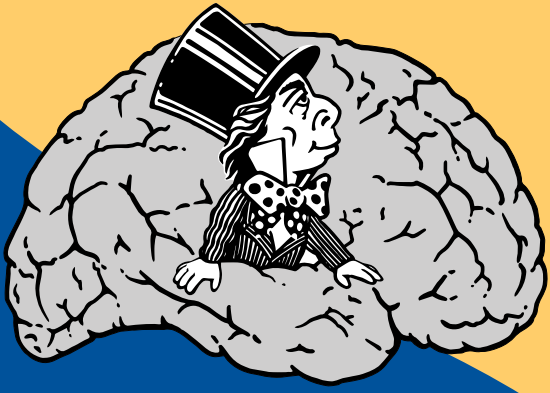
КОГНИТИВНАЯ НАУКА

В МОСКВЕ



НОВЫЕ ИССЛЕДОВАНИЯ

МАТЕРИАЛЫ КОНФЕРЕНЦИИ 2019

Под ред. Е.В. Печенковой, М.В. Фаликман

УДК 159.9
ББК 88.25
К57

Когнитивная наука в Москве: новые исследования. Материалы конференции 19 июня 2019 г. Под ред. Е. В. Печенковой, М. В. Фаликман. – М.: ООО «Буки Веди», ИППиП. 2019 г. — 656 стр.

ISBN 978-5-4465-2346-7

УДК 159.9
ББК 88.25

ISBN 978-5-4465-2346-7

СИСТЕМА ПОИСКА В МУЛЬТИКАНАЛЬНОМ КОРПУСЕ «РАССКАЗЫ И РАЗГОВОРЫ О ГРУШАХ»

Н. А. Коротаев* (1, 2), Г. Б. Добров (2), А. Н. Хитров (3)
n_korotaev@hotmail.com
1 — РГГУ, Москва; 2 — Институт языкознания РАН, Москва; 3 — Институт русского языка им. В. В. Виноградова РАН, Москва

Аннотация. В докладе описывается поисковая система, разрабатываемая на материале мультиканального корпуса «Рассказы и разговоры о грушах». Поиск реализован на сайте проекта (http://multidiscourse.ru/search/) и опирается на единую схему мультиканальной аннотации. В рамках этой схемы различные действия участников естественной коммуникации (вербальные, просодические, жестикуляционные, глазодвигательные и проч.) размечаются по общим принципам: выделяются единицы одного или нескольких уровней, для которых в дальнейшем указываются дополнительные свойства. Итоговые разметки хранятся в формате eaf, предназначенном для работы в программной среде ELAN. Благодаря этому становится возможным опираться на движок ELAN в серверной части поисковой системы. В свою очередь, клиентская часть представляет собой разработанное «с нуля» одностраничное приложение, в котором процесс составления запроса сделан более интуитивным и дополнительно адаптирован под особенности проекта. Формируя запрос, пользователь может ограничить область поиска отдельными записями или их этапами, выбрать тип единицы и указать ее свойства, а также составить сложную цепочку из нескольких единиц, связанных между собой различными отношениями на единой временно́й шкале. В выдаче по запросу каждый найденный результат представлен в текстовом формате, а также сопровождается соответствующим видеофрагментом. Насколько нам известно, разрабатываемый ресурс не имеет аналогов в российской лингвистике.

Ключевые слова: корпус, поиск, мультиканальная коммуникация, веб-интерфейс, разметка корпусов, лингвистические программные продукты, мультимодальность

Работа выполнена при финансовой поддержке РФФИ, проект № 18-00-01598 КОМФИ.

Доклад посвящен технологическим аспектам создания и использования мультиканального корпуса русского языка, в первую очередь — вопросам, связанным с разработкой поисковой системы по корпусу. Мультиканальность — это одно из базовых свойств естественной коммуникации. Несмотря на то что лингвисты в течение долгого времени ограничивали объект своего изучения вербальным сигналом, в действительности участники коммуникации, решая свои задачи, прибегают не только к речи, но и к другим каналам: просодии, жестикуляции, движению глаз и т.д. (Adolphs, Carter, 2013; Müller et al., 2014 и др.). Среди вопросов, решаемых в рамках мультиканального подхода,

возможно, центральное место занимает вопрос о характере координации различных ресурсов при воплощении коммуникативного замысла. Так, например, широко известна гипотеза «единой точки роста», согласно которой мануальная жестикуляция и речь выступают репрезентациями некоторой общей внутренней программы, что проявляется, в частности, в опережающем характере жестикуляции (McNeill, 1992; Гришина, 2017). Если для проверки этого и других подобных предположений опираться не на экспериментальные данные, а на данные естественной коