МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Филологический факультет

АКТУАЛЬНЫЕ ПРОБЛЕМЫ ЛИНГВИСТИКИ И ЛИТЕРАТУРОВЕДЕНИЯ

Сборник материалов VII (XXI)
Международной научно-практической конференции
молодых учёных
(16–18 апреля 2020 г.)

Выпуск 21

Томск
Издательский Дом Томского государственного университета
2020

КОМПЬЮТЕРНАЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА

DOI: 10.17223/978-5-94621-901-3-2020-29

ОБРАБОТКА ПОЛЕВЫХ ДАННЫХ ДЛЯ ДОКУМЕНТИРОВАНИЯ ЯЗЫКОВ

Погодаева Е.Н.

Томский государственный университет, магистрант

DIGITAL PROCESSING OF FIELD DATA FOR LANGUAGES DOCUMENTATION

Pogodaeva E.N.

Tomsk State University, master student

В статье представлены технические аспекты обработки языковых данных корпуса тюркско-русских билингвов Южной Сибири. При анализе звуковых и текстовых данных корпуса были выявлены технические задачи, решение которых является важным для создания оптимальной траектории работы в проекте по документированию языков.

Ключевые слова: документирование языка, билингвизм, бимодальный корпус, MyStem, ELAN.

The article presents technical aspects of linguistic data processing. The data were aggregated when creating The Oral Speech Corpus of Russian-Turkic Bilinguals of Southern Siberia. Analyzing of audio and text data has revealed technical problems, and to solve these problems is crucial in order to create an optimal workflow for a language documentation project.

Key words: language documentation, bilingualism, bimodal corpus, MyStem, ELAN.

Научный руководитель: З.И. Резанова, д-р филол. наук, профессор ТГУ.

Документирование языков – область лингвистики, целью которой является долговременная многоцелевая фиксация данных естественного языка для дальнейшего изучения его разных аспектов. Основой для таких проектов являются тексты разных жанров, снабжённые многоуровневыми аннотациями, однако первичным материалом выступают аудио- и видеозаписи речи респондентов.

Корпус русской речи тюркско-русских билингвов (RuTuBiC) создаётся в рамках проекта «Языковое и этнокультурное разнообразие Южной

Сибири в синхронии и диахронии: взаимодействие языков и культур» [1]. 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 институциональной коммуникации [2]. Корпус содержит три подкорпуса: шорско-русский, татарско-русский, хакасско-русский.

При подготовке к созданию корпуса исследователи не всегда учитывают технические аспекты работы, так как к выполнению задач на каждом этапе нередко привлекают специалистов из разных областей, а сами этапы работы часто удалены друг от друга во времени.

Первым этапом работы по созданию корпуса является запись аудиоматериала. Анализ собранных аудиозаписей выявил большое количество случаев наложения звучания речи собеседников. При создании текстового корпуса это не является серьёзной проблемой, но при сегментации аудио для создания аннотаций невозможно разделить две реплики, звучащие одновременно, поэтому для каждого конкретного случая решение о принадлежности реплики принимается индивидуально. В инструкцию по сбору аудиоматериалов были включены следующие рекомендации: избегание помех (шум, чужие разговоры); замена подбадривающих реплик кивком головы или другим жестом; также реплики интервьюера не должны накладываться на речь респондента, следует дожидаться паузы.

Расшифровка интервью является вторым этапом создания корпуса. На данном этапе необходима стандартизация фиксирования явлений спонтанной устной речи. Расшифровка аудиозаписей осуществляется в соответствии с правилами русской орфографии с фиксацией отклонений от речевого стандарта. Речь респондента передаётся именно так, как она звучит, в том числе в случаях, противоречащих нормам письменного литературного русского языка, что соответствует принятой практике, например, в проектах «Рассказы о сновидениях» [3] или «Корпус контактно-обусловленной русской речи билингов-носителей малых языков Севера Сибири и Дальнего Востока» [4]. Для более точного отражения некоторых особенностей произношения респондентов используются знаки международного транскрипционного алфавита. Для маркирования отклонений от норм ударения используются заглавные буквы русского алфавита.

