Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)
САЕ «Институт человека цифровой эпохи»
Автономная магистерская программа
«Компьютерная и когнитивная лингвистика»

ДОПУСТИТЬ К ЗАЩИТЕ В ГЭК

Руководитель ООП
д-р филол. наук, профессор
_____ З. И. Резанова
«26» июня 2020 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

СИСТЕМА РАЗМЕТКИ И МЕТАРАЗМЕТКИ В КОРПУСЕ РУССКОЙ УСТНОЙ РЕЧИ ТЮРКСКО-РУССКИХ БИЛИНГВОВ (RuTuBiC)

по основной образовательной программе подготовки магистров

направление подготовки 45.04.03–Фундаментальная и прикладная лингвистика

Погодаева Елена Николаевна

Научный руководитель,
д-р. филол. наук, доцент
_____ З.И. Резанова
подпись
«26» июня 2020 г.

Автор работы
студент группы № 131883
_____ Е.Н. Погодаева
подпись

Томск-2020

Оглавление

Введение

Во второй половине двадцатого века в отечественной и зарубежной лингвистике выделилось направление, цель которого – изучение языковых закономерностей на материале структурированных собраний текстов. Основной отличительный признак таких собраний текстов – возможность осуществления поиска по заданным параметрам для дальнейшего анализа полученных результатов. Такое собрание текстов, объединённых по какому-либо признаку называют текстовым корпусом. Раздел лингвистики, использующий текстовые корпуса для анализа данных с помощью корпусных методов, называется корпусная лингвистика. Являясь исходным материалом для корпусной лингвистики, корпус текстов в то же время является её основным продуктом. На данном этапе развития технологий сбора, обработки и хранения данных наибольшей ценностью обладают корпуса текстов, снабжённых лингвистической и метаинформацией, позволяющей в дальнейшем осуществлять различные виды анализа в соответствии с исследовательскими задачами, которые ставят перед собой создатели корпуса. В данной работе представлена попытка создания системы лингвистической разметки и метаразметки в корпусе русской устной речи тюркско-русских билингвов RuTuBiC.

Актуальность данной работы определяется потребностью в создании корпуса, содержащего примеры реального языкового употребления и предоставляющего возможность для фиксации и анализа типов речевых отклонений на всех уровнях языковой системы в соотнесении с типами языкового контактирования. Необходимость изучения языкового контактирования и создания подобных корпусов обусловлена, прежде всего, усиливающимися тенденциями глобализации на основе русского языка, являющегося единственным официальным языком на территории Российской

Федерации и определяющего социальную реальность сосуществования представителей разных народов в многонациональном государстве.

Существует ряд российских и зарубежных проектов, посвящённых изучению влияния материнских языков населения разных регионов на устную и письменную речь билингвов, использующих русских язык. Следует отметить, что респонденты, чья речь была использована в качестве материала корпуса RuTuBiC, являются именно носителями естественного билингвизма, однако так же существуют и корпуса, созданные для изучения учебного билингвизма. Среди российских проектов особенный интерес представляют такие корпуса как корпус контактно-обусловленной русской речи носителей Севера Сибири и Дальнего Востока, материалы которого расшифрованы и размечены П.С. Плешак, Н.М. Стойновой и И.А. Хомченковой и Русский учебный корпус[1] — проект НИУ ВШЭ, осуществляемый Лабораторией по корпусным исследованиям под руководством Е.В. Рахилиной. Целью создания обоих корпусов было изучение разных аспектов устной и письменной речи носителей естественного и учебного билингвизма. Кроме того, интерес представляют и корпуса, содержащие русскую устную речь монолингвов, прежде всего корпуса, созданные в рамках проекта «Рассказы о сновидениях и другие корпуса звучащей речи»[2]. Каждый из перечисленных корпусов имеет мотивированную систему разметки и описание технических аспектов создания корпуса, представленные в публикациях[3].

1 URL: http://web-corpora.net/RLC.

2 URL: http://spokencorpora.ru/.

3 Khomchenkova I. A., Pleshak P. S., Stoynova N. M. The Corpus of Contact-Influenced Russian of Northern Siberia and the Russian far East // Papers from the Annual International Conference “Dialogue”. M.: RSUH. 2019. P. 253-264. Rakhilina E., Vyrenkova A., Mustakimova E., Ladygina A., Smirnov I. Building a learner corpus for Russian // Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umea, 16th November 2016. Рассказы о сновидениях: корпусное исследовение русского устного дискурса / Под ред. А.А. Кибрика и В.И. Подлесской, М.: Языки славянских культур, 2009. — 736 с.: ил.

Корпусная лингвистика является разделом компьютерной лингвистики, поэтому работа с корпусами текстов тесно связана с использованием компьютерных технологий и программного обеспечения, дающего исследователям возможность быстрого получения и обработки информации. Развитие и совершенствование технологий обработки и хранения записей аудиофайлов также сделало возможным создание корпусов звучащей речи. Несомненным преимуществом звукозаписывающих технологий – возможность повторного воспроизведения аудиофрагмента, что при расшифровке позволяет отразить все особенности употребления различных единиц языка. В зависимости от цели создания корпуса звучащей речи могут содержать как короткие фрагменты текстов, записанные в лабораторных условиях, так и записи неподготовленной устной речи в ситуациях реального общения. Записи лекций или других публичных выступлений в таких случаях не являются достаточно репрезентативным материалом, так как в них представлена «озвученная» норма письменного литературного языка.

Корпус русской устной речи тюркско-русских билингвов RuTuBiC создаётся в рамках проекта «Языковое и этнокультурное разнообразие Южной Сибири в синхронии и диахронии: взаимодействие языков и культур»[1], цель которого — выявление закономерностей исторического развития и современного состояния языков и культур Южной Сибири в аспекте их взаимодействия на основании современных языковых, антропологических и психолингвистических данных с применением корпусных и психолингвистических методов исследования. Основной отличительной особенностью корпуса является бимодальность — сочетание двух типов внешних стимулов, воспринимаемых человеком, которому адресован текст. Русский язык, являясь вторым, не материнским, активно используется авторами текстов корпуса во многих сферах, прежде всего в

1 URL: http://p220.ru/labs/laboratoriya-lingvisticheskoy-antropologii/.

институциональной коммуникации[1]. Корпус содержит три подкорпуса: шорско-русский, татарско-русский, хакасско-русский.

Основной **материал** корпуса — полевые записи устной речи, собранные в течение трёх лет г. Шерегеш и Таштагол, пос. Большая Суета Кемеровской области, г. Абакан, с. Аскиз, с. Чиланы, д. Юрт-Оры и Акбалык. В настоящее время корпус включает в себя более пятисот часов звучания. В записи интервью участвовали респонденты трёх возрастных групп.

Объектом исследования в данной работе является корпус русской устной речи тюркско-русских билингвов Южной Сибири.

Предмет исследования – система лингвистической разметки и метаразметки мультимодального корпуса.

Цель данной работы – описание типологически релевантных признаков создаваемого корпуса и инструментов, используемых в процессе его создания.

Для достижения цели были поставлены следующие **задачи**:

1. дать общую характеристику корпуса;
2. описать типы разметки, применяемые к материалам корпуса;
3. выбрать и апробировать подходящие методы автоматической обработки текстовых и аудиоматериалов.

В работе используется следующее программное обеспечение для работы с материалами корпуса: автоматический морфологический анализатор Mystem[2] и программа для создания и аннотирования мультимодальных файлов ELAN[3].

---

1 Резанова З.И. Подкорпус устной речи русско-тюркских билингвов Южной Сибири: типологически релевантные признаки // Вопросы лексикографии. 2017. No 11. С. 105–118.

2 URL: https://yandex.ru/dev/mystem/.

3 URL: https://archive.mpi.nl/tla/elan.

Теоретической основой исследования являются работы, посвящённые типологии корпусов[1], аннотированию мультимодальных данных[2], а также исследования в области семантики[3] и компьютерной лингвистики[4].

Методологическую основу исследования составляют методы разных наук: методы автоматической и полуавтоматической обработки текстовых и звуковых данных, корпусные методы, методы исследования семантики.

Теоретическая значимость работы определяется её вкладом в дальнейшую разработку системы разметки корпуса. Апробированные инструменты обработки текстовых и аудиоматериалов могут быть использованы в дальнейшем для решения задач морфологического анализа и синхронизированного представления звуковых и текстовых данных. **Практическая значимость** данной работы заключается в том, что построенная последовательность действий при работе над корпусом может быть использована при работе над корпусом с использованием методов автоматической обработки текстовых и аудиоматериалов.

Структура работы: работа включает в себя введение, две главы, заключение и список литературы.

Во введении обосновывается актуальность исследования, формулируются объект, предмет, цель и решаемые задачи, приводятся примеры литературы, освещающие историю вопроса, определяются теоретическая и практическая значимость.

---

1 См. например: Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн., – СПб.: СПбГУ. РИО. Филологический факультет, 2013. URL: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

2 См. например: Ide N., Pustejovsky J. Handbook of Linguistic Annotation / Springer 2017.

3 См. например: Бабенко Л.Г. Русские глагольные предложения: экспериментальный синтаксический словарь. Под общ. ред. Л.Г. Бабенко. М.: Флинта: Наука, 2002.

4 См. например: Марчук Ю.Н. Компьютерная лингвистика: учеб. пособие. М.: АСТ: Восток — Запад, 2007. Adolphs S., Knight D. Building a spoken corpus: What are the basics? // The Routledge Handbook of Corpus Linguistics / ed. by Anne O'Keeffe and Michael McCarthy, 2010. Jurafsky, Daniel & Martin, James. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2008.

В первой главе описано развитие методологии корпусной лингвистики и типология корпусов, а также дана характеристика корпуса RuTuBiC согласно основных типологически релевантных признаков.

Вторая глава посвящена обоснованию выбора и описанию работы инструментов автоматической обработки текстовых и аудиоматериалов и их применению в осуществлении некоторых видов анализа текста, описана репрезентационная схема мультимодальной разметки. Также во второй главе описано многоуровневое аннотирование отклонений от речевого стандарта.

В заключении подводятся основные итоги исследования и определяются перспективы его дальнейшего развития.

Апробация работы. Основные этапы работы над корпусом были представлены на научно-практических мероприятиях: VI (XX) Международная конференция молодых учёных «Актуальные проблемы лингвистики и литературоведения», НИ ТГУ, 18–20 апреля 2019; научная школа «Контактирование языков: лингвистический, социолингвистический, психолингвистический аспекты», (32 часа), 13-16 мая 2019; XXX ежегодная Международная научная конференция «Язык и культура», 16-19 сентября 2019 года; Двадцатые филологические чтения «Интерпретационный потенциал языковой системы и творческая активность говорящего: взаимодействие лексической и грамматической семантики», 17-18 октября 2019 года; Международная конференция молодых учёных и педагогов «Проблемы сохранения культурно-языкового разнообразия Российской Федерации», апрель 2020 (тезисы доклада прошли экспертизу); VII (XXI) Международная научно-практическая конференция молодых ученых «Актуальные проблемы лингвистики и литературоведения», НИ ТГУ, 16-18 апреля 2020 года.

Глава 1. RuTuBiC: лингвистически размеченный корпус текстов – определение понятий

1.1. Развитие методологии корпусной лингвистики

Корпусная лингвистика является относительно новым направлением лингвистической науки. На данный момент корпусную лингвистику принято рассматривать как раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий[1]. В отличие от разделов лингвистики, имеющих своей целью описание языка или оценку языковой структуры, корпусная лингвистика представляет собой совокупность методов, процедур и ресурсов работы, применяемых к разным аспектам языковых исследований. Также исследователи определяют корпусную лингвистику, как деятельность, необходимую для составления или использования корпуса и направленную на исследование естественного употребления языка.

Современная корпусная лингвистика восходит к историко-лингвистическим исследованиям конца XVIII — начала XIX веков. В сравнительно-историческом языкознании использовали корпусные методы исследования языков, так как языковой материал был представлен прежде всего в виде значительных по объёму собраний текстов, относящихся к разному времени и месту создания. Именно цитаты из текстов являлись основным иллюстративным материалом для описания истории и грамматики. Представители младограмматизма утверждали, что общение с помощью языка возможно именно благодаря тому, что естественный язык имеет общественную, а не биологическую природу. Выбранный в связи с этим путь развития лингвистики, как науки, занимающейся изучением не только

1Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн., – СПб.: СПбГУ. РИО. Филологический факультет, 2013.

мёртвых, но и функционирующих в данный момент языков и их диалектов, так же во многом повлиял на развитие корпусной лингвистики не только как науки, но и методологии исследования.

Ещё одним важным направлением, повлиявшим на современное состояние корпусной лингвистики, является составление грамматик и словарей. Так Герман Пауль, используя в качестве примеров для описания немецкой грамматики произведения своих современников — классиков немецкой литературы, утверждал, что научный подход к языку возможен не только при изучении стадий развития одного и того же языка, но и при сопоставлении имеющегося современного материала, представленного данными нескольких родственных языков или диалектов одного языка[1]. Подобный подход можно проследить и в наше время, однако для получения более достоверных сведений составители современных национальных корпусов используют тексты разных жанров, что позволяет, в частности, исследовать грамматические особенности устной речи. Так в Национальном корпусе русского языка представлены несколькими подкорпусами, в числе которых основной корпус, содержащий тексты, иллюстрирующие русский литературный язык, газетный корпус (корпус современных СМИ), корпус региональной и зарубежной прессы, корпус параллельных текстов, корпус диалектных текстов, корпус поэтических текстов, корпус устной речи, акцентологический корпус и мультимедийный корпус[2].

