.

31. Яхонтов С.Е. Письменный и разговорный китайский язык в VII–XIII вв. н.э. // Проблемы китайского и общего языкознания. К 90-летию С.Е. Яхонтова / отв. ред. Е.Н. Колпачкова. СПб.: Изд-во «Студия «НП-Принт», 2016. С. 182–195.

32. Baker P., Hardie A., McEnery T. A glossary of corpus linguistics. Edinburgh: Edinburgh University Press Ltd, 2006. 192 p.

33. Beijing Language and Culture University Corpus Center [Электронный ресурс]. 2004. URL: http://bcc.blcu.edu.cn/ (дата обращения: 15.03.2024).

34. Finegan E. LANGUAGE: its structure and use. N.Y.: Harcourt Brace College Publishers, 2004. 613 p.

35. Glazkova A.V. Building a text corpus for automatic biographical facts extraction from Russian texts // International Journal of Open Information Technologies. 2019. 7 (1), P. 97–103.

36. Glosbe [Электронный ресурс]. 2023. URL: https://glosbe.com/ (дата обращения: 21.05.2024).

37. Li Ch., Wang X. Building large Chinese corpus for spoken dialogue research in specific domains. International Joint Conference on Natural Language Processing. 2017. P. 320–324.

38. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge University Press. 2012. 312 p.

39. Syndict 薪 典 [Электронный ресурс]. 2022. URL: https://www.syndict.com/index.htm (дата обращения: 21.05.2024).

40. 戴斯谨。梅县客家方言人称代词的句法及篇章衔接功能研究。广西师范大学，2012。[Дай Сидзин. Исследование синтаксиса и роль личных местоимений в тексте диалекта Хакка уезда Мэй].

41. 邓海龙。客家方言语料库建设与英语学习中的母语迁移研究。江西：赣南师范学院，2011。[Дэн Хайлун. Исследование создания корпуса диалекта Хакка и передачи родного языка при изучении английского языка].

42. 邓海龙。赣南客家方言语音语料库建设的必要性分析。江西：赣南师范学院，2016。[Дэн Хайлун. Анализ необходимости создания речевого корпуса диалекта Хакка провинции Цзянси].

43. 梅州市梅县区人民政府网站：客家方言。 [Официальный сайт народного правительства округа Мэйчжоу уезда Мэй: диалект Хакка]. [Электронный ресурс]. 2012. URL: http://www.gdmx.gov.cn/zjmx/kjwh/content/post_65161.html (дата обращения: 04.12.2023).

44. 谢留文，黄雪贞。客家方言的分区（稿）。北京：中国社会科学院语言研究所，2007。[Се Лювэнь, Хуан Сюэчжэнь. Разделение диалектов хакка (рукопись)].

45. 谭科宏。梅县客家话口语中常用名词的变异现象。广东：嘉应学院，2019。[Тань Кэхун. Явление вариативности общеупотребительных существительных в диалекте Хакка уезда Мэй].

46. 郝鹏飞。广西贺州市桂岭镇客家话研究。广西师范大学，2011。[Хао Пэнфэй. Исследование диалекта хакка в городе Гуйлин, город Хэчжоу, Гуанси].

47. 钱曾怡。推广普通话和保护方言。山东大学，2006。[Цянь Цзэньи. Распространение общенародного китайского языка и защита диалектов].

48. 张雪，杨梓蓉。普通话和梅县客家话的词汇比较研究。澳门科技大学国际学院，2022。[Чжан Сюэ, Ян Цзыжун. Сравнительное исследование лексики путунхуа и Хакка уезда Мэй].

49. 张宇。大学开设方言课：传承文化未尝不可。红网，2019。[Чжан Юй. Университеты ввели уроки диалектов: осуществление передачи культурного наследия].

50. 钟舟海。浅议客家方言俗语的研究与保护。江西：江西理工大学，2022。[Чжун Чжоухай. Краткая дискуссия об исследовании и сохранении пословиц диалекта Хакка].

51. 郑秋晨。梅县客家话对古汉语语音的传承。广州：广州涉外经济职业技术学院，2016。[Чжэн Цючэнь. Наследование древнекитайского произношения диалектом Хакка уезда Мэй].

52. 杨悦。“方言梗”如何留住方言文化。南方日报，2021。[Ян Юэ: Как шутки на тему различий в диалектах сохраняют культуру].

53. 闫淑惠。近四十年来客家方言研究的：历史经验与当代反思。江西：赣南师范大学学报，2020。[Янь Шухуэй. Исследование диалекта Хакка за последние 40 лет: исторический опыт и современное отражение].

Министерство науки и высшего образования РФ
Федеральное государственное автономное
образовательное учреждение высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт филологии и языковой коммуникации
Кафедра восточных языков

УТВЕРЖДАЮ
Заведующий кафедрой
_____ И.Г. Нагибина
«__13__» ____июня____ 2024 г.

БАКАЛАВРСКАЯ РАБОТА

45.03.02 Лингвистика

КОНСТРУИРОВАНИЕ ТЕКСТОВОГО КОРПУСА ДИАЛЕКТА ХАККА

Научный руководитель _____ ст. преп. каф. ВЯ Ю. Чжан

Выпускник _____ Е.А. Скомороха

Нормоконтролер _____ И.А. Рабцевич

Красноярск 2024

РЕФЕРАТ

Тема бакалаврской работы – «Конструирование текстового корпуса диалекта Хакка». Выпускная квалификационная работа представлена в объеме 67 страниц, включает в себя 13 таблиц, а также список использованной литературы, состоящий из 53 источников, 22 из которых на иностранных языках.

Ключевые слова: КОРПУСНАЯ ЛИНГВИСТИКА, КОРПУС, КОНКОРДАНС, КИТАЙСКИЙ ЯЗЫК, ДИАЛЕКТ, ХАККА, СОХРАНЕНИЕ ЯЗЫКОВОГО РАЗНООБРАЗИЯ.

Цель: разработка текстового корпуса китайского языка на основе диалекта Хакка и последующий анализ корпуса.

Задачи: 1) подготовка описательной характеристики теоретической базы корпусной лингвистики; 2) анализ проблемы создания корпуса в рамках корпусной лингвистики; 3) формулировка цели создания корпуса на материале китайского языка; 4) описание теоретических основ русской и китайской диалектологии; 5) подбор материалов для создания корпуса китайского языка; 6) конструирование корпуса текстов китайского языка диалекта Хакка на материале видеороликов платформы Bilibili

Актуальность выбранной темы и сферы ее исследования обусловлена тем, что на сегодняшний день китайский язык занимает место в списке наиболее изучаемых языков. Вопрос сохранения диалектов является одной из важнейших проблем, связанных с исследованиями китайского языка. Диалект Хакка представляет особый интерес, так как наименее исследован российскими учеными. Создание корпуса на материале Хакка позволит упростить процесс работы с языковыми особенностями данного диалекта.

Основные выводы и результаты исследования:

1. Корпусная лингвистика – актуальное научное направление, в рамках которого происходит конструирование корпусов с целью упрощения работы ученых в сфере работы с большими объемами данных.

2. Диалектная ситуация в современном Китае неоднозначна, однако тенденция к защите культурного языкового наследия, в том числе диалектов, превалирует над тенденцией к вытеснению диалектов.

3. Большинство масштабных корпусов строятся на основе публицистических или литературных текстов, в связи с чем создание корпуса на основе диалекта является новым направлением в корпусной лингвистике.

4. Создание параллельного корпуса позволяет проводить сравнительный анализ двух языковых систем, в нашем случае диалекта Хакка и путунхуа.

5. Отличительными особенностями диалекта Хакка является устройство фонетической и лексической системы, что тесно связано с влиянием соседних диалектов и древнекитайского языка. Грамматический строй Хакка во многих аспектах совпадает с правилами, принятыми в путунхуа.

Перспективы дальнейшего исследования: 1) увеличение объема текстового корпуса диалекта Хакка, 2) углубленное исследование связей диалекта Хакка с древнекитайским языком, 3) добавление новых языков в корпус.