

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81'32 : 811.112.28

Т.А. Архангельский, О.А. Созинова

Мультимедийный корпус языка идиш*

Представлен созданный авторами мультимедийный корпус языка идиш. Первая версия корпуса содержит 10 часов аудио- и видеоматериалов, синхронизированных с расшифровками. Для корпуса были созданы поисковая платформа и веб-интерфейс, использующие базу данных типа NoSQL, веб-фреймворк Django и ряд модулей на языке JavaScript. Веб-интерфейс позволяет выполнять лексико-грамматические запросы и просматривать результаты в виде расшифровки, подсветка которой синхронизирована с проигрываемым мультимедийным материалом. Описаны отличия мультимедийного корпуса идиша от аналогичных мультимедийных корпусов и преимущества созданной поисковой платформы.

Ключевые слова: идиш, мультимедийный корпус, документация языков

ВВЕДЕНИЕ

Идиш является крайне интересным объектом исследования для лингвистов по целому ряду причин. Лексика идиша включает в себя несколько пластов: германский, славянский и семитский; идиш имеет большое количество диалектов, которые характеризуются значительными различиями в фонетике и грамматике; морфология и синтаксис идиша из-за долгих языковых контактов приобрели ряд славянских черт. Однако именно лингвистических работ, посвящённых исследованиям идиша, крайне мало. Во многом это связано с отсутствием материалов и инструментов для исследования. Единственный корпус языка идиш (<http://web-corpora.net/YNCL/search/>) появился лишь несколько лет назад и включает только письменные тексты [1]; для анализа современной устной речи исследователям необходимо искать информантов (что особенно сложно, если требуется получить данные нескольких диалектов).

Целью нашей работы было создание мультимедийного корпуса языка идиш, т.е. собрания расшифрованных видео- и аудиозаписей с грамматической разметкой и возможностью поиска посредством онлайн-интерфейса. Мультимедийные корпуса языков – сравнительно недавнее явление, и немногие языки обладают в настоящий момент таким ресурсом. Это связано с тем, что составление подобных корпусов требует больших затрат времени и сил, чем создание корпуса письменного языка, из-за меньшей доступности видео- и аудиозаписей, трудоёмкости расшиф-

ровки и более сложного устройства такого корпуса с технической точки зрения. Мы надеемся, что создание такого ресурса позволит увеличить количество исследований, использующих идишский материал.

Описываемый проект включал в себя сбор и расшифровку текстов, разработку принципов функционирования онлайн-интерфейса корпуса и создание поисковой платформы корпуса. В настоящий момент корпус содержит около 10 часов расшифрованных мультимедийных материалов, завершена разработка первой версии поисковой системы и онлайн-интерфейса, который будет размещён в Интернете до конца 2014 г.

ОБЗОР МУЛЬТИМЕДИЙНЫХ КОРПУСОВ

Первым этапом работы, предшествующим разработке собственной поисковой платформы мультимедийного корпуса, стало изучение существующих онлайн-интерфейсов мультимедийных корпусов. Нас интересовали технологии, которые использовались при создании корпусов, а именно возможности поиска, реализация проигрывания видео- и аудиофайлов и синхронизация мультимедийного содержимого с текстом. Всего было рассмотрено 10 корпусов, находящихся в свободном доступе в Интернете.

Из рассмотренных корпусов наиболее мощным поиском обладает мультимедийный подкорпус Национального корпуса русского языка [2] (<http://www.ruscorpora.ru/search-murco.html>). Разметка корпуса позволяет задать не только грамматические характеристики, но и ряд других параметров, например, количество говорящих в клипе, типы речевых действий, ориентацию ладони, направление движения и многое другое. В корпусе использован язык

* Данное научное исследование (14-05-0074) выполнено при поддержке Программы «Научный фонд НИУ ВШЭ» в 2014 г.

программирования JavaScript, видео воспроизводится с помощью плеера Yandex Player 13.101-123 в формате FLV, есть возможность скачать видео в формате MP4. Текст выдачи можно расширить, чтобы посмотреть на контекст. Синхронизации текста выдачи с видеофайлом нет.

Одним из самых первых и известных мультимедийных корпусов является корпус детской речи CHILDES, являющийся частью системы TalkBank [3]. Этот корпус содержит записи детской речи, самые ранние из которых относятся к 1960-м годам, на нескольких языках. В онлайн-интерфейсе корпуса CHILDES Transcript Browser (<http://childes.psy.cmu.edu/browser/index.php>) для поиска используется специально разработанный для этого проекта язык запросов CLAN. На сайте находятся директории с файлами различного типа: аудио-, видео- и текстовые файлы. Видеоматериалы воспроизводятся при помощи плагина проигрывателя QuickTime, который должен быть предварительно установлен у пользователя. В тексте выдачи все строки пронумерованы для синхронизации видео с текстом.

Ещё один корпус, использующий плагин плеера QuickTime для воспроизведения аудиофайлов, – Scottish Corpus of Texts & Speech [4] (<http://www.scottishcorpus.ac.uk/>). Кроме простого текстового поиска в этом корпусе предусмотрен поиск по метаинформации: имеется возможность задать автора, тип текста и некоторые другие параметры с помощью чекбоксов. Аудиофайл открывается в новом всплывающем окне, текст синхронизирован с аудиодорожкой. И текст, и аудиофайл можно скачать. Корпус написан на PHP и JavaScript.

Устные корпуса испанского языка Spanish Learner Language Oral Corpora (<http://www.splloc.soton.ac.uk/search.php>) и французского языка French Learner Language Oral Corpora (<http://www.flloc.soton.ac.uk/search.php>) построены по технологии корпуса CHILDES. В корпусе предусмотрен простой текстовый поиск с возможностью задать несколько параметров. Аудио открывается в том же окне. Корпуса написаны с использованием JavaScript.

Корпус жестов National Center for Sign Language and Gesture Resources (NCSLGR) corpus [5] (<http://secrets.rutgers.edu/dai/queryPages/querySelection.php>) обладает только простым поиском с использованием чекбоксов по разным свойствам жестов. В качестве результата поиска выдаётся таблица с названиями жестов и именами людей. В пересечении строк и столбцов находится ссылка на видео с жестом, а также на видео с контекстом. Поисковая страница написана на PHP, также использован JavaScript. Видео проигрывается при помощи встроенного плеера JW Player 5.1.897.

Мультимедийный корпусный ресурс с одним из самых красивых интерфейсов, Corpus Project in Colloquial Japanese Sign Language [6] (<http://research.nii.ac.jp/jsl-corpus/en/>), почти полностью написан на Flash. По сути, корпус представляет собой базу данных видео с жестами без возможности поиска. Видео появляется во всплывающем окне, во время воспроизведения аннотация к жестам появляется внутри самого видео.

В Диалектологическом корпусе Бассейна реки Устья (Архангельская область, Устьянский район)

(<http://www.slavist.de/Pushkino/login.php>) есть три поля для поиска: простой поиск, лексико-грамматический поиск и поиск с помощью CQP-запросов. В выдаче присутствуют аудиофайлы и текст. Результаты поиска можно экспортировать в форматах CSV и XML, аудиофайлы можно скачать в формате WAV. В корпусе есть возможность получить расширенный контекст для результатов запроса, однако аудио при этом не изменяется. Синхронизации аудио с текстом нет. Корпус полностью написан на PHP.

Педагогический корпус BACKBONE Pedagogic corpora for content & language integrated learning [7] (<http://webapps.ael.uni-tuebingen.de/backbone-search/faces/search.jsp>), созданный в Университете Тюбингена, является базой данных, которая содержит видеointервью с носителями шести языков. Видео- и аудиофайлы нельзя проиграть на сайте, их можно только скачать (формат WVX). Поиск осуществляется по категориям в базе данных, а также по коллокациям. Корпус написан с использованием JavaScript.

Ещё один проект Университета Тюбингена – English Language Interview Corpus as a Second-Language Application (http://www.uni-tuebingen.de/elisa/html/elisa_index.html) содержит интервью с носителями английского языка. По реализации этот корпус очень похож на BACKBONE. В нём также возможно только скачивание видео (формат SMIL) и просмотр текста. Есть возможность загрузить результаты в формате XML, а также просмотреть частотные списки слов.

Как видно из обзора, большинство корпусов даёт возможность простого текстового поиска, а также поиска по метаинформации и дополнительной аннотации, тип которой зависит от предназначения корпуса (например, поиск по жестам). Большинство корпусов не имеет синхронизации текста и проигрываемого аудио- или видеофрагмента, а некоторые из них позволяют только скачивать мультимедийное содержимое в виде отдельных файлов, не предоставляя возможности проигрывания его на странице.