Лексикографы XVIII века при составлении толковых словарей в качестве иллюстраций подбирали цитаты из книг известных авторов таким образом, чтобы контекст облегчал понимание значения лексической единицы. Сэмюэл Джонсон создал первый подобный словарь английского языка, к которому в наши дни обращаются не только в учебных и исследовательских

---

1 Резанова, З. И. История языкознания: XIX - первая половина XX века: Хрестоматия : учебное пособие : в 2 частях / З. И. Резанова. — 2-е изд., стер. — Москва : ФЛИНТА, [б. г.]. — Часть 1 : 2 — 2012. — 264 с.

2URL: http://www.ruscorpora.ru/new/corpora-structure.html.

целях, но и в юридической практике для более подробного уточнения значений слов[1]. Несмотря на то, что современные компьютерные технологии в значительной степени ускоряют поиск слова в подходящем контексте, сама идея использования минимального контекста для иллюстрации слова совпадает с принципами составления толковых словарей в доцифровую эпоху. Однако корпусная лингвистика в своём современном состоянии следует скорее дескриптивной, чем прескриптивной традиции, так как описывает в том числе и случаи языкового употребления, не относящиеся к письменному литературному языку. Этому также поспособствовало развитие полевой лингвистики, так как основным материалом в данном случае являются аудио- и видеозаписи, и их расшифровки. Как правило, респондентами в таких случаях являются люди разного возраста и уровня образования, а коммуникация выстроена на основе тем, отражающих подробности повседневной жизни респондентов.

Описанная грамматика и лексика какого-либо языка являются основой для его изучения. В лингводидактике долгое время существовала традиция изучения языка через усвоение его нормативного варианта, в то время как аутентичное языковое употребление может не вполне соответствовать принятой норме. На данный момент фокус в том числе в отечественной лингводидактике всё больше смещается в направлении изучения реального языкового употребления. С развитием современных технологий сбора и обработки не только текстовых, но и аудиовизуальных данных такой подход на сегодняшний день имеет достаточное количество реализаций не только в виде учебных пособий для разных категорий обучающихся, но и в виде программного обеспечения, используемого для поддержки интерактивной учебной деятельности. Примером такого программного обеспечения может

1URL: https://johnsonsdictionaryonline.com/

являться Обучающий подкорпус Национального корпуса русского языка, ориентированный на преподаваение русского языка в школе[1].

Следующее направление, оказавшее влияние на формирование методологии корпусной лингвистики как отдельной науки — составление карт диалектов и сборников диалектных выражений. Среди современных проектов в области корпусной лингвистики и лингвистической типологии можно назвать «Всемирный атлас языковых структур»[2] – одну из наиболее полных открытых баз данных, содержащую информацию о географическом распределении структурных признаков большинства языков мира.

Корпусная лингвистика в её нынешнем виде сохранила описанную методологию в том числе из-за смены общетеоретических приоритетов лингвистической науки – смещения фокуса с «языка» на «речь» или перехода от изучения лингвистических конструктов к изучению лингвистических фактов[3]. Однако результаты многих исследований стали намного достовернее благодаря доступности значительных объёмов данных в электронном виде и быстрому развитию современных информационных технологий.

1.2. Лингвистически размеченный корпус текстов – определение понятия

Несмотря на давнее использование корпусных методов в лингвистике, сами термины «корпусная лингвистика» и «лингвистический корпус» появились во второй половине XX века.

Существуют разные определения корпуса текстов. В русскоязычной литературе наиболее часто цитируемыми являются определения В. П.

1Савчук С. О, Сичинава Д. В. Обучающий корпус русского языка и его использование в преподавательской практике // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 317—334.

2 URL: https://wals.info/

3Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении, 2008, No. 16 (2), 7—20.

Захарова и М. В. Копотева. Первое определение гласит: «Под лингвистическим, или языковым, корпусом текстов понимается большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач»[1]. Второе определение уточняет некоторые важные особенности современных корпусов: «Корпус — собрание языковых материалов (текстов, аудио- и видеозаписей и т. д.), собранных в соответствии с определенными принципами, размеченных по определенному стандарту и обеспеченных специализированной поисковой системой. Корпусом (неразмеченным корпусом) могут называть любое собрание текстов, объединенных каким-то общим признаком (языком, жанром, автором, периодом создания текстов)»[2]. Так во втором определении его автор указывает на возможность наличия в корпусе не только текстовых, но и мультимедийных данных. Также из этого определения следует, что в понятие «корпус текстов» входит также и корпус-менеджер — специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления пользователю результатов в удобной форме. Анализ определений позволяет в том числе отследить и развитие корпусной лингвистики как науки. Приведённый М. В. Копотевым вариант определения корпуса как любого собрания текстов, отобранных по какому-либо признаку, отсылает нас к первым предшественникам современных корпусов — исследованиям священных текстов разных религий. Самым известным подобным исследованием являются конкордации — алфавитные постраничные перечни

1Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн., – СПб.: СПбГУ. РИО. Филологический факультет, 2013.

2 Копотев М. Введение в корпусную лингвистику. – Прага, Animedia Company, 2014.

слов и фраз, встречающихся в Библии[1]. Впервые конкордация была создана в XIII веке католическим монахом-францисканцем Антонио Падуанским. Определения из зарубежных источников содержат те же положения. Так Т. МакЭнери и Э. Вилсон определяют корпус как «собрание языковых фрагментов, соответствующих чётким языковыми критериям отбора для использования в качестве представления наиболее точной модели языка»[2].

Во всех приведённых определениях можно проследить общую тенденцию к определению корпуса как ограниченного по объёму массива языковых данных, отобранных по определённым критериям. Быстрое развитие компьютерных технологий во второй половине XX века позволило в разы ускорить процесс сбора и обработки данных корпуса, поэтому большую роль стали иметь именно критерии организации данных в корпусе в соответствии с его назначением. Вместо задачи по сбору достаточного количества текстового материала на первый план начинают выходить задачи, связанные с интерпретацией полученных данных с помощью методов лингвистического и математического анализа.

1.3 Типология лингвистически размеченных корпусов

Первый электронный корпус текстов, подвергнутый статистической обработке был создан в 1960-е гг. в США. The Brown Corpus of Standard English включал в себя 500 текстов по 2000 словоупотреблений из книг, газет и журналов, опубликованных не ранее 1961-ого года. Целью создания данного корпуса было изучение жанров письменного литературного английского языка. В 1961 году аналогичный корпус был создан в Великобритании — The Lancaster-Oslo-Bergen Corpus. При создании обоих корпусов была намечена важная тенденция — несмотря на ограничения,

1Adolphs S., Knight D. Building a spoken corpus: What are the basics? // The Routledge Handbook of Corpus Linguistics / ed. by Anne O'Keeffe and Michael McCarthy, 2010.

2 McEnery T., Wilson A. Corpus linguistics. Edinburgh: Edinburgh University Press, 1996.

заданные статистическими критериями отбора, при отборе текстов создателям корпуса приходилось в том числе опираться на профессиональную интуицию. Этот факт говорит о том, что в процессе создания больших корпусов для достижения оптимальных результатов в равной степени важны как стандартизация, так и ведение научной дискуссии.

На примере двух первых больших электронных корпусов так же можно выделить основные характеристики любого лингвистического корпуса, обеспечивающие достоверность результатов, полученных на его материале. Свойство корпуса, заключающееся в статистически достоверном представлении языка или его части, и достигаемое за счёт необходимого объема и жанрового разнообразия текстов, называется репрезентативность[1].

Так как корпусы текстов могут быть созданы для разных задач, выделяют следующие типы репрезентативности: 1) корпусы первого типа, отражающие в себе всё многообразие речевой деятельности; 2) корпусы второго типа, отражающие бытование некоторого культурного или языкового явления в речевой практике, построенные для определённой цели.

Первые электронные корпуса относились к корпусам первого типа, так как при их создании была предпринята попытка отразить всё многообразие письменного литературного английского языка. Тем не менее, в наше время их нельзя было бы назвать репрезентативными, так как жёсткое ограничение по объёму и типу данных не давало возможности показать всё многообразие языка. На сегодняшний день большинство корпусов первого типа являются репрезентативными, так как с момента создания первых корпусов технологии сбора, обработки и хранения информации были существенно улучшены, что даёт возможность для сбора данных разного типа. Корпусы второго типа, как правило, репрезентативны, так как изначально предполагают полное представление только заданных аспектов языкового употребления.

1 Копотев М. Введение в корпусную лингвистику. – Прага, Animedia Company, 2014.

Под сбалансированностью корпуса понимают представленность в нём разных типов текстов в равных или соответствующих реальному употреблению пропорциях. Следует отметить, что оценивать репрезентативность и сбалансированность корпуса нужно с учётом тех задач, для решения которых он был создан.

В мировой практике существует устоявшийся подход к классификации корпусов. Среди основных параметров классификации корпусов выделяют следующие.

Корпус может быть моноязычным, мультиязычным или смешанным. Моноязычный корпус может содержать тексты разных вариантов или диалектов одного языка. Мультиязычные корпуса, в свою очередь, делятся на два типа согласно содержанию: тексты, порождённые в ситуации многоязычного общения и тексты, имеющие одинаковое содержание, но переведённые на разные языки. Второй тип мультиязычных корпусов называют параллельными или двуязычными корпусами. Такие корпуса составляют преимущественно для задач машинного перевода. Некоторые исследователи при описании типологии корпусов называют описанный критерий именно критерием параллельности[1].

Согласно типу текстов корпуса разделяют на письменные, устные и мультимодальные. Данный классификационный параметр определяют согласно типу данных, которые содержатся в корпусе. Основным материалом для корпусов письменной речи является письменный текст. С развитием технологий оцифровки документов стало возможным создание электронных диахронических корпусов, содержащих тексты доцифровой эпохи. Один из таких корпусов — корпус старославянских текстов Corpus Cyrillo-Methodianum Helsingiense[2]. Корпуса устной речи содержат аудиозаписи,

1Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн., – СПб.: СПбГУ. РИО. Филологический факультет, 2013.

2 URL: http://www.helsinki.fi/slaavilaiset/ccmh/

которые могут сопровождаться расшифровкой, отражающей особенности произнесения слов. Детализация расшифровки в данном случае зависит от исследовательских задач, которые решают с помощью корпуса. Мультимодальные корпуса могут содержать как аудио-, так и видеозаписи, однако в отличие от корпусов устной речи, при разметке в фокусе находится тезис о том, что естественный дискурс по своей природе мультимодален[1] и состоит из синхронизированных между собой каналов коммуникации.

В соответствии с жанрами текстов выделяют литературные, диалектные, разговорные, публицистические, исторические корпуса и корпуса второго языка.

По степени представленности языкового материала выделяют полнотекстовые и фрагментированные корпуса. Частными случаями фрагментированного корпуса является корпус n-грамм и конкорданс.

Следующий важный классификационный параметр — тип разметки корпуса. Выделяют метатекстовую, лингвистическую, экстралингвистическую разметку. На сегодняшний день практически в каждом крупном корпусе присутствуют все три вида разметки, однако выбор параметров каждого типа разметки, как правило, зависит от цели создания корпуса.

Согласно определению М. Копотева, аннотация — это приписанная всем единицам выбранного уровня (текст, предложение, словоформа) соответствующая лингвистическая информация[2]. Аннотирование корпуса — это процесс приписывания текстам и их компонентам специальных тегов в соответствии с разными типами разметки[3].

1Adolphs S., Knight D. Building a spoken corpus: What are the basics? // The Routledge Handbook of Corpus Linguistics / ed. by Anne O'Keeffe and Michael McCarthy, 2010.

2 Копотев М. Введение в корпусную лингвистику. – Прага, Animedia Company, 2014.

3Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн., – СПб.: СПбГУ. РИО. Филологический факультет, 2013.

Термин «аннотация» может быть применён по отношению ко всем типам информации, передающим и описывающим акт коммуникации, в том числе транскрипту. Термины «аннотация» и «разметка» в русскоязычных работах по корпусной лингвистике являются взаимозаменяемыми.

В соответствии с принятой мировой практикой, для корпусов, представленных в открытом доступе, существуют следующие основные требования к разметке: мотивированность и теоретическая нейтральность. Под мотивированностью разметки понимают наличие чёткой схемы анализа текста, в которой обозначена связь каждого из параметров разметки с задачами, которые ставят перед собой создатели корпуса. Теоретически нейтральной разметкой считается разметка, опирающаяся на существующие стандарты корпусной лингвистики и использующая систему понятий и обозначений, приближенных к принятым в мировой практике. Таким образом, при соблюдении перечисленных требований создателями корпуса, пользователям корпуса должна быть доступна инструкция пользователя, содержащая стандартизированную систему понятий.

Метатекстовая разметка содержит «данные о языке» или «данные о данных», характеризующие коммуникативный акт по заданным параметрам. По мнению исследователей, именно метатекстовая разметка позволяет сделать корпус репрезентативным и сбалансированным[1]. Метаразметка имеет ряд в структуре корпуса ряд других важных функций: формирование архитектуры корпуса, контроль информационного наполнения корпуса, обеспечение возможности поиска и отбора текстов пользователем для составления подкорпусов с заданными свойствами[2].

1Sinclair, J. (1996) EAGLES. Preliminary recommendations on Corpus Typology. EAG–TCWG–CTYP/P. Version of May, 1996.

2Савчук С. О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 62—88.

Метаразметка корпуса тюркско-русских билингвов основана на данных двух социолингвистических анкет и включает в себя информацию о времени и месте рождения, проживания, обучения, профессиональной деятельности, сведений о родственниках по разным типам родства, о способе приобретения и использования языков.

На данный момент в мультимедийной разметке представлено четыре параметра: тип речи, тип дискурса, жанровая и тематическая принадлежность.