Следующий этап, определяющий поисковые возможности корпуса – аннотирование. Аннотирование осуществляется в программе ELAN, разработанной в Институте психолингвистики им. Макса Планка для многоуровневого ручного и полуавтоматического аннотирования аудио- и видеоматериалов [5]. Система аннотирования состоит из девяти уровней, содержащих текст или теги: полный текст интервью, текст речи респондента, морфологическая разметка, разметка ошибок, нормативный вариант, тип дискурса, тип речи, жанр речи, тема текста.

Морфологическая разметка, осуществляемая с помощью морфологического анализатора MyStem, требует ручной коррекции [6]. За время работы над корпусом были выявлены типичные случаи коррекции, которая может быть выполнена с помощью технических средств разного уровня, от автозамены в текстовом редакторе для отдельного файла до обработки нескольких файлов средствами языков программирования.

Аннотации, в которые вносят текст и теги разметки, создаются путём сегментирования аудио в программе ELAN. Несмотря на существование готовых технических решений для автоматической синхронизации аудио и текста, результаты, полученные с их помощью, требуют ручной коррекции из-за наличия фоновых шумов и речевых сбоев в речи респондента и интервьюера.

Анализ мировой практики создания бимодальных корпусов показал, что перечисленные этапы работы над корпусом, возникающие проблемы и средства их решения являются типовыми. Разметка звучащей речи и создание структуры корпуса требует значительного количества времени, поэтому для систематизации работы и накопления полезных практик была создана онтология предметной области создания и разметки корпуса, включающей в себя задачи, методы их решения, инструменты, технические требования к данным и типы данных.

Литература

1. Языковое и этнокультурное разнообразие Южной Сибири в синхронии и диахронии: взаимодействие языков и культур. URL: <http://p220.ru/home/projects/item/1201-lab-lang> (дата обращения: 02.04.2020).
2. Резанова З.И. Подкорпус устной речи русско-тюркских билингвов Южной Сибири: типологически релевантные признаки // Вопросы лексикографии. 2017. № 11. С. 105–118.
3. Рассказы о сновидениях и другие корпуса звучащей речи. URL: <http://spokencorpora.ru/> (дата обращения: 02.04.2020).

4. Корпус контактно-обусловленной русской речи билингвов-носителей малых языков Севера Сибири и Дальнего Востока. URL: http://web-corpora.net/tsakorpus_russian_nonst/corpus.html (дата обращения: 02.04.2020).

5. The Language Archive (TLA) of the Max Planck Institute for Psycholinguistics. URL: <https://tla.mpi.nl/> (дата обращения: 02.04.2020).

6. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Publications: Web Mining and Search, 2003. URL: <https://research.yandex.com/publications/36> (дата обращения: 02.04.2020).

DOI: 10.17223/978-5-94621-901-3-2020-30

РЕФЕРИРОВАНИЕ НОВОСТНЫХ ТЕКСТОВ ПРИ ПОМОЩИ КВАНТИТАТИВНЫХ МЕТОДОВ (TF-IDF)

Белогорская Д.В.

Томский государственный университет, студент

SUMMARIZING NEWS TEXTS USING QUANTITATIVE METHODS (TF-IDF)

Belogorskaya D.V.

Tomsk State University, student

Актуальной проблемой современного общества является стремительный рост объема информации, который именуется «информационным взрывом». Одним из способов решения данной проблемы является автоматическое реферирование текстов, осуществляемое с помощью разных методов. В статье представлены результаты первого этапа исследования, полученные на основе применения квантитативного метода TF-IDF при решении задачи реферирования новостных текстов.

Ключевые слова: автоматическое реферирование, TF-IDF, новости.

Currently, modern society is suffering from the rapid increase of information, which is called the «information explosion». One of the ways to solve this problem is automatic text summarization, which can be done using different methods. This article presents the results of the first step of the research obtained on the basis of these methods.

Keywords: automatic summarization, TF-IDF, news.

Научный руководитель: З.И. Резанова, д-р филол. наук, профессор ТГУ.

Стремительный рост объема информации именуется «информационным взрывом». Большая часть информации представлена в текстовом формате, поэтому для решения проблемы «информационного взрыва» необходимо применение методов автоматической обработки