Проанализировав достоинства и недостатки существующих платформ для мультимедийных корпусов, мы решили руководствоваться следующими принципами.

- В качестве языка клиентских скриптов будет использоваться JavaScript; для удобства обработки запросов и результатов в онлайн-интерфейсе при обмене информацией между модулями будет использоваться формат JSON, поддерживаемый языком JavaScript.
- В интерфейсе не будут использоваться Flash и другие технологии, требующие установки дополнительных плагинов к браузеру, существующие не для всех систем.
- Корпус будет содержать встроенный проигрыватель. Это продиктовано как соображениями удобства (пользователь может сразу просмотреть видео, относящееся к найденному фрагменту расшифровки), так и соображениями авторского права: мы не имеем возможности предоставить пользователям для скачивания полные версии некоторых видеофайлов.
- Текст выдачи будет двусторонне синхронизирован с видео: при проигрывании в тексте будет под-

свечиваться расшифровка проигрываемого фрагмента, а при нажатии на какой-либо фрагмент будет запускаться соответствующий ему фрагмент видео.

ХАРАКТЕРИСТИКИ КОРПУСА

Перед началом разработки корпусной платформы мы сформулировали требования, которым корпус должен удовлетворять с точки зрения поисковых возможностей, принимая в расчёт предполагаемую целевую аудиторию корпуса и задачи, для решения которых он будет использоваться. Поскольку все расшифровки аудио- и видеоматериалов, составляющих мультимедийный корпус, также будут содержаться и в Корпусе языка идиш («основном» корпусе), мы посчитали разумным выделить те группы задач, которые невозможно решить без аудио- и видеосопровождения, и при разработке мультимедийного корпуса ориентироваться в первую очередь на эти задачи.

Предполагаемыми пользователями корпуса являются исследователи, которых интересует просодия, фонетика, жесты и социолингвистика идиша. Для проведения корпусных исследований в этих областях необходимо иметь возможность параллельного просмотра видеоматериала и сопровождающей его расшифровки. При этом в корпусе должен быть реализован поиск по расшифровкам. Для проведения типологических и грамматических исследований хорошо приспособлен основной корпус, оснащённый мощным поиском. В большинстве описанных выше случаев, однако, хватает простого поиска по тексту или по лемме. Метаинформация же в мультимедийном корпусе должна быть более детальной: например, без информации о диалекте говорящего будет невозможно осуществлять социолингвистические и диалектологические исследования (разница в произношении отдельных слов носителями разных диалектов), а без детального описания жанра и типа текста станет сложным исследование жестов.

Исходя из этих соображений, мы решили реализовать в первой версии корпуса следующий поисковый функционал:

1. Поиск по словоформе (с использованием регулярных выражений).
2. Поиск по лемме (с использованием регулярных выражений).
3. Поиск по грамматике (с использованием логических функций).
4. Поиск по метаинформации.

При одновременном задании нескольких параметров они должны объединяться в один запрос с помощью конъюнкции. Поиск конструкций, состоящих из нескольких слов, реализованный в основном корпусе, в мультимедийном корпусе реализован не будет, так как он необходим в первую очередь для грамматических исследований.

Как и в основном корпусе, мы не имеем возможности предоставлять пользователям доступ к полным версиям аудио- и видеоматериалов. Одной из причин являются авторские права: авторы видео дают разрешение использовать видео в корпусе, но не размещать его целиком. Другая причина – ограниченные серверные мощности и пропускная способность ка-

нала, которых может не хватить, если поисковая выдача будет включать полные версии видео. Тем не менее, при исследовании устной речи обычно требуется доступ к более широкому контексту каждого результата поиска, чем при работе с письменными текстами. Мы решили эту проблему разбиением видеофайлов на пересекающиеся фрагменты длиной 3 минуты тремя разными способами: с отступом 0, 1 и 2 минуты. В итоге каждый кадр исходного видео содержится в трёх разных фрагментах и как минимум в одном из них он имеет отступ не менее одной минуты от начала и конца фрагмента. При индексации расшифровок каждому сегменту расшифровки присписывается адрес фрагмента видео, оптимального с точки зрения широты контекста. Если полная версия видео доступна на сайте его авторов или на каком-либо видеохостинге, результаты поиска сопровождаются ссылкой на полную версию.