Несмотря на то, что интервью относят к диалогическим речевым жанрам, в некоторых текстах присутствуют монологические фрагменты, так как респондент не только даёт ответы на вопросы социолингвистических анкет и анкет языкового опыта, но и рассказывает о релевантных бытовых ситуациях.

Полилог в системе разметки выделен в отдельную категорию, однако встречается сравнительно редко и представляет собой реплики других респондентов или интервьюеров, относящиеся к обсуждаемой теме. Полилог также можно считать сигналом перехода от институционального дискурса к личностному, так как интервью является диалогическим жанром речи.

Определение типа дискурса зависит от выбора подхода к классификации дискурсов. Согласно распространённой практике, при анализе текста с позиций социолингвистики различают два типа дискурса, институциональный и личностный. Институциональный дискурс задаёт статусно-ролевые границы коммуникации, в которой говорящие выступают представителями определённого социального института.

Несмотря на то, что даже в рамках научного интервью общение не является полностью обезличенным, противопоставление институционального и личностного дискурса является даёт дополнительные возможности для анализа социокультурных ситуаций общения, типов коммуникативных личностей и способов организации текста. В

противоположность институциональному, личностный дискурс представляет говорящего носителем своего внутреннего мира.

Основными характеристиками личностного дискурса являются спонтанность, выраженная субъективность, а также употребление сниженной лексики и нечёткое беглое произношение[1].

Так как метаинформация не является целевой для корпуса интерферентных отклонений в устной речи билигвов, создателями корпуса было принято решение исключить микрожанры из перечня, что позволило бы повысить уровень абстракции и теоретической нейтральности разметки. В корпусе представлены такие жанры как интервью, беседа, разговор, рассказ, история, описание.

Относительно перечисленных уровней метаразметки тематическая разметка представлена наиболее подробно и на данный момент содержит двенадцать тем: рассказ о памятном периоде жизни, значимое событие в жизни, рассказ о типовом образце протекания жизни, характеристики окружающих людей, семья, рассказ о себе, рассказ об окружающем мире, рассказы о различных аспектах социального существования респондента, рассказы о народной культуре, развёрнутое выражение мнения, обмен мнениями, комментарий к обстоятельствам диалога.

Лингвистическая разметка – это все типы лингвистической информации приписываемой единице языка. Существующие типы лингвистического аннотирования соответствуют уровням языка и имеют соответствующие минимальные единицы[2]. Сетка разметки содержит три уровня, содержащих теги и словоформы.

Морфологическая разметка содержит информацию частеречной принадлежности слова и его морфологических категориях. Это основной тип

1Карасик В.И. О типах дискурса // Языковая личность: институциональный и персональный дискурс: Сб. науч. тр. Волгоград: Перемена, 2000. - С.5-20.

2 Копотев М. Введение в корпусную лингвистику. – Прага, Animedia Company, 2014.

разметки для большинства крупных корпусов, так как морфологический анализ является основой для дальнейших форм анализа.

В традиционном языкознании под морфологией понимают только то, что характеризует внешнюю форму слова. Однако в автоматической обработке текста морфологический анализ на основе данных о форме слова даёт сведения о разных уровнях языковой структуры[1].

Морфологический анализ текста включает следующие этапы: 1) лексическая обработка фразы с использованием словаря основ для выделения исходных слов и распределения их по морфологическим и семантическим принципам; 2) идентификация окончания.

Основная функция морфологической разметки в корпусе - обеспечение поиска заданных слов, словоформ и конструкций.

Морфологическая разметка осуществляется автоматически с помощью программы Mystem. Система морфологических признаков, заложенная в данной программе, соответствует стандартам русского письменного литературного языка, лежащим в основе «Грамматического словаря русского языка» А.А. Зализняка.

В текстовом материале корпуса отражены особенности устной речи, которые при осуществлении морфологического анализа фиксируются программой Mystem как отклонения.

Следующий уровень лингвистической разметки — аннотирование отклонений от речевого стандарта. Основное отличие создаваемого корпуса от существующих корпусов учебных и билингвальных корпусов заключается в том, что система разметки ошибок ориентирована на стандарты устной речи. Теги ошибок охватывают четыре уровня языковой системы. фонетику, морфологию, лексику и синтаксис.

1 Марчук Ю.Н. Компьютерная лингвистика: учеб. пособие. М.: АСТ: Восток — Запад, 2007

Перечисленные группы тегов представлены с разной степенью детализации, что обусловлено исследовательскими задачами, для которых создаётся корпус.

Каждому тегу группы фонетических, морфологических и лексических отклонений соответствует аннотация, содержащая форму коррекции.

Соотнесение параметров метаразметки и лингвистической разметки позволяет осуществлять точный поиск по заданным параметрам для соотнесения типа языковой интерференции и социокультурных характеристик респондента[1].

1.4 Корпус RuTuBiC: типологические характеристики

В зависимости от цели и назначения корпуса при отборе текстов для корпуса в центре внимания могут находиться разные классификационные параметры.

Гипотеза авторов проекта состоит в том, что устная речь билингвов содержит проявление интерференции материнских языков тюркской группы, поэтому целью создания корпуса является репрезентативное представление устной речи билингвов[2].

По назначению корпус является исследовательским, так как в отличие от иллюстративных корпусов он создается для решения широкого круга типологических и психолингвистических задач.

Создаваемый корпус является смешанным. Несмотря на наличие слов и реплик на материнском языке корпус не является параллельным, так как целевым языком корпуса является русский. Корпус также нельзя назвать

1Резанова З. И., Веснина Г. Ю. Подкорпус русской речи билингвов лингвистического корпуса «Томский региональный текст»: принципы разметки и метаразметки корпуса // Вопр. лексикографии. – 2016. – No 1 (9). – С. 29–39.

2Резанова З. И. Корпус устной речи русско-тюркских билингвов Южной Сибири: разметка отклонений от речевого стандарта // Вопросы лексикографии, No. 15, 2019, С. 127- 140.

моноязычным, так как цель его создания заключается в отслеживании и исследовании переключения кодов.

Корпус является полнотекстовым, так как текстовая часть материала представлена в виде полных расшифровок интервью с респондентами, включающих как речь респондента, так и речь интервьюера. Использование специального программного обеспечения для многоуровневого аннотирования корпуса позволяет осуществлять поиск по n-граммам - цепочкам идущих подряд токенов, где n - их количество. Практически во всех современных корпусах также предусмотрена возможность вывода результатов поиска в формате KWIC (key words in context), являющимся одним из способов отображения конкорданса - списка вхождений заданного токена или леммы.

Объём корпуса составляет пятьсот пятьдесят часов звучания, что является достаточным для репрезентативного представления исследуемых лингвистических явлений.

Так как корпус содержит аудио-, а не видеозаписи, возможно его последующее представление в открытом доступе. Записи интервью велись в соответствии с нормами Этического комитета Международного центра исследований развития человека ТГУ, все информанты подписывали листы информированного согласия о конфиденциальности информации. В корпусе тексты представлены как анонимные, данные об информантах хранятся в зашифрованном виде. Шифр интервью позволяет получить информацию о материнском языке респондента, порядковом номере интервью в каждом отдельном подкорпусе и порядковом номере интервью с конкретным респондентом в случае, если с одним респондентом было записано несколько интервью.

Дискурсивное, жанровое и тематическое разнообразие текстов, представленных в корпусе, ограничено вопросами анкет, используемых для

сбора метатекстовой информации, а также обсуждениями тем и ситуаций, значимых для респондента.

Ключевым дифференциальным параметром создаваемого корпуса является тип соотношения языков, используемых автором при порождении текста[1]. Согласно данному критерию, автор текста может являться или не являться носителем языка. Носителями языка не принято считать респондентов, изучающих язык как иностранный, регулярно использующих неродной язык для общения в разных ситуациях, а также билингвов. В соответствии с перечисленными категориями не носителей языка, существуют корпуса ученических текстов, корпуса лингва франка и эритажные корпуса. Тексты могут быть отобраны для корпуса согласно своему функциональному положению в коммуникации билингва, так как предметом изучения может быть как материнский (унаследованный) язык, так и изучаемый второй язык. Разметка корпусов “неносителей” или же “носителей нескольких языков” предполагает наличие разметки ошибок, содержащей информацию о языковом уровне, формальных особенностях и источнике ошибки[2].

Русский язык, являясь целевым языком корпуса, не является материнским языком для авторов текстов корпуса. При этом уровень владения русским языком равен или превышает уровень владения материнским языком. Материнские языки контактирования - шорский, хакасский и татарский. Важным отличием от других корпусов подобного типа также является то, что материалом корпуса являются не учебные письменные работы или заранее подготовленные устные сообщения, а

1Резанова З. И. Подкорпус устной речи русско-тюркских билингвов Южной Сибири: типологически релевантные признаки // Вопросы лексикографии. 2017. No 11. С. 105–118.

2Резанова З. И., Веснина Г. Ю. Подкорпус русской речи билингвов лингвистического корпуса «Томский региональный текст»: принципы разметки и метаразметки корпуса // Вопр. лексикографии. – 2016. – No 1 (9). – С. 29–39.

свободные беседы на темы, близкие респондентам, освещающие различные аспекты их повседневной жизни[1].

Следующим важным классификационным параметром, требующим подробного рассмотрения, является форма речи, определяемая средствами выражения высказывания. В данном случае различают три типа текстов корпуса: письменные, устные и мультимодальные. Последний тип представляет особый интерес, так как естественный дискурс по своей природе мультимодален[2].

Материалом корпусов такого типа могут являться аудио- и видеозаписи. Главное отличие разметки мультимодального корпуса от корпуса письменных текстов заключается в том, что она связана не только с цепочками символов, но и с единицами, не вложенными в письменный текст (жестами или направлением взгляда). Наиболее простым вариантом мультимодального корпуса является корпус звучащей речи, содержащий аудиофрагменты и орфорграфическую транскрипцию.

Следует отметить, что форма реализации речи (устная или письменная) — формальный критерий, не являющийся достаточным для разграничения сфер функционирования языка. Поэтому исследователи выделяют «подлинную» и «неподлинную» устную речь[3]. Многие лингвисты отмечают, что подготовленное публичное выступление следует рассматривать как часть кодифицированного литературного языка и, соответственно, относить к книжным стилям — научному, деловому или публицистическому. В данном случае именно они, а не форма реализации речи в большей степени

1Резанова З. И., Веснина Г. Ю. Подкорпус русской речи билингвов лингвистического корпуса «Томский региональный текст»: принципы разметки и метаразметки корпуса // Вопр. лексикографии. – 2016. – No 1 (9). – С. 29–39.

2Adolphs S., Knight D. Building a spoken corpus: What are the basics? // The Routledge Handbook of Corpus Linguistics / ed. by Anne O'Keeffe and Michael McCarthy, 2010.

3Земская Е. А., Китайгородская М. В., Ширяев Е. Н. Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис / Е. А. Земская и др. М: Наука, 1981.

определяют характер используемых языковых средств. Таким образом, доклады, лекции и другие виды публичных выступлений, подготовленные монологи и диалоги в кино и театре следует относить к «неподлинной» устной речи, так как все эти жанры основаны на письменной кодифицированной речи. К «подлинной» устной речи в таком случае относится именно разговорная речь.

Выделяют следующие различия между двумя разновидностями русского литературного языка[1]. 1) Реализация разговорной речи и кодифицированного литературного языка определяется экстралингвистическими условиями. 2) Разговорная речь — это сфера коммуникации особого рода, реализующая специфическую языковую систему, которой свойственно отсутствие ограничений узуального и лексического характера в построении единиц разных уровней и реализации грамматических значений, черты аналитизма и взаимодействие вербальных средств и невербальными. На лингвистические особенности разговорной речи оказывают влияния условия её протекания. 3) Нормы разговорной речи отличаются высокой вариативностью. Это объясняется зависимостью разговорной речи от речевого этикета и поведенческих норм. 4) Разговорную речь нельзя считать сниженным стилем литературного языка, так как он является не самостоятельной языковой системой, а набором средств, используемых для придания тексту непринуждённости. 5) Наличие в разговорной речи разнородных элементов объясняется неограниченностью тематического диапазона и допустимостью использовать различного рода инкрустации и заимствования (просторечие, жаргонизмы, диалектизмы, профессионализмы, канцеляризмы, шаблоны публицистического стиля).

По мнению составителей Национального корпуса русского языка, представленность устной речи в корпусе не только в текстовом виде, но и в

1 Земская Е. А., Китайгородская М. В., Ширяев Е. Н. Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис / Е. А. Земская и др. М: Наука, 1981.

аудиоформате важна для понимания процессов, происходящих в современном русском языке. Авторы статьи «Зачем нужен национальный корпус русского языка?» отмечают, что нормы письменной и устной речи для языков с давней письменной традицией расходятся достаточно сильно, так как письменная речь, как правило, более консервативна[1]. Это утверждение справедливо не только для русского, но и для многих европейских и азиатских языков с древней литературной традицией.

Так как по назначению корпус является исследовательским, представленные типы разметки были выбраны в соответствии с задачами, для решения которых он создаётся. Разметка мультимодального корпуса представляет собой несколько связанных между собой слоёв разметки, объединённых в рамках одного акта коммуникации. На данный момент разметка корпуса состоит из девяти слоёв, три из которых относятся к лингвистическому аннотированию, четыре к метатекстовому и ещё два являются базовыми для семи перечисленных, так как содержат полную текстовую запись интервью и отдельно текстовую запись речи респондента.

Выводы по 1 главе

В данной главе были рассмотрены основные теоретические положения корпусной лингвистики, даны определения ключевых терминов, относящихся к типологии корпусов и процессу аннотирования. Также было описано развитие методологии корпусной лингвистики. Данное описание позволяет сделать вывод о том, что корпусная лингвистика как наука сохранила описанную методологию, однако с развитием современных информационных технологий достоверность результатов многих исследований существенно возросла.

1Плунгян В. А. Зачем нужен национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005.

Центральным понятием корпусной лингвистики является лингвистический корпус. Анализ определений, приведённых в разных исследованиях, помог выделить общую тенденцию к определению корпуса как ограниченного по объёму массива языковых данных, отобранных по заданным критериям.

Типологические признаки лингвистически размеченных корпусов определяют критерии сбора и организации материалов корпуса. Среди прочих значимых классификационных параметров особое место занимает аннотирование, так как от точности выбора параметров лингвистической разметки и метаразметки точность поисковой выдачи в корпусе.

Описание типологических характеристик и параметров разметки корпуса русской устной речи тюркско-русских билингвов позволило получить чёткое представление об особенностях материала корпуса и дальнейших этапах работы с ними.

Глава 2. Многоуровневое аннотирование бимодального корпуса

2.1. Автоматическая морфологическая разметка

Выбор вида разметки в корпусе зависит от цели его создания и типов данных, составляющих его. Можно выделить виды разметки, являющиеся необходимым для функционирования большинства крупных корпусов. В число таких видов разметки входит морфологическая, для получения которой необходимо осуществление морфологического анализа.

Морфологический анализ — это получение леммы или основы заданного токена, а при необходимости, морфологических параметров[1]. Данное определение морфологического анализа содержит основные понятия, фигурирующие в литературе, посвящённой теории и практике морфологического анализа: лемма, токен, морфологические параметры. Токен является минимальной единицей морфологического анализа, что в автоматической обработке текстовых данных может рассматриваться как цепочка символов без разделителя. Лемма — это нормальная форма токена, в словаре соответствующая начальной форме слова.

В настоящее время существует три основных подхода к осуществлению автоматического морфологического анализа: подход с использованием правил, статистические методы машинного обучения и гибридные методы, использующие и правила, и статистику.

В соответствии с основными терминами, можно выделить основные этапы морфологического анализа, токенизацию и лемматизацию.

Токенизация — это совокупность операций графематического анализа текста. Разные этапы токенизации могут осуществляться независимо друг от друга вручную или с помощью средств автоматического морфологического анализа, однако такой подход скорее свойственен отработке учебных задач,

1 Большакова Е.И.. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. Пособие— М.: МИЭМ, 2011.

при реализации морфологического анализа, результаты которого затем войдет в корпус, осуществляется весь перечень шагов. Перед тем, как их перечислить, необходимо уточнить, что данный перечень актуален только для языков, в которых на письме используется пробел. Разделение поданного на вход текста на цепочки символов предполагает, что алгоритм распознает как токен не только реально существующие слова естественного языка, но и разделители, например, знаки препинания. Затем происходит отделение небуквенных символов для их последующего удаления. Одной из главных проблем при таком подходе является наличие сокращений, так как сокращённая форма одного слова может являться полной формой другого: «рис.» не распознаётся алгоритмом как сокращённая форма от имени существительного «рисунок», но распознаётся как полная форма имени существительного «рис». В некоторых исследованиях последних лет[1] предлагается решение таких проблем посредством использования лингвистических правил и описания повторяющихся токенов с помощью регулярных выражений. Аналогичный подход также применяется в случае некорректной токенизации слов с дефисом.

Для каждого корректно идентифицированного токена можно определить начальную форму — лемму. В основе реализации этого процесса лежит использование словаря основ. Модули морфологического анализа, используемые для работы с русским языком, базируются на грамматическом словаре А.А. Зализняка, содержащем более ста тысяч слов с указанием особенностей их изменения. Частным случаем лемматизации является стемминг, однако в процессе стемминга не используется словарь основ, в данном случае алгоритм выделяет и отсекает изменяемую часть слова, оставляя для обработки предполагаемую основу.

1 Ермакович М. В. Автоматическое определение границ слова в русском языке / Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2017» Москва, 31 мая — 3 июня 2017.

Как лемматизацию, так и стемминг можно применить к большому количеству распространённых естественных языков, однако при выборе между лемматизацией и стеммингом следует учитывать все особенности целевого языка и имеющиеся в распоряжении технические ресурсы. Стеммеры имеют более высокую скорость работы, однако точность результатов при анализе флективных языков, в том числе русского, зависит от возможного количества флексий для отдельно взятого языка. Однако на сегодняшний день стеммеры, изначально разработанные с учётом особенностей морфологии английского языка, например, Стеммер Портера (Snowball) имеют достаточно успешные реализации, пригодные для использования на русскоязычных текстах.

Однако в крупных корпусных проектаха, например, при создании Национального корпуса русского языка, принято использовать инструменты, изначально разработанные для анализа текста на русском языке. Так при морфологической разметке НКРЯ использовались системы автоматического морфологического анализа Mystem и Dialing.

Несмотря на то, что токенизация и лемматизация не дают информации о частеречной принадлежности слова и его морфологических характеристиках, эти этапы являются обязательными для осуществления морфологического анализа.

Для осуществления морфологической разметки в корпусе RuTuBiC была выбрана разработка компании «Яндекс» — автоматический морфологический анализатор Mystem, относящийся к группе инструментов морфологического анализа, работающих на основе правил. В качестве входных данных можно использовать как текстовую строку, так и текстовых файл в нужном формате и кодировке. Так как словарной базой для Mystem является словарь А.А. Зализняка, закрытый для пополнения при осуществлении морфологического анализа, идентификация слов, не входящих в словарь, возможна только с помощью подключения

пользовательского словаря, оформленного в соответствии с правилами, прописанными в документации[1] Mystem.

Mystem может быть использован как консольная программа, так и с помощью программного интерфейса модулей некоторых языков программирования, например, Python, R или Java. На рисунке 1 представлен пример запуска Mystem как консольной программы в терминале.

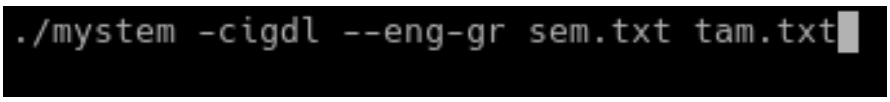


Рис.1

Выбор опций для использования определяет формат вывода и наличие той или иной информации о леммах. Так для морфологического аннотирования корпуса RuTuBiC используется формат вывода, предусматривающий отображение каждого предложения на отдельной строке и присваивание грамматической информации каждой лемме без отображения в файле вывода исходных словоформ. После получения результата морфологического анализа необходимо составление пользовательского словаря, содержащего лексические единицы, не входящие в словарь А.А. Зализняка. Так как при создании транскрипта передаются отклонения от литературной нормы письменного русского языка, в пользовательский словарь, как правило, попадают токены, передающие слова русского языка в искажённой форме или имеющие тюркское происхождение. В первую группу могут попасть практически любые лексические единицы в зависимости от особенностей произношения конкретного респондента и темпа речи. В данном случае алгоритм Mystem может построить гипотезу относительно частеречной принадлежности и морфологических характеристик слова. При обнаружении ошибочной гипотезы, слово вносится в пользовательский словарь (рисунок 2).

1 URL: https://yandex.ru/dev/mystem/doc/index-docpage/

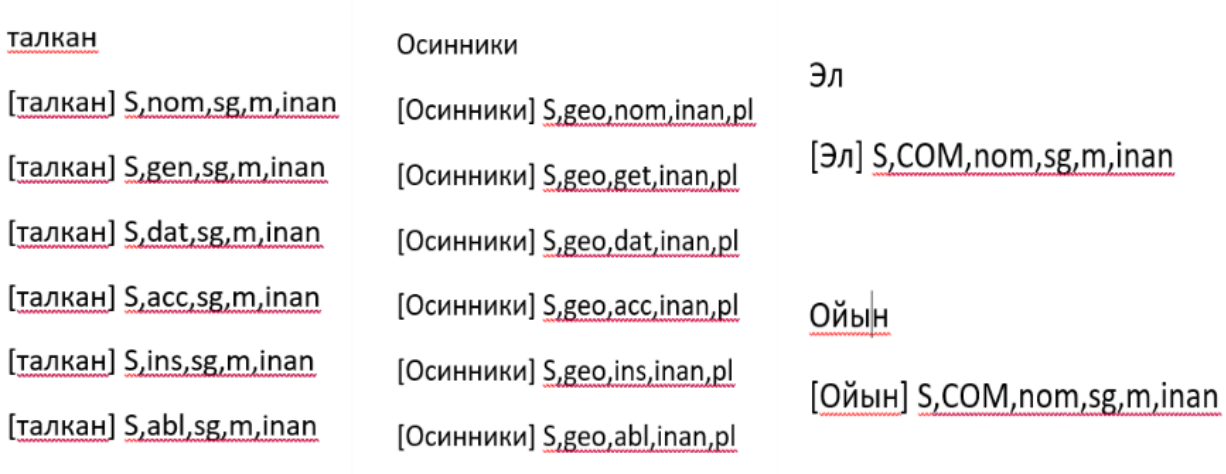


Рис.2

На рисунке выше представлены наиболее часто встречающиеся примеры групп лексических единиц, включаемых в пользовательский словарь : слова тюркского происхождения, слова, при анализе которых некорректно сработал модуль снятия контекстной омонимии, а также некоторые слова с дефисом. Коррекция падежной и частеречной омонимии осуществляется вручную.

Результаты морфологического анализа текста могут быть использованы не только для собственно уровня морфологического аннотирования, но и для уточнения маркирования некоторых отклонений от речевого стандарта.

Так для маркирования речевых сбоев, выраженных дискурсивными маркерами, необходим чёткий алгоритм выделения языковой единицы с нужным синтагматическим значением. Согласно определению из словаря лингвистических терминов, синтагматика — это исследование языка, заключающееся в последовательном разделении текста на все менее протяженные соположенные единицы, которые сосуществуют, сочетаются между собой, но отличаются одна от другой; включение их в ряды «по горизонтали» (слово в пределах высказывания, морфема в пределах слова, звук в пределах звукосочетания)[1]. Из этого определения ясно, что для поиска

1Розенталь Д. Э., Теленкова М. А. Словарь-справочник лингвистических терминов. — Изд. 2-е. —:Просвещение, 1976.

нужного синтагматического значения слова нужно исследовать его окружение в каждом отдельном высказывании. Для изучения совокупности вариантов окружения и сочетаний единицы языка в семантике применяют дистрибутивный метод.

Для поиска закономерности в употреблении того или иного дискурсивного маркера анализ его окружения начинается с предиката — главного компонента семантической модели предложения. В качестве базы семантических моделей был использован «Экспериментальный синтаксический словарь» под редакцией Л. Г. Бабенко[1]. Словарь содержит алфавитный указатель глаголов русского языка с указанием номера лексико-семантической группы, номера лексического варианта модели, совмещенной семантической модели или образной семантической модели. Каждая словарная статья содержит информацию о семантическом типе предиката, типовой семантике ситуации и основных предикатах, передающих идею типовой ситуации в общем виде. Грамматическая информация для анализа конкретного предложения получена с помощью автоматического морфологического анализатора Mystem. В Таблице 1 ниже представлена выборка предложений, иллюстрирующих употребление наречия «там» в качестве самостоятельного члена предложения и в качестве дискурсивного маркера, несущего оттенок значения неопределённости.

ЛВМ — лексический вариант модели. МСС — минимальная структурная схема.

1 Бабенко Л.Г. Русские глагольные предложения: экспериментальный синтаксический словарь.Под общ. ред. Л.Г. Бабенко. М.: Флинта: Наука, 2002.

Пример	**ЛВМ**	**МСС**	**Mystem**
Вот и **там были** тоже.	Предложения, отображающие ситуацию бытия-существования в определенном времени и пространстве. Живое существо находится где-либо.	***N1Vf*;** односоставное определенно-личное предложение.	{вот=PART=} {и=CONJ=} **{там=ADVPRO=}** **{быть=V,нп=прош,мн,изъяв,несов}** {тоже=PART=}.
там могут быть от пяти до двенадцати струн...	Несколько субъектов находятся где-л., помещаясь, располагаясь каким-л. образом.	***Inf*;** односоставное инфинитивное предложение	**{там=ADVPRO=}** **{мочь=V,несов,нп=непрош,мн,изъяв,3-л}** **{быть=V,нп=инф,несов}** {от=PR=} {пять=NUM=(пр|дат|род)} {до=PR=} {двенадцать=NUM=(пр|дат|род)} {струна=S,жен,неод=род,мн}...
мы там не только выступали с инструментами и с песнями, но ещё и танцевали с девочками, да…	Человек, группа лиц, творческий коллектив исполняет музыкальное произведение на музыкальном инструменте	***N1Vf*;** двусоставное предложение с номинативным подлежащим и глагольным сказуемым.	{мы=SPRO,мн,1-л=им} **{там=ADVPRO=}** {не=PART=} {только=PART=} **{выступать=V,нп=прош,мн,изъяв,несов}** {с=PR=} {инструмент=S,муж,неод=твор,мн} {и=CONJ=} {с=PR=} {песня=S,жен,неод=твор,мн}, {но=CONJ=} {еще=ADV=} {и=CONJ=} {танцевать=V,несов,пе=прош,мн,изъяв} {с=PR=} {девочка=S,жен,од=твор,мн}, {да=PART=}…

Пример	**ЛВМ**	**МСС**	**Mystem**
у нас в деревне была школа, **там** до четырёх, ну четыре года, **обучали** только в начальной школе.	Человек (группа людей, организация) дает кому-л. знания, умения, навыки для осуществления какой-л. деятельности.	***N1Vf*;** односоставное определенно-личное предложение	{у=PR=} {мы=SPRO,мн,1-л=(пр|вин|род)} {в=PR=} {деревня=S,жен,неод=(пр,ед|дат,ед)} {быть=V,нп=прош,ед,изъяв,жен,несов} {школа=S,жен,неод=им,ед} **{там=ADVPRO=}** {до=PR=} {четыре=NUM=(пр|род|вин,од)}, {ну=PART=} {четыре=NUM=(им|вин,неод)} {год=S,муж,неод=(вин,мн|род,ед|им,мн)}, **{обучать=V,пе=прош,мн,изъяв,несов}** {только=PART=} {в=PR=} {начальный=A=(пр,ед,полн,жен|дат,ед,полн,жен|род,ед,полн,жен|твор,ед,полн,жен)} {школа=S,жен,неод=(пр,ед|дат,ед)}.
там соотношение хакасов и русских было очень значительным	Предложения, отображающие ситуацию бытия-существования в определенном времени и пространстве. Неодушевленный предмет находится (размещается) где-л. каким-л.	***N1Vf;*** двусоставное предложение с номинативным подлежащим и глагольным сказуемым.	**{там=ADVPRO=}** **{соотношение=S,сред,неод=(вин,ед|им,ед)}** {хакас=S,муж,од=(вин,мн|род,мн)} {и=CONJ=} {русский=A=(пр,мн,полн|вин,мн,полн,од|род,мн,полн)} **{быть=V,нп=прош,ед,изъяв,сред,несов}** {очень=ADV=} {значительный=A=(дат,мн,полн|твор,ед,полн,муж|твор,ед,полн,сред)}

Пример	ЛВМ	МСС	Mystem
	образом.		
У нас было парочка танцев таких, как модерн, **там**, такие, но в основном мы танцевали народные танцы.			{у=PR=} {мы=SPRO,pl,1p=(abl|acc|gen)} {быть=V,intr=praet,sg,indic,n,ipf} {парочка=S,f,inan=nom,sg} {танец=S,m,inan=gen,pl} {такой=APRO=(abl,pl|gen,pl| acc,pl,anim)}, {как=CONJ=} {модерн=S,m,inan=(acc,sg| nom,sg)}, **{там=PART=}**, {такой=APRO=(nom,pl|acc,pl,inan)}, {но=CONJ=} {в=PR=} {основное=S,sg,n,inan=abl} {мы=SPRO,pl,1p=nom} {танцевать=V,ipf,tran=praet,pl,indic} {народный=A=(acc,pl,plen,inan| nom,pl,plen)} {танец=S,m,inan=(acc,pl|nom,pl)}.

В предложениях 1-5 «там» является наречием места, обозначающим признак действия. В колонке «Mystem» «там» в этих предложениях обозначено как местоименное наречие. Однако в предложении 6 «там» не связано с предикатом и является частицей.

В приведённом фрагменте исследования с использованием грамматической информации, полученной с помощью Mystem, предикат взят за основу, так как значимые единицы языка распределены в речи по определённым прогнозируемым закономерностям. В данном случае применение дистрибутивного метода позволяет спрогнозировать вероятность появления в речи одного элемента, основываясь на информации о другом.

2.2. Аннотирование мультимодального файла

Мультимодальные данные являются альтернативой текстовым данным в чистом виде. В лингвистике и других гуманитарных науках мультимодальные данные включают в себя оцифрованный аудио- или видеосигнал. Отличительная особенность мультимодальных данных — привязка к определённому временному промежутку. Такие данные являются основным материалом в исследованиях устной речи, невербальной коммуникации и жестового языка.

В основе мультимодального аннотирования лежит идея о том, что временной поток можно разделить на сегменты, каждый из которых имеет начальную и конечную точку[1]. Несмотря на то, что в настоящее время существует ряд технических средств для автоматической обработки аудио- и видеосигнала, в большинстве случаев аннотирование мультимодального корпуса осуществляется полностью или частично вручную, так как инструменты автоматической обработки цифрового сигнала не дают точности результатов, необходимой для представления данных в корпусе и последующего осуществления исследований.

Аннотирование мультимодального корпуса требует соблюдение ряда условий при работе с текстовыми данными, так как текстовое представление корпуса должно быть синхронизировано с аудио- сигналом. Выбор программного обеспечения для расшифровки аудиозаписи и её синхронизации с текстом зависит перечня исследовательских задач, для решения которых создаётся корпус. Однако независимо от ограничений, связанных с типом данных, и выбором программного обеспечения для работы, аннотирование мультимодального файла имеет чёткую последовательность.

1 Ide N., Pustejovsky J. Handbook of Linguistic Annotation / Springer 2017.

Первым этапом работы является создание транскрипта — текстовой версии аудио- или видеозаписи. Говорящие при этом могут быть маркированы шифром или инициалами. Ниже представлен фрагмент такого транскрипта одной из аудиозаписей, входящей в корпус RuTuBiC, иллюстрирующий маркирование участников диалога и минимальную разметку. В примере ниже участники диалога маркированы в формате ФИО, ремарки неречевых действий с двух сторон отмечены знаком решётки, а наложение реплик — круглыми скобками.

БАС@ Ну они готовились на филологический целенаправленно.

КАВ@ #смеётся# Наверное, я просто, я всегда считала себя таким достаточно начитанным человеком, что ну по сравнению со своими одноклассниками я считала себя очень умной #смеётся#. Когда я приехала сюда, я как-то подумала, что я такая, как все, и ничего такого в моих знаниях нет, что мне ещё вот читать и читать, и читать. Вот. Ещё мне нравится Стивен Кинг. Тоже люблю фа… (БАС@ Мм, фантастика) да, такую литературу.

Как правило, подобный транскрипт создаётся в текстовом процессоре, из которых чаще всего используют Microsoft Word или LibreOffice Writer. Подобные программы имеют удобный функционал для маркирования тегов разметки цветом и внесения комментариев, что часто является важной частью работы в команде. Недостатком использования текстовых процессоров является необходимость копирования текста или конвертации файла в формат, подходящий для работы с морфологическим анализатором. Альтернативой текстовым процессорам являются текстовые редакторы. Для работы над данным проектом были использованы текстовые редакторы Sublime Text и Notepad++. Подобные программы не предоставляют такого широкого функционала для форматирования текста, как текстовые процессоры, однако являются более оптимальным решением с точки зрения

использования на следующих этапах работы, так как предоставляют возможность сохранения текста в формате и кодировке, необходимых для корректного осуществления процедуры автоматического морфологического анализа.

В некоторых случаях в транскрипт включают временной код или тайм-код — цифровые данные о времени в специальном формате, записываемый вместе со звуком или изображением. При расшифровке данных корпуса наличие тайм-кода позволяет создателю транскрипта вернуться к нужному моменту аудио- или видеозаписи для уточнения или коррекции текста. Также именно наличие таймкодов в дальнейшем определяет возможность синхронизированного представления текстовых и аудио- или видеоматериалов корпуса. Как правило, минимальным сегментом при создании таймкодов является реплика, однако при работе над корпусом на этапе создания транскрипта с таймкодами необходимо представлять не только перечень исследовательских задач, для решения которых создаётся данный корпус, но и перечень тех технических этапов работы, которые будут выполняться в дальнейшем на основе созданного транскрипта. Для корпуса RuTuBiC такими задачами являются автоматическая морфологическая разметка, разметка отклонений от речевого стандарта, а также метаразметка.

При разработке траектории работы над корпусом было апробировано три варианта последовательности работы. Первый вариант предполагает создание транскрипта в текстовом процессоре или текстовом редакторе, таймкоды же в данном случае создаются отдельно с использованием специального программного обеспечения. Второй же вариант предполагает одновременное создание текста и таймкодов с помощью специального программного обеспечения, предназначенного для создания субтитров к аудио- или видеофайлам. В обоих случаях таймкоды были созданы с помощью программы Aegisub.

Третий вариант требует дополнительной подготовки, но позволяет на выходе получить готовый шаблон мультимодального файла в формате, пригодном для следующих этапов аннотирования. В данном случае необходимо в первую очередь создать иерархическую структуру слоёв разметки, один из которых будет использован для записи временных меток, соответствующих границами минимальных сегментов — реплик, предложений, словосочетаний или слов. Для работы с корпусом RuTuBiC была использована программа ELAN[1], апробированная при разметке мультимодальных файлов для многих проектов, связанных в том числе с документированием языков и психолингвистическими исследованиями[2]. Несмотря на то, что ELAN предоставляет функционал, достаточный для большинства проектов по созданию мультимодальных корпусов с аннотированием разного уровня сложности, необходимо учитывать особенности звучащей речи респондентов, являющейся основным материалом корпуса. Так как материалом корпуса являются аудиозаписи спонтанной речи, такие звуковые файлы не вполне подходят для применения методов автоматической сегментации звучащей речи или автоматической сихронизации аудио- и текстовых материалов из-за наличия фоновых шумов, наложения реплик и наличия речевых сбоев в речи респондента и интервьюера. Так как сегменты, полученные на первом этапе работы, должны

1 URL: https://archive.mpi.nl/tla/elan

2 URL: http://nflrc.hawaii.edu/ldc/.

Aguera P. et al. ELAN: A Software Package for Analysis and Visualization of MEG, EEG, and LFP Signals.

Blokland R. et al. Language Documentation meets Language Technology.

Brugman H., Russel A. Annotating Multi-media / Multi-modal resources with ELAN.

Partanen N. et al. Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region.

Soldner F., Perez-Rosas V., Mihalcea R. Box of Lies: Multimodal Deception Detection in Dialogues.

Wittenburg P. et al. ELAN: a Professional Framework for Multimodality Research.

Литвиненко О.А., Николаева Ю.В., Аннотирование русских мануальных жестов: теоретические и практические вопросы, 2017.

Шерстинова Т.Ю. Лингвистические мультимедийные архивы и национальный Фонд звучащей речи «Голоса народов России».

иметь максимально точные границы, таймкоды, полученные в результате работы встроенного в ELAN модуля автоматической сегментации звучащей речи не могут быть использованы для создания аннотаций. В связи с описанными особенностями звуковых данных, было принято решение осуществлять сегментацию вручную. После создания сегментов во встроенном редакторе можно осуществлять транскрибирование, по завершении которого можно экспортировать полученный текст в формате и кодировке, подходящих для автоматического морфологического анализа.

С позиций выбора оптимальной последовательности работы и инструментов на этапе создания транскрипта звучащей речи, наиболее подходящим кажется третий вариант, позволяющий благодаря функционалу используемого программного обеспечения получить представление о структуре мультимодального файла и связи разных уровней разметки. Однако опыт работы над корпусом показал, что выбор оптимальной последовательности и инструментов для работы над каждой задачей зависит в том числе от навыков, которыми владеет или готов освоить человек, работающий над её выполнением, так как к работе каждом этапе часто привлекают специалистов из разных областей, а сами этапы работы часто удалены друг от друга во времени.

Независимо от того, какая последовательность работы была выбрана на этапе транскрибирования, следующий важный этап создания мультимодального корпуса — создание файла, в котором разные уровни аннотирования должны быть иерархически связаны. Существует ряд программ, предназначенных для создания мультимодальных файлов, их аннотирования и анализа.

Для работы над корпусом RuTuBiC была выбрана программа ELAN, разработанная Институтом психолингвистики Общества Макса Планка в Неймегене.[1] Выбор программного обеспечения для работы над корпусом

1 URL: https://archive.mpi.nl/tla/elan

мотивирован особенностями материалов корпуса и этапов работы над ним. В литературе, аккумулирующей лучшие практики в области корпусной лингвистики и аннотирования корпусов разных типов[1], авторы упоминают ряд программ, используемых для задач мультимодального аннотирования. Анализ документации к перечисленному программному обеспечению показал, что несмотря на кажущуюся взаимозаменяемость каждая из программ подходит для выполнения ограниченного определённых задач. Можно выделить две группы программ: программы предназначенные для транскрибирования, синхронизации и разметки данных, а так же программы, предназначенные для анализа данных. К первой группе относятся такие программы как ANVIL[2], ChronoViz[3], ELAN[4], EXMARaLDA Partitur Editor[5]. ANVIL и ChronoViz были разработаны специально для работы с видеофайлами, в то время как ELAN и EXMARaLDA Partitur Editor имеют схожий функционал, предназначенный для работы как с аудио-, так и видеофайлами и позволяющий создавать многоуровневую структуру связанных аннотаций. ELAN и EXMARaLDA Partitur Editor были использованы в ряде проектов по созданию мультимодальных корпусов[6]. Несмотря на схожий функционал, выбор программного обеспечения для работы был обусловлен в том числе тем, что большинство мультимодальных корпусов, созданных в рамках российских проектов документирования языка[7] и социолингвистических исследований[8], были созданы с использованием программы ELAN. Ко второй группе программ относится Praat[9] —

1Adolphs S., Knight D. Building a spoken corpus: What are the basics? // The Routledge Handbook of Corpus Linguistics / ed. by Anne O'Keeffe and Michael McCarthy, 2010.

2 URL: https://www.anvil-software.org/.

3 URL: https://chronoviz.com/.

4 URL: https://archive.mpi.nl/tla/elan.

5 URL: https://exmaralda.org/en/partitur-editor-en/.

6 URL: https://exmaralda.org/en/projects/.

7 URL: http://web-corpora.net/tsakorpus_russian_nonst/corpus.html.

8 URL: http://www.ord-corpus.spbu.ru/SocialStudies/p_00_001.html.

9 URL: https://www.fon.hum.uva.nl/praat/.

программное обеспечение, предназначенное для осуществленя подробного анализа речевых данных. Функционал данного программного пакета не предполагает создание мультимодального файла с иерархической структурой, однако в дальнейшем он может быть использован для уточнения использования некоторых тегов фонетических отклонений, в том числе обусловленных межъязыковой интерференцией.