ТЕКСТЫ

В первую версию корпуса включено около 10 часов мультимедийного материала, основную часть которого составляют экспедиционные записи (юго-восточный диалект) и лекции носителей идиша (северо-восточный диалект). Поскольку было решено синхронизировать текст с видео, при расшифровке было необходимо указывать, к какому фрагменту видео относится расшифровываемый фрагмент. Для этого была использована система ELAN, а итоговые файлы расшифровок хранятся в XML-формате eaf. Разметка фрагментов для привязки к видео преследовала исключительно практическую цель обеспечения синхронизации в корпусе и поэтому в общем случае не отражает какого-либо лингвистически осмысленного деления текста (например, на предложения или элементарные дискурсивные единицы), хотя в большинстве случаев границей фрагментов является пауза. Каждый фрагмент имеет длину несколько секунд.

Орфография

При выборе правил расшифровки мы руководствовались, во-первых, спецификой материала, а во-вторых, поисковыми задачами корпуса.

Специфика идишской устной речи по сравнению с письменной состоит, во-первых, в большей вариативности и частом проявлении диалектных черт в фонетике и в лексике, а во-вторых, во взаимодействии идиша с другими языками (например, русским). Последнее выражается в обилии спонтанных заимствований и переключений кодов и вызвано тем, что носители, фигурирующие в видеоматериалах, владеют и другими языками как родными, а для некоторых участников диалогов – например, интервьюеров или слушателей публичных лекций – идиш вообще не является родным, что побуждает носителей переключаться на общий для говорящих и слушающих код.

Задачи корпуса как поискового инструмента, с другой стороны, требуют сведения вариативности к минимуму. При поиске какого-либо слова оно должно находиться во всех диалектных и фонетических вариантах, поскольку одна из основных задач корпуса – предоставить пользователям возможность сравнить

произнесение одного и того же слова носителями разных диалектов. Этого можно добиться, если каждая словоформа будет записана в корпусе в соответствии с правилами литературного языка. Ещё одним обстоятельством, требующим наличия стандартизированной записи, является использование морфологического парсера. Парсер, разработанный в рамках проекта «Корпус языка идиш» Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика», работает с литературным языком и не способен распознавать диалектные варианты словоформ. Одним из решений было бы сопровождение точной расшифровки «нормализованным» вариантом, который был бы представлен в виде дополнительного слоя информации, выровненного с первым. Такое решение предлагалось, например, в [8] для русского диалектного корпуса. Этот подход позволяет сохранить больше информации, но является намного более ресурсоёмким и не имеет серьёзных преимуществ при наличии аудио- и видеосопровождения расшифровок. Поэтому мы, вслед за создателями текущей версии Диалектного подкорпуса НКРЯ [9] и аудиокорпуса Устьянского говора [10], решили придерживаться орфографического подхода: при расшифровке во всех возможных случаях словоформы записываются в стандартизированном виде, соответствующем литературной норме.

В случае с идишем, однако, такое решение требует дополнительных уточнений. Дело в том, что на практике для записи идишских текстов существуют десятки различных орфографических систем (см. [11] об истории идишского правописания), основанных на еврейском письме. Стандартный вариант орфографии был предложен институтом YIVO в 1937 г., однако он до сих пор не соблюдается многими издательствами, а кроме того, не даёт исчерпывающего описания всей орфографии, оставляя в некоторых случаях выбор за пишущим. Также институтом YIVO была разработана стандартная система транслитерации латиницей.

При расшифровке мы остановились на использовании стандартной латинской транслитерации, а не одной из орфографий, использующих еврейскую письменность. Причин такому решению несколько. Во-первых, использование транслитерации не только не приводит к потере информации, но и, наоборот, добавляет новую информацию: если в стандартной орфографии гебраизмы – заимствования из иврита и арамейского – пишутся по правилам языка-оригинала, без огласовок, то транслитерация позволяет их прочитать и тем, кто не знаком с этим пластом идишской лексики. Во-вторых, применение стандартной орфографии может быть затруднено при записи некоторых заимствований, которыми изобилуют тексты. В-третьих, такой подход позволяет передавать переключения кодов с помощью переключения систем письма (например, русский текст может быть записан кириллицей), что вызвало бы проблемы при использовании оригинальной орфографии из-за разницы в направлении письма.

Основываясь на описанной концепции, мы составили инструкцию для разметчиков, в которой были более подробно рассмотрены отдельные пункты. Ни-

же приведены основные решения, которые было необходимо разработать для сложных случаев:

1. В расшифровках не приводятся к литературной форме слова, имеющие существенные грамматические отличия от литературного идиша, например, использующие другую морфему для выражения какого-либо значения.

2. Если в диалекте одной литературной лексеме соответствует два варианта с разным значением, это различие отмечается. Например, слову *shul* ‘школа; синагога’ в некоторых диалектах со стандартным фонетическим переходом *u* → *i* соответствует пара слов *shil* (с переходом) и *shul* (без перехода), одно из которых имеет значение «школа», а другое — «синагога». В таком случае различие *shul/shil* должно сохраняться и, соответственно, форма *shil* не должна нормализовываться.

3. Переключения кодов планируется записывать в орфографии языка, на который переключается говорящий (в частности, переключения на русский, довольно частые в нашем корпусе, записываются кириллицей). Однако заимствования зачастую бывает сложно или невозможно отличить от переключения кодов, из-за чего некоторые исследователи даже предлагают отказаться от этого различия [12]. Следовательно, для достижения согласованности расшифровщиков необходимо разработать чёткие правила, позволяющие выделить переключения кодов. Такими при расшифровке считаются только фрагменты длиной более одного слова, которые невозможно интерпретировать как заимствования по грамматическим или фонологическим соображениям.

АРХИТЕКТУРА ПОИСКОВОЙ ПЛАТФОРМЫ

Поисковая платформа разрабатывалась в соответствии с требованиями, сформулированными в п. 2. Архитектура платформы включает три основные части: пользовательский интерфейс, выполненный на HTML/JavaScript, поисковый механизм, включающий в себя базу данных и модуль на языке Python, через который можно задавать поисковые запросы и получать результаты, и веб-сервер, написанный также на языке Python с использованием веб-фреймворка Django. Последний отвечает, среди прочего, за связь интерфейса с поисковым механизмом и авторизацию пользователей. Данные между этими тремя модулями передаются через json-объекты.

Веб-сервер

В функции веб-сервера входит, во-первых, выдача веб-страниц по запросу, а во-вторых, организация взаимодействия между интерфейсом и поисковой системой. Сервер корпуса, как и поисковый механизм, написан на языке Python с использованием одного из самых распространённых веб-фреймворков – Django. Программа, написанная с использованием этого веб-фреймворка, может либо самостоятельно обслуживать запросы, либо работать через посредство другого сервера, установленного в системе. Первой из этих возможностей мы пользовались во время тестирования и отладки интерфейса корпуса; версия,

доступная для пользователей, получает запросы и отправляет ответы через веб-сервер Apache.

Дополнительное звено между интерфейсом и поисковым механизмом позволило реализовать несколько функций, невозможных при использовании более простой структуры. Веб-сервер отвечает за следующие функции:

1. Преобразование поискового запроса пользователя в json-запрос к базе данных. Пользователь вводит информацию о словоформе, которую он хочет найти, в несколько текстовых полей. В настоящий момент это поля «Лемма», «Словоформа» и «Грамматика». Если первые два поля содержат обычные строки, то поле «Грамматика» может содержать сложные выражения, содержащие логические функции: конъюнкцию, дизъюнкцию и отрицание (например, «найти все глаголы в форме причастий или прилагательные»). Строка запроса в этом случае должна быть преобразована в json-объект древовидной структуры (см. п. 5.2), что удобнее делать на сервере, чем на компьютере клиента.

2. Получение ответа от поискового сервера и отправка его в интерфейс по частям. Поскольку по запросу пользователя может быть найдено очень много результатов, мы решили придерживаться стандартной схемы выдачи результатов, согласно которой они делятся на страницы (по умолчанию 20 результатов на странице). Если бы все результаты загружались сразу, это привело бы к большой нагрузке на сервер и чрезмерно долгому ожиданию ответа пользователем. Сервер, получив ответ от поискового механизма, отправляет часть полученных результатов и информацию об общем количестве результатов клиенту, после чего пользователь может запросить другую часть результатов, нажав на ссылку с номером страницы.