Основным элементом мультимодального файла, характеристики которого определяют структуру разметки и поисковые возможности корпуса, является слой. Слой представляет собой группу аннотаций одного типа, имеющих одинаковые характеристики и определённым образом связанных с аннотациями других слоёв в иерархической структуре разметки. Выбор программы ELAN для осуществления мультимодального аннотирования обусловлен в том числе и тем, что данное программное обеспечение предоставляет более широкий по сравнению с аналогичными программами функционал, с помощью которого можно задать характеристики каждого слоя. Основной параметр, определяющий тип связи между аннотациями одного слоя и между разными слоями — Linguistic Type Stereotype. Для слоёв корпуса RuTuBiC были использованы стереотипы None и IncludedIn. Стереотип None является характеристикой родительского слоя, не имеющего никаких иерархических и структурных ограничений кроме запрета на наложение сегментов внутри слоя. Стереотип IncludedIn отвечает за создание дочерних слоёв, границы сегментов в которых связаны с границами сегментов родительских слоёв, также слой с таким стереотипом может содержать промежутки между сегментами. С более подробной характеристикой программных опций, отвечающих за настройку данного параметра, можно ознакомиться в Приложении. В корпусе RuTuBiC на данный момент существуют следующие слои (см. Таблица 2).

Название слоя	**Тип аннотаций**	**Родительский или дочерний по отношению к соседним слоям**	**Linguistic Type Stereotype**
text	текст	parent (transcript_resp)	None
transcriprt_resp	текст	referring (text) /parent (word_morph, mistake, Discourse, CT, Theme)	IncludedIn
word_morph	текст	referring (transcript_resp)	IncludedIn
mistake	теги	referring (transcript_resp)/parent (word_correct)	IncludedIn
word_correct	текст	referring (mistake)	IncludedIn
Discourse	теги	referring (transcript_resp)	IncludedIn
CT	теги	referring (transcript_resp)	IncludedIn
Theme	теги	referring (transcript_resp)	IncludedIn
Genre	теги	referring (transcript_resp)	IncludedIn

Таблица 2

2.3. Мультимодальное аннотирование корпуса RuTuBiC

Аннотирование аудиозаписей вручную - трудоёмкий процесс, требующий количества времени, превышающего длительность аудиозаписи.

ELAN – многофункциональное программное обеспечение для мультимодального аннотирования, разработанный Институтом психолингвистики Общества Макса Планка в Неймегене[1].

ELAN используют в проектах по документированию исчезающих языков, транскрибированию жестового языка и исследованию мультимодальной коммуникации, однако функциональные возможност,и

1 URL: https://archive.mpi.nl/tla/elan

программного обеспечения позволяют использовать его вне перечисленных направлений.

В процессе работы генерируется файл в формате EAF - XML формат, представляющий собой модель данных, отражающую связи отдельных элементов размеченного файла. Аннотации, синхронизированные с медиафайлом, представлены в виде иерархии слоёв, содержащих данные разного типа.

Корпус содержит девять слоёв, из которых два содержат текстовые данные, один - теги морфологической разметки, один - теги отклонений от речевого стандарта, один - коррекцию отклонений в текстовом формате и четыре - теги метаразметки, описывающие тип речи, дискурс, жанр и тему текста.

Иерархия слоёв (рисунок 2) в данном случае нужна не только для организации данных в корпусе, но и для того, чтобы ускорить процесс создания аннотаций и получить более точные результаты.

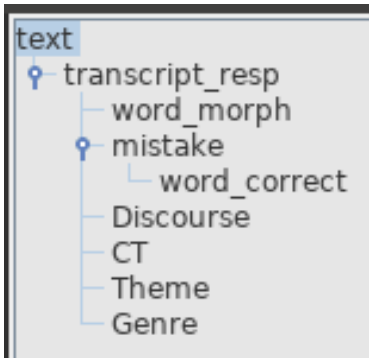


Рис.2 Иерархия слоёв разметки

Первый слой содержит текстовые данные, синхронизированные с аудиофайлом и является родительским по отношению ко всем остальным. В текстовой версии интервью представлена как речь респондента, так и речь интервьюера (рисунок 3).

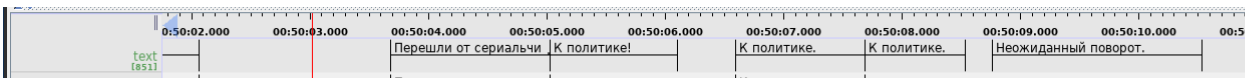


Рис.3

Включение в корпус текстовой записи речи интервьюера обусловлено тем, что метаразметка описывает акт коммуникации в целом. Аннотации этого слоя разметки создаются с помощью сегментирования аудиофайла с последующим заполнением этих сегментов текстовыми данными. Несмотря на существование готовых технических решений для автоматической сегментации звучащей речи для синхронизированного представления аудио и текста, результаты, полученные с их помощью, требуют ручной коррекции из-за наличия фоновых шумов и речевых сбоев в речи респондента и интервьюера. В связи с этим сегментация аудиофайла производится вручную в программе ELAN (рисунок 4).

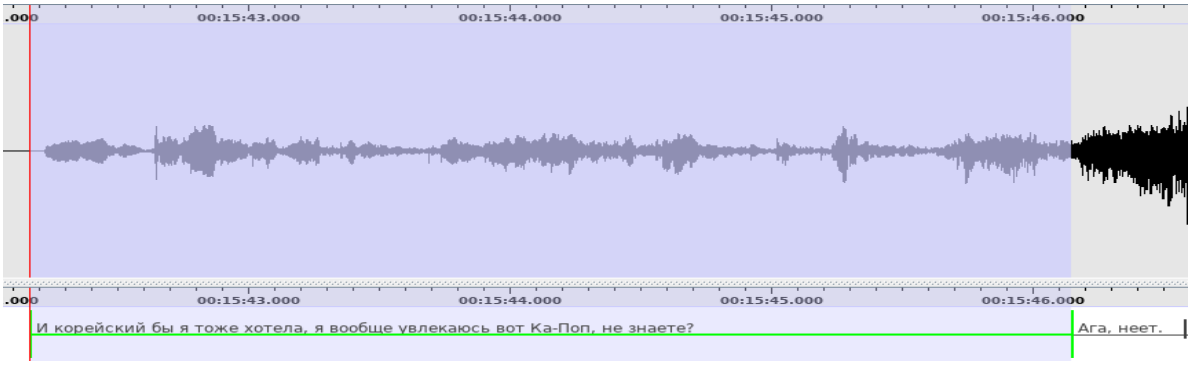


Рис.4

Минимальной единицей для данного слоя является предложение, поэтому каждый созданный сегмент соответствует предложению. В созданные сегменты в режиме транскрибирования вносят текстовые данные.

Второй слой разметки содержит текстовое представление речи респондента (рисунок 5). Этот слой является зависимым по отношению к первому. Аннотации этого слоя создаются путём копирования из родительского слоя, поэтому повторная сегментация аудио для этого и остальных слоёв, минимальной единице для которых является предложение, не требуется. Аннотации, содержащие речь интервьюера удаляются вручную, так как количество сегментов в репликах респондента и интервьюера не ограничено.

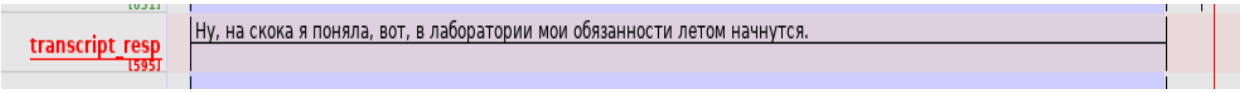


Рис.5

После внесения текста в аннотации второго слоя, можно экспоритировать речь респондента в формате, подходящем для работы с Mystem для осуществления автоматического морфологического анализа текста.

Слой с морфологической разметкой зависит от второго слоя, так как как морфологический разбор осуществляется только для речи респондента. Аннотации этого слоя содержат имплементированные вручную леммы и теги частеречной принадлежности и морфологических категорий[1], которые алгоритм автоматического морфологического анализатора Mystem присваивает каждой исходной словоформе (рисунок 6).

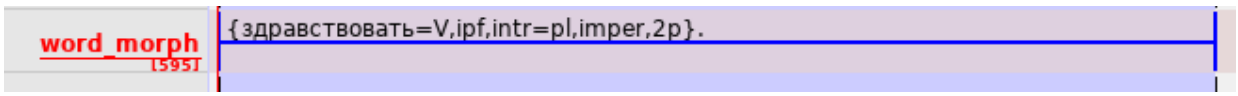


Рис.6

В процессе автоматического морфологического анализа можно выделить некоторые случаи **отклонений от речевого стандарта**. Теги отклонений (Приложение 1) заранее вносят в списки возможных значений аннотаций для того, чтобы в процессе разметки выбор тега из выпадающего списка минимизировал риск отступления от принятых в корпусе обозначений. Для создания аннотаций в этом слое необходимо повторное выборочное сегментирование аудиофайла, так как в зависимости от типа отклонения минимальной единицей может являться слово, словосочетание или предложение (рисунок 7).

1 URL: https://yandex.ru/dev/mystem/doc/grammemes-values-docpage/

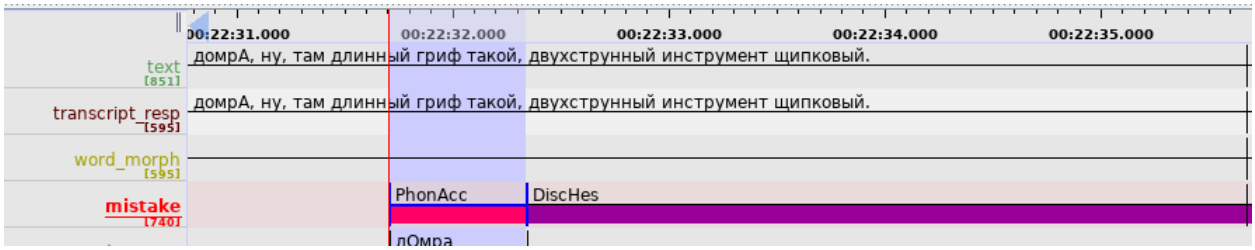


Рис.7

Наличие соответствующей аннотации в слое **формы коррекции** предполагается не для каждой аннотации слоя, содержащего маркирование отклонений от речевого стандарта. В процессе работы над системой маркирования отклонений от речевого стандарта были выделены следующие группы тегов: фонетические, морфологические, деривационные, синтаксические, лексические, а также теги, маркирующие речевые сбои и переключение кодов. На материале аннотаций размеченного мультимодального файла рассмотрим некоторые примеры соотнесения аннотаций в слоях маркирования отклонений от речевого стандарта и формы коррекции.

Маркирование фонетических отклонений от речевого стандарта включает в себя теги нарушения норм ударения и все остальные отклонения от норм литературного произношения, свойственные спонтанной устной речи. Оба типа фонетических отклонений предполагают наличие формы коррекции, так как существуют вполне однозначные варианты нормативного произношения, зафиксированные в орфоэпических словарях (рисунки 8 и 9).

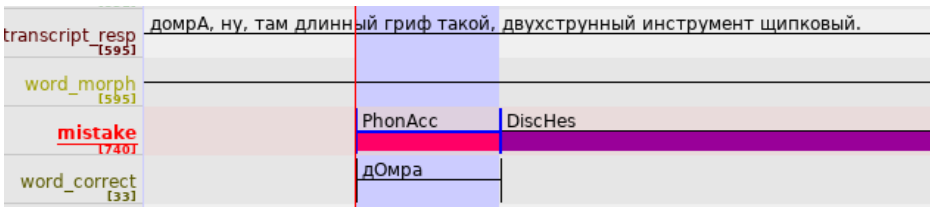


Рис.8

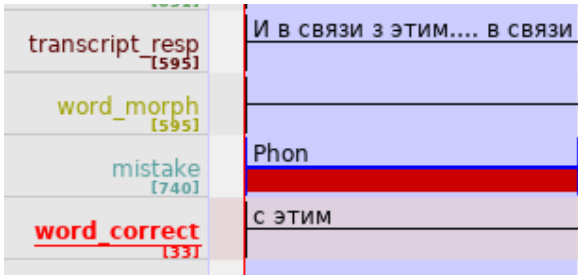
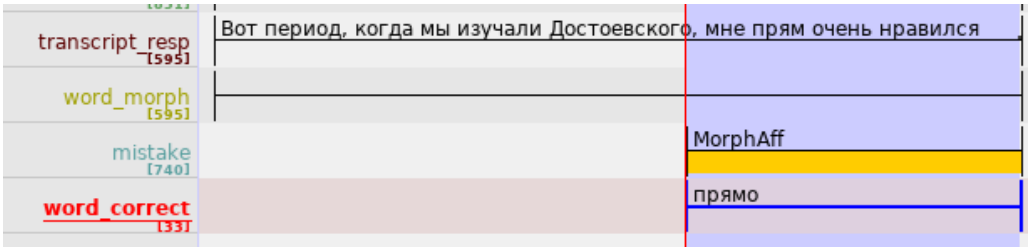


Рис.9

Морфологические теги также в большинстве случает имеют аннотацию, содержащую форму коррекции (рисунок 10), так как в данном случае речь идёт о нарушении определённых грамматических норм.

Рис.10



Нарушения норм деривации представлены только одним тегом, которому в большинстве случаев, соответствует аннотация с формой коррекции (рисунок 11).