3. Поддержка нескольких языков интерфейса. Было решено выполнить интерфейс корпуса на двух языках – русском и английском (с перспективой добавить в дальнейшем интерфейс на идише). Наиболее удобный и масштабируемый вариант решения этой задачи состоит в том, что в HTML-коде страниц вместо надписей помещаются специальные шаблоны, на место которых сервер вставляет текст из соответствующего языкового файла.

4. Панель администрирования. В платформу было решено добавить несколько функций, доступных администраторам, таких как просмотр истории поисковых запросов пользователей и загрузка новых текстов. Фреймворк Django хорошо подходит для этой задачи, поскольку имеет встроенные средства для регистрации пользователей и создания панели администратора.

Поисковый механизм

В большинстве существующих корпусных поисковых систем используются либо реляционные базы данных (например, в Corpus of Contemporary American English и Corpus del Español [13] или в Восточно-армянском национальном корпусе [14] и других корпусах, использующих ту же систему, в том числе в основном идишском корпусе), либо специальный веб-сервер с индексными файлами, не хранящимися в

базе данных (например, в системе Corpus Workbench [15] и в Национальном корпусе русского языка [16]). В нашей поисковой системе мы решили пойти по третьему пути и использовать нереляционные (NoSQL) базы данных. Такой подход уже применялся на практике (например, CouchDB в комбинации с Apache Lucene в [17] и MongoDB в [18]), но ещё не получил большого распространения в создании корпусов из-за новизны этой технологии.

В качестве базы данных была выбрана MongoDB. В отличие от реляционных баз данных, в MongoDB единицей хранения является BSON-документ (аналог строки в реляционных базах); документы объединяются в коллекции (аналог таблицы). В отличие от реляционных баз, задающих жёсткие рамки в виде фиксированного числа и типа столбцов в каждой таблице, объекты в MongoDB могут иметь любое количество свойств и иметь сложную внутреннюю структуру. Поисковый запрос к MongoDB представляет собой json-объект специального вида. Несмотря на намного более свободную организацию хранения, нереляционные базы данных позволяют создавать индексы по любым полям объектов и их сочетаниям и использовать их для быстрого поиска.

В текущем варианте поисковой системы база данных содержит коллекции текстов, предложений и словоформ. Каждый документ в коллекции текстов содержит метаданные для одного из текстов корпуса. Документы коллекции слов содержат информацию обо всех уникальных словоформах, имеющихся в корпусе, вместе с их грамматическими разборами и всей остальной информацией, доступной для поиска. Эта коллекция играет роль обратного индекса: каждый документ содержит список идентификаторов предложений, в которых содержится соответствующая словоформа. Документы коллекции предложений описывают предложения (или другие элементарные фрагменты текста), представляющие собой массивы словоформ и содержащие дополнительную информацию, в частности, идентификаторы соседних предложений и идентификатор документа. Коллекция предложений, в отличие от двух других коллекций, предназначена в первую очередь для выдачи результатов, а не для поиска, поэтому словоформы в ней могут содержать дополнительные поля, содержимое которых должно быть показано в выдаче, но не должно быть доступно для поиска.

Ниже приводится начальный фрагмент одного из предложений коллекции:

```
{ "video_source": "video1.mp4",  
  "doc": "interview_1.eaf",  
  "start_offset": "1069.49",  
  "words": [  
    { "wf": "zaynen",  
      "type": "word",  
      "ana": [  
        { "lex": "zayn",  
          "transl_en": "be",  
          "gr": { "tense": "pres",  
                  "pos": "V",  
                  "number": "pl" } }  
      ]  
    }  
  ]  
}
```

```

{"wf": "linkse",
 "type": "word"},
{"wf": ",",
 "type": "punc"},
...

```

В примере выше показаны ссылки на видеофайл и исходный файл с расшифровкой, время начала предложения в виде отступа от начала видео в секундах и начало массива словоформ, составляющих предложение. Первое слово имеет единственный грамматический разбор, включающий поля *lex* (лемма), *transl_en* (перевод на английский язык) и поле *gr* (грамматика), которое, в свою очередь, распадается на несколько полей, соответствующих грамматическим категориям. Вторая словоформа (*linkse*) не имеет разбора, а третья является знаком пунктуации, что отражено в поле *type* (тип токена).