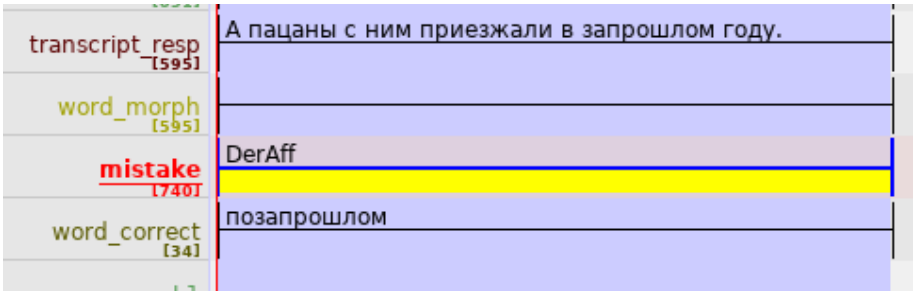


Рис.11

Теги, маркирующие синтаксические отклонения от речевого стандарта, являются наиболее подробно представленной группой тегов.

В случаях, когда указание коррекции необходимо, соответствующие аннотации копируют с слой, содержащий коррекцию ошибок. Данный слой является зависимым по отношению к слою с тегами отклонений от речевого стандарта и содержит текстовые данные.

Метаразметка

Аннотации четырёх слоёв метаразметки содержат соответствующие теги (Приложение 2), внесённые в списки значений аннотации. Аннотации слоёв, содержащих информацию о типе речи, дискурсе, речевом жанре и теме

текста, являются зависимыми по отношению к слою с речью респондента и создаются путём копирования готовых пустых сегментов из второго слоя в шестой, седьмой, восьмой и девятый слои соответственно (рисунок 12).

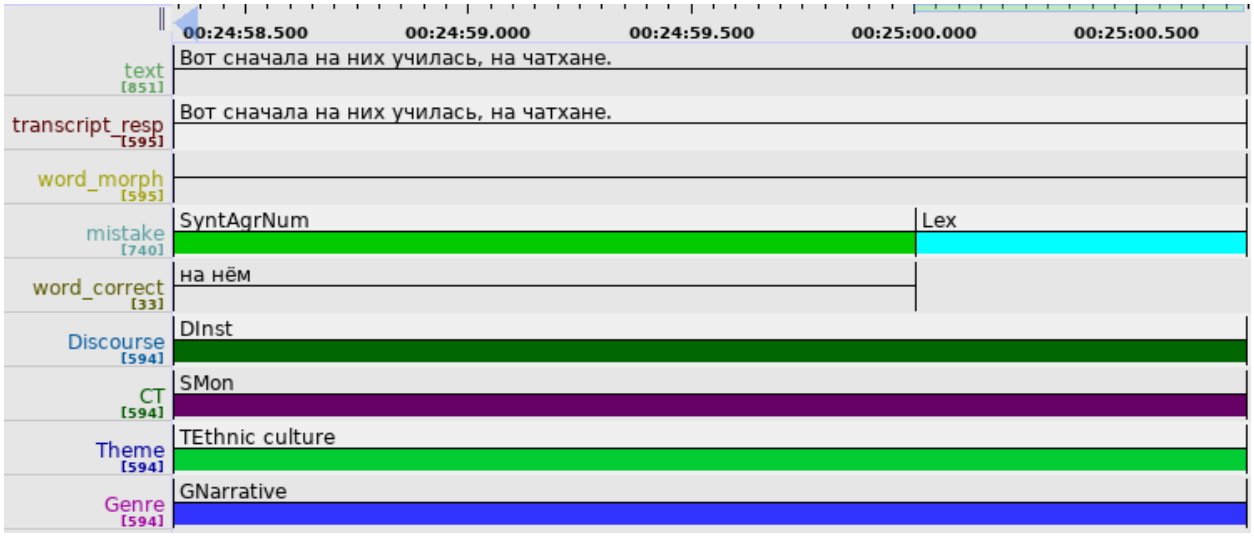


Рис.12

Собрание файлов в формате EAF ещё не является корпусом, однако функциональные возможности программы ELAN позволяют осуществлять поиск по заданным параметрам и получать статистику аннотаций (рисунок 13, рисунок 14)

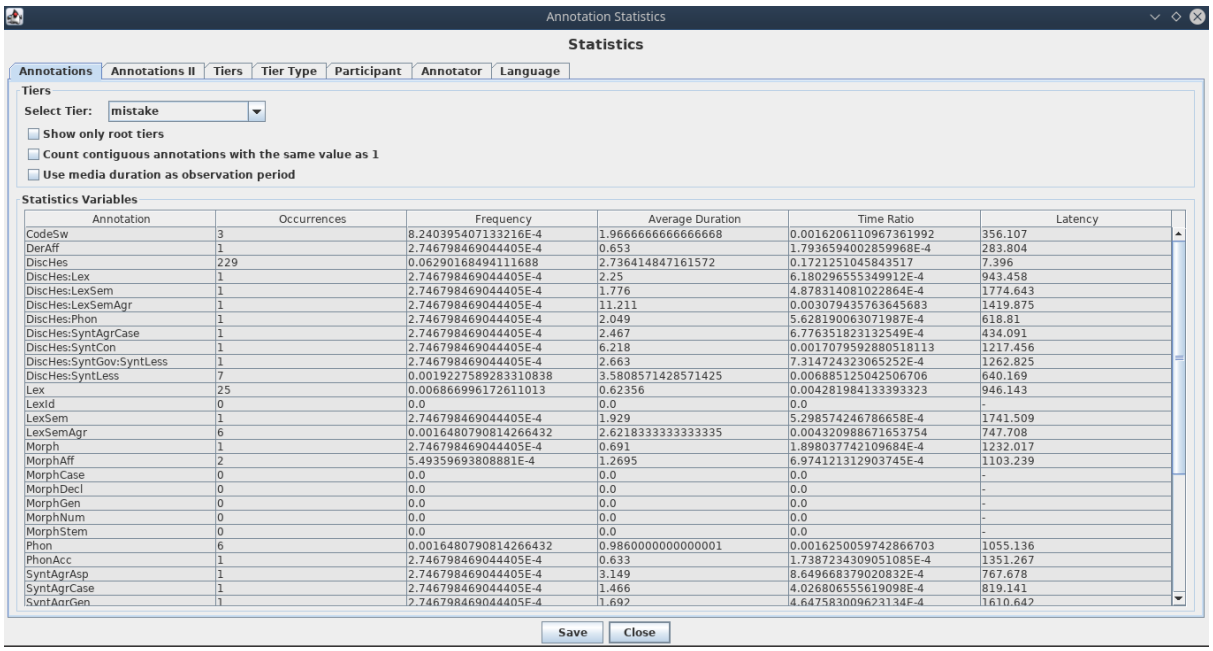


Рис.13

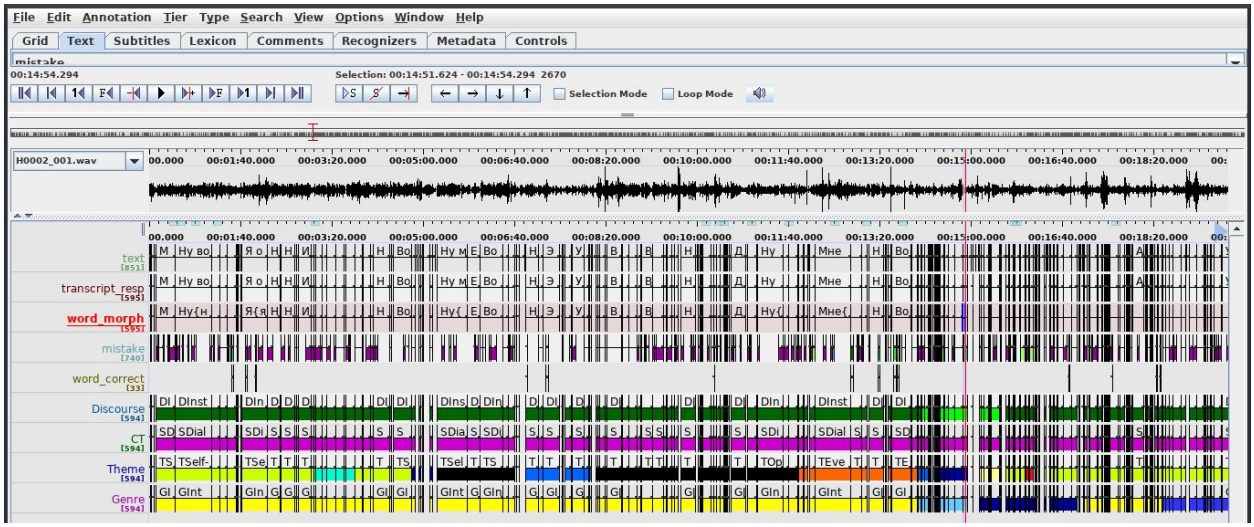


Рис.14

Выводы по 2 главе

В данной главе были описаны практические аспекты многоуровневого аннотирования бимодального корпуса. Выбор вида аннотирования в корпусе зависит от цели его создания и типов данных, составляющих его. К видам разметки, являющимися необходимым для функционирования большинства крупных корпусов. В число таких видов разметки входит морфологическая, для получения которой необходимо осуществление морфологического анализа, результаты которого могут быть использованы не только для собственно уровня морфологического аннотирования, но и при маркировании некоторых отклонений от речевого стандарта.

Особенности мультимодальных файлов определяют технические требования требования к данным и применяемое программное обеспечение.

Заключение

Данное исследование показало, что создание корпуса текстов действительно требует объединения лингвистической теории и информационных технологий, позволяющие решить задачи, связанные с обработкой большого количества данных.

Описанные информационные технологии являются простыми в реализации, что понижает порог вхождения в процесс создания корпуса. Использование автоматического морфологического анализатора позволяет отказаться от приписывания словам частеречных и морфологических характеристик вручную и избежать расхождения в употреблении тегов в рамках одного корпуса. Мультимодальные файлы, созданные в программе ELAN дают возможность составить наглядное представление о структуре корпуса и его поисковых возможностях, а также реализовать его главный дифференциальный параметр – бимодальность.

На данный момент одной из сложностей является ручное составление пользовательских словарей для коррекции результатов работы автоматического морфологоического анализатора Mystem, так как стандрарт, заложенный там, ориентирован на нормы письменного литературного языка и не отражает всех особенностей устной речи. С особенностями беглой спонтанной речи так же связана другая сложность – при создании мультимодального файла в программе ELAN требуется ручная сегментация аудиофайла, так как существующие инструменты автоматической сегментации звучащей речи не дают достаточно точных результатов.

Данная работа была выполнена в рамках проекта Лаборатории лингвистической антропологии НИ ТГУ «Языковое и этнокультурное разнообразие Южной Сибири в синхронии и диахронии: взаимодействие языков и культур»

ЛИТЕРАТУРА

1. Архипов А. В. Документирование малых языков: научные и технические аспекты // Языковое разнообразие в киберпространстве: российский и зарубежный опыт. - М., 2008. - С. 76-83.

2. Бабенко Л. Г. Русские глагольные предложения: экспериментальный синтаксический словарь / Под общ. ред. Л. Г. Бабенко. М.: Флинта: Наука, 2002.
3. Богданова Н. В. и др. Звуковой корпус русского языка «Один речевой день»: пути пополнения и первые результаты исследования // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). М.: Издательство РГГУ, 2010.

4. Большакова Е. И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. Пособие / Большакова Е. И. и др. — М.: МИЭМ, 2011.
5. Гришина Е. А. Два новых проекта для Национального корпуса: мультимедийный подкорпус и подкорпус названий // Национальный корпус русского языка: 2003—2005. М.: Индрик, 2005, 233—250.

6. Гришина Е. А. Устная речь в Национальном корпусе русского языка // Национальный корпус русского языка: 2003—2005. М.: Индрик, 2005, 94—110.

7. Гришина Е. А. О маркерах разговорной речи (предварительное исследование подкорпуса кино в Национальном корпусе русского языка) // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог'2007» (Бекасово, 30 мая — 3 июня 2007 г.), 147—156

8. Гришина Е. А. Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 175—214.

9. Гришина Е. А., Савчук С. О. Корпус устных текстов в НКРЯ: состав и структура // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 129—149.

10. Ермакович М. В. Автоматическое определение границ слова в русском языке / Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2017» Москва, 31 мая — 3 июня 2017.

11. Захаров В. П. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн., / В. П. Захаров, С. Ю. Богданова. – СПб.: СПбГУ. РИО. Филологический факультет, 2013. — 148 с.

12. Зеленков Ю. Г., Сегалович И. В., Титов В. А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара «Диалог — 2005». М., 2005. С. 188–197.

13. Земская Е. А. Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис / Е. А. Земская и др. М: Наука, 1981.

14. Карасик В. И. О типах дискурса // Языковая личность: институциональный и персональный дискурс: Сб. науч. тр. Волгоград: Перемена, 2000.

15. Кибрик А. А. Рассказы о сновидениях: корпусное исследовение русского устного дискурса / Под ред. А. А. Кибрика и В. И. Подлесской, М.: Языки славянских культур, 2009. — 736 с.: ил.

16. Копотев М. В. Введение в корпусную лингвистику / М. В. Копотев – Прага, Animedia Company, 2014.

17. Литвиненко О. А. Николаева Ю. В., Аннотирование русских мануальных жестов: теоретические и практические вопросы, 2017.

18. Ляшевская О. Н. К проблеме лемматизации несловарных слов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2007», 407—412.

19. Ляшевская О. Н., Плунгян В. А., Поляков А. Е., Савчук С. О., Сичинава Д. В. Обработка текстов для Национального корпуса русского языка: технологическая цепочка. Международная конференция «Корпусная лингвистика-2004». Тезисы докладов. СПб.: СПбГУ, 54—56

20. Марчук Ю. Н. Компьютерная лингвистика: учеб. Пособие / Ю. Н. Марчук — М.: АСТ: Восток — Запад, 2007

21. Плунгян В. А. Зачем нужен национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005.

22. Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении, 2008, No. 16 (2), 7—20.

23. Рахилина Е. В. Корпус как творческий проект // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 7—26.

24. Резанова, З. И. История языкознания: XIX - первая половина XX века: Хрестоматия : учебное пособие : в 2 частях / З. И. Резанова. — 2-е изд., стер. — Москва : ФЛИНТА, [б. г.]. — Часть 1 : 2 — 2012. — 264 с.

25. Резанова З. И. Корпус устной речи русско-тюркских билингвов Южной Сибири: разметка отклонений от речевого стандарта // Вопросы лексикографии, No. 15, 2019, С. 127- 140.

26. Резанова З. И., Веснина Г. Ю. Подкорпус русской речи билингвов лингвистического корпуса «Томский региональный текст»: принципы разметки и метаразметки корпуса // Вопр. лексикографии. – 2016. – No 1 (9). – С. 29–39.

27. Резанова З. И. Подкорпус устной речи русско-тюркских билингвов Южной Сибири: типологически релевантные признаки // Вопросы лексикографии. 2017. No 11. С. 105–118.

28. Розенталь Д. Э., Теленкова М. А. Словарь-справочник лингвистических терминов. — Изд. 2-е. —:Просвещение, 1976.

29. Савчук С. О, Сичинава Д. В. Обучающий корпус русского языка и его использование в преподавательской практике // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 317—334.

30. Савчук С. О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 62—88.

31. Сичинава Д.В. Национальный корпус русского языка: очерк предыстории. 2005.

32. Труды международной конференции «Корпусная лингвистика – 2011» 27–29 июня 2011 г., Санкт-Петербург. – СПб.: СПбГУ. Филологический факультет, 2011. – 348 с.

33. Шерстинова Т. Ю. Лингвистические мультимедийные архивы и национальный Фонд звучащей речи «Голоса народов России».

34. Шерстинова Т.Ю. «Один речевой день» на временной шкале: о перспективах исследования динамических процессов на материале звукового корпуса // Филология. Востоковедение. Журналистика. Серия 9. – СПб., 2009.

35. Adolphs S., Knight D. Building a spoken corpus: What are the basics? // The Routledge Handbook of Corpus Linguistics / ed. by Anne O'Keeffe and Michael McCarthy, 2010.

36. Aguera P. et al. ELAN: A Software Package for Analysis and Visualization of MEG, EEG, and LFP Signals.

37. Biber D., Conrad S., Reppen R. Corpus Linguistics. Investigating language structure and use. Cambridge University Press, 1998.

38. Blokland R. et al. Language Documentation meets Language Technology.

39. Brugman H., Russel A. Annotating Multi-media / Multi-modal resources with ELAN.

40. Ide N., Pustejovsky J. Handbook of Linguistic Annotation / Springer 2017.

41. Jurafsky, Daniel & Martin, James. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2008.

42. Khomchenkova I. A., Pleshak P. S., Stoynova N. M. The Corpus of Contact-Influenced Russian of Northern Siberia and the Russian far East // Papers from the Annual International Conference “Dialogue”. M.: RSUH. 2019. P. 253-264.

43. Kilgarriff, A., Grefenstette, G.: 2003, Introduction to the special issue on web as corpus // Computational Linguistics. 2003. No 29 (3). P. 333-347.

44. McEnery T., Wilson A. Corpus linguistics. Edinburgh: Edinburgh University Press, 1996.

45. Partanen N. et al. Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region.

46. Rakhilina E., Vyrenkova A., Mustakimova E., Ladygina A., Smirnov I. Building a learner corpus for Russian // Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umea, 16th November 2016.

47. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine linguistics]. Prague: Animedia.

48. Sinclair, J. (1996) EAGLES. Preliminary recommendations on Corpus Typology. EAG–TCWG–CTYP/P. Version of May, 1996.

49. Soldner F., Perez-Rosas V., Mihalcea R. Box of Lies: Multimodal Deception Detection in Dialogues.

50. Wittenburg P. et al. ELAN: a Professional Framework for Multimodality Research.

51. URL: http://web-corpora.net/RLC.

52. URL: http://spokencorpora.ru/.

53. URL: https://archive.mpi.nl/tla/elan

54. URL: http://p220.ru/labs/laboratoriya-lingvisticheskoy-antropologii/.

55. URL: https://yandex.ru/dev/mystem/.

56. URL: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

57. URL: http://www.ruscorpora.ru/new/corpora-structure.html.

58. URL: https://johnsonsdictionaryonline.com/

59. URL: https://wals.info/

60. URL: http://www.helsinki.fi/slaavilaiset/ccmh/

61. URL: https://www.anvil-software.org/.

62. URL: https://chronoviz.com/.

63. URL: https://exmaralda.org/en/partitur-editor-en/.

64. URL: http://web-corpora.net/tsakorpus_russian_nonst/corpus.html.

65. URL: http://www.ord-corpus.spbu.ru/SocialStudies/p_00_001.html.

66. URL: https://www.fon.hum.uva.nl/praat/.

ПРИЛОЖЕНИЕ 1

Тип тэга	Пояснение/определение
Phon	Особенности произношения отдельных слов. Отклонения от норм реализации системных фонетических явлений.
PhonAcc	Отклонения от норм ударения.
Morph	Лексико-грамматические ошибки.
MorphStem	Морфологические варианты на стыке морфем, варианты форм слова (чаще неизменяемые части речи).
MorphAff	Отклонения от норм образования грамматических аффиксов, грамматических форм слова, выбор синонимичных формообразовательных аффиксов.
MorphDecl	Отклонения от нормы нулевого склонения.
MorphNum	Отклонения от норм категоризации по числу
MorphCase	Использование особых падежных форм, характерных для разговорной речи, например «новый звательный падеж»
MorphGen	Отклонение от норм образования категории рода
DerAff	Отклонения от норм словообразования, использование синонимичного деривационного аффикса, другого способа словообразования
SyntGov	Отклонения от норм падежного управления
SyntPrep	Отклонение от норм использования предлогов
SyntNum	Отклонение от норм использования форм числа Согласование по числу и использование формы сущ. в другом числе – разные вещи. В тюркских языках в

	обощенном значении – ед. число. В русском возможны оба варианта. Человек смертен. Люди смернты.
SyntAgrGen	Отклонения от норм согласования по роду
SyntAgrNum	Отклонения от норм согласования по числу
SyntAgrAsp	Отклонение от норм согласования по виду
SyntAgrCase	Отклонение от норм согласования по падежу
SyntCon	Нарушение норм семантико-синтаксического сочетания элементов. В том числе неверное употребление союзов.
SyntStruct	Неверное построение словосочетаний, нарушение структуры предложения
Lex	Использование диалектизмов, просторечных лексем, заимствований из тюркских языков
LexId	Идиоматические выражения, свойственные первому языку или диалектно-просторечные русского языка
LexSem	Использование общерусского слова в ином значении. В том числе просторечие
LexSemAgr	Нарушение норм лексико-семантического согласования
CodeSw	Переключение кодов, переход с одного языка на другой. В том числе при намеренном употреблении слов другого языка. В том числе морфемы и звуки.
DiscHes	Хезитация — речевое колебание, связанное со спонтанностью речи: текст рождается непосредственно в момент речи, возникает проблема выбора речевых единиц (слов и грамматических структур) и планирования предложения в целом.

	Использование различного рода маркеров дискурсивной связности, заполнителей пауз, особенностей ритмической организации речи; например, использование в качестве средств ритмической организации речи маркеров авторизации.

ПРИЛОЖЕНИЕ 2

Тег	Комментарий
Типы речи	
[SMon]	Речь одного конкретного человека, *которая может прерываться репликами окружающих его людей (является противоположностью диалогу), что не нарушает коммуникативного доминирования одного коммуниканта.* Характерная черта - коммуникативная активность одного из участников коммуникации, не рассчитанная на активную одновременную реакцию слушателя. Для монолога типичны значительные по размеру отрезки текста/речи, состоящие из структурно и содержательно связанных между собой высказываний.
[SDial]	Обмен репликами между двумя лицами, характеризуется сменой активных коммуникативных позиций двух участников коммуникации. Реплика – формально-структурная единица диалога, фрагмент дискурса одного говорящего, отграниченный речью другого участника коммуникации.
[SPoly]	Обмен репликами между тремя и более лицами, характеризуется сменой реплик активных участников диалога, и слушающего (третьего, четвертого и т.д.) участников коммуникации, функции которых могут меняться, задавая новые направления развития темы.
Дискурсы	
[DInst]	**Институциональный дискурс -** речевая практика, обусловленная в своем осуществлении особенностями конститутивных параметров коммуникативного акта (цель, тип коммуникантов, социальный психологический контекст и др. параметры), порождающая своеобразный вариант речи/текстов.
[DPers]	**В персональном дискурсе** коммуниканты выступают как личности, индивидуальности, связанные с многими социальными институтами, но в данном случае не выступающие как «рупоры» какого-либо социального института, личностные дискурсы характеризуются многотемьем, разнообразием речевых структур, обращением к бытовой тематике.
Жанры речи	
Интервью **[GInt]**	Диалогический жанр институционального дискурса, в корпусе –

	речевой жанр научного дискурса, личное общение исследователя с опрашиваемым, при котором исследователь (или его полномочный представитель) задает вопросы и фиксирует ответы в соответствии с заранее разработанным планом.
Беседа [GTalk]	Жанр диалогического типа речи, целью которого может являться обмен мнениями и сведениями, имеющими личностную направленность, выражение точки зрения, установление или поддержание личностных отношений с партнером по коммуникации. Для беседы свойственна частая смена тем, оценочность по отношению к теме, солидарность во мнениях. Такие элементарные жанры как сплетня, похвала, одобрение, комплимент, искреннее признание, шутка являются макрожанрами беседы.
Разговор [GConv]	Жанр диалогического типа речи, целью которого является информирование собеседника. Одним из основных мотивов разговора является заинтересованность говорящего получить нужную информацию. Для разговора свойственно вопросно-ответное реплицирование и ориентация на одну тему. Роль лидера, направляющего ход разговора, в данном случае играет спрашивающий.
Рассказ [GNarrative]	Жанр монологического типа речи, целью которого является информирование о событиях, фактах, которые произошли с рассказчиком или кем-либо другим. Темой рассказа могут быть любые события и факты, объединённые одной макротемой. Для рассказа свойственна целостность передаваемой информации, обеспечиваемая связностью отдельных фрагментов. Макрожанрами разговора могут являться оценка и ирония.
История [GStory]	Жанр монологического типа речи, целью которого является передача сведений о происшедших ранее событиях. История включает подведение смыслового итога, резюме, сопоставление с оценкой современных событий и фактов. Кроме того, важный прагматический фактор речи при рассказе «истории» – память. Макрожанром являются реплики собеседников. В истории проявляется тематическая фрагментарность, ассоциативные

	отступления от сюжета повествования, эллиптированные конструкции, вопросно-ответные ходы. Экспрессивность лексических элементов обусловлена культурным фоном ситуации общения, отражает спонтанность, неподготовленность повествования, поэтому в речи наблюдается обилие конкретизирующих лексем, а также вводных слов, показывающих контроль говорящего над ходом изложения и способом выражения.
Описание [GDescription]	Жанр монологической речи, целью которого является описание какого-то явления, характерные черты его протекания, условия, обстоятельства, участник. Макрожанр – краткие реплики собеседников, не прерывающие тематического единства, направленные на поддержку говорящего, уточнение аспектов описания и оценки сказанного.
Тема текста	
[TPeriod of life]	Рассказы о жизни, воспоминания о её памятных периодах и фрагментах (детство, учеба в школе) Ключевые показатели – рассказ касается событийного ряда некоторого периода жизни, сами события характеризуются обобщенно, по каким-то важным чертам.
[TEvent]	Рассказы о крупных памятных событиях: свадьба, развод, рождение детей, встречи с детьми, внуками, поездки. Рассказчик характеризует событие, особенности его протекания, участников, не сосредоточиваясь на подробной характеристики какого-либо компонента или обстоятельства события.
[TMode of life]	Рассказы о бытовых обычных делах:– охота, рыбалка, работа по хозяйству. Рассказчик характеризует не отдельное событие, не рассказывает о нем историю, но описывает это как типовые образцы протекания жизни.
[TFact]	Краткий рассказ-представление о каком-либо конкретном факте, выступающем характеристикой больших событий. Небольшая история о случае.

[THuman characteristics]	Характеристики людей (друзья, соседи, отношения с родственниками). В фокусе находится человек, его личностные особенности, типовые поступки, роль в жизни рассказчика и т. д.
[TSelf-characterization]	В фокусе находится сам рассказчик, его личностные особенности, типовые поступки.
[TFamily]	Характеристика членов семьи.
[TNatural environment]	Рассказы об окружающем мире. Природа, погода, описание деревни, городского пространства, бытовых обстоятельств жизни. Внимание рассказчика и говорящего сосредоточено на окружающих обстоятельствах жизни, ее фрагмента, возможно переключение с описания события.
[TSocial environment]	Рассказы о различных аспектах социального контекста существования информантов в разные периоды жизни, характеристика социального контекста существования людей, в среде которых в какой-то период жизни оказывается информант.
[TEthnic culture]	Рассказы о народной культуре, праздниках, обычаях, этносах, этнических языках.
[TOpinion]	Развернутое выражение мнения. Входит в жанр беседы.
[TOpinion exchange]	Фрагмент диалога, в котором собеседники выражают оценочные мнения относительно явлений и предметов.
[TComment]	Комментарий к обстоятельствам диалога – окружающее говорящих пространство, появляющиеся персоны, изменяющиеся в течение диалога обстоятельства, что-то может привлечь вимание и отрефлексировано говорящими.