Очевидным преимуществом представленного способа хранения является то, что извлечённые из базы данных документы, соответствующие предложению, можно почти без изменений передавать в интерфейс, поскольку они являются json-объектами. Модификации, которым эти объекты подвергаются перед передачей в интерфейс, включают удаление избыточной информации (например, ссылка на исходный файл расшифровки не нужна при просмотре результатов пользователем) и подсветку словоформ, являющихся результатом запроса (что реализуется добавлением к словоформе атрибута *highlight*).

Запрос к такой базе данных тоже имеет вид json-объекта. В самом простом варианте запрос представляет собой словарь, где ключам, представляющим названия атрибутов, соответствуют их желаемые значения. Для выполнения более сложных запросов, включающих логические функции или регулярные выражения, используется специальный язык запросов MongoDB. Ниже приводится пример такого запроса:

```

{"$or": [{"ana.gr.case": "gen"},
 {"ana.gr.number": "pl"}], "ana.lex": {"$regex": "^z.*"}}

```

В приведённом запросе используются специальные обозначения "\$or", обозначающее дизъюнкцию, и "\$regex", обозначающее поиск по регулярному выражению. Весь запрос предназначен для поиска в коллекции словоформ и означает «найти все словоформы, грамматический разбор которых имеет тэг родительного падежа или тэг множественного числа и лемма которых начинается с буквы z».

Информация, которая вводится пользователем в поля поисковой формы, передаётся серверу в виде GET-запроса и преобразуется сервером в поисковый json-запрос описанного выше вида, после чего отправляется на обработку в поисковую систему. Ответ, полученный от поисковой системы, передаётся обратно в интерфейс в виде json-объекта.

Пользовательский интерфейс

Онлайн-интерфейс корпуса по существу представляет собой одну страницу, которая может выполнять функции задания поискового запроса и отображения результатов. Все операции, связанные с

передачей данных между сервером и клиентом, выполняются с помощью технологии Ajax, что позволяет изменять содержимое страницы (например, при задании нового запроса или переходе на следующую страницу выдачи) без её перезагрузки. Эта страница состоит из шапки и поисковых полей. В шапке находятся ссылки на информацию о проекте, инструкцию по использованию корпуса, а также на вход в панель администратора. Также там расположены кнопки для смены языка страницы.

В поисковых полях пользователь может задать лемму, словоформу (в том числе с помощью регулярных выражений), а также грамматические характеристики (либо с помощью чекбоксов, либо в виде строки, содержащей тэги и логические функции). Поисковое поле можно свернуть для удобного просмотра результатов выдачи.

Выдача подгружается на главную страницу под поисковым полем. С левой стороны расположен видеоплеер, в который загружаются необходимые видеофайлы. Из видеофайлов формируется плейлист, за счёт чего на странице находится только одно окно с видео, а не таблица из видео, как это сделано во многих других мультимедийных корпусах.

Для нашего корпуса мы использовали плеер Projekktor V1.3.09 (<http://www.projekktor.com/>). Этот плеер обладает рядом встроенных функций, которые показали нам полезными для нашего проекта, – формирование плейлиста и установка курсора на определённые фрагменты видео, необходимые для синхронизации с текстом.

Справа от видео расположен текст выдачи. Окно с текстом имеет такую же высоту, что и видеоплеер, и сопровождается полосой прокрутки. Полоса прокрутки, а также сворачивание поискового окна реализованы с помощью специального модуля, использующего функционал библиотеки jQuery.

В поисковой выдаче словоформы, являющиеся результатом поиска, выделены оранжевым цветом, при наведении курсора на каждое слово появляется морфологическая разметка в маленьком всплывающем окне. Во время воспроизведения видео синхронизированные фрагменты текста подсвечиваются.

ПЕРСПЕКТИВЫ

Одним из приоритетных направлений развития мультимедийного корпуса является увеличение количества текстов. С точки зрения разметки существует несколько возможностей для улучшения. Во-первых, в будущем может быть добавлен отдельный слой расшифровки с точной фонетической записью, который будет сопровождать основной слой с орфографической записью. Во-вторых, с помощью автоматических систем принудительного выравнивания (*forced alignment*) фрагменты, являющиеся единицами синхронизации, могут быть уменьшены до одного слова или даже до одной фонемы, что упростит фонетические исследования с помощью корпуса. Наконец, в-третьих, к грамматической разметке может быть добавлена другая разметка, релевантная для мультимедийных материалов, например, разметка жестов.

СПИСОК ЛИТЕРАТУРЫ

1. Кирьянов Д.П., Лучина Е.С., Панова Т.А., Тагабилева М.Г. Корпус языка Идиш: теория и практика // Тирош – труды по иудаике. – 2014. – № 14. – С. 78–90.
2. Гришина Е.А. Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. – СПб.: Нестор-История, 2009. – С. 175–214.
3. MacWhinney B. From CHILDES to TalkBank // Research on Child Language Acquisition / eds. M. Almgren, A. Barreña, M. Ezeizaberrera, I. Idiazabal & B. MacWhinney. – Somerville, MA: Cascadia, 2001 – P. 17–34.
4. Anderson J., Beavan D., Kay C. SCOTS: The Scottish corpus of texts and speech // Creating and Digitizing Language Corpora / eds. J. Beal, K. Corrigan, H. Moisl. – 2007. – Vol. 1: Synchronic Databases. – Basingstoke: Palgrave Macmillan. – P. 17–34.
5. Neidle C., Vogler C. A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface // The 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012. – Istanbul, Turkey, May 27, 2012.
6. Mayumi Bono, Kouhei Kikuchi, Paul Cibulka, Yutaka Osugi. A Colloquial Corpus of Japanese Sign Language: Linguistic Resources for Observing Sign Language Conversations // Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014. – Reykjavik, Iceland, May 26–31, 2014.
7. Kohn K. The BACKBONE project: pedagogic corpora for content and language integrated learning. Objectives, methodological approach and outcomes // Eurocall Review. – 2012. – Vol. 20, № 2.
8. Качинская И.Б. Корпус Диалектных Текстов в Национальном корпусе русского языка: состояние и перспективы // Лексический атлас русских народных говоров (Материалы и исследования). – СПб., 2009. – С. 57–68.
9. Летучий А.Б. Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003–2005. – М.: Индрик, 2005. – С. 215–232.
10. von Waldenfels R., Daniel M., Dobrushina N. Why Standard Orthography? Building the Ustyia River Basin Corpus, an online corpus of a Russian dialect // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). Вып. 13 (20). – М.: Изд-во РГГУ, 2014.
11. Schaechter Mordkhe. Fun Folkshprakh tsu Kulturshprakh [The History of the Standardized Yiddish Spelling]. – New York: League for Yiddish/YIVO, 1999. – 111 с.
12. Muysken P. Bilingual speech: A typology of code-mixing. – Cambridge University Press, 2000. – 306 с.
13. Davies M. The advantage of using relational databases for large corpora: speed, advanced queries and unlimited annotation // International Journal of Corpus Linguistics. – 2005. – Vol. 10, № 3. – P. 307–334.
14. Даниэль М.А., Поляков А.Е., Рубаков С.В., Левонян Д.В., Плунгян В.А., Хуршудян В.Г. Восточноармянский национальный корпус // Армянский гуманитарный вестник. Вып. 2/3-II. – М.; Ереван: Зангак-97, 2009. – С. 9–33.
15. Evert S., Hardie A. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. – 2011.
16. Аброскин А.А. Поиск по корпусу: проблемы и методы их решения // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. – СПб.: Нестор-История, 2009. – С. 277–282.
17. Sliwkanich T. et al. Towards scalable summarization and visualization of large text corpora // Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. – ACM, 2012. – P. 863.
18. Niekler A., Wiedemann G., Heyer G. Leipzig Corpus Miner – A Text Mining Infrastructure for Qualitative Data Analysis // Terminology and Knowledge Engineering 2014. – 2014.

Материал поступил в редакцию 12.12.14.

Сведения об авторах

АРХАНГЕЛЬСКИЙ Тимофей Александрович – кандидат филологических наук, преподаватель факультета филологии Национального исследовательского института «Высшая школа экономики», Москва
e-mail: tarkhangelskiy@hse.ru

СОЗИНОВА Ольга Андреевна – студент факультета филологии Национального исследовательского университета «Высшая школа экономики», Москва
e-mail: oa.sozinova@gmail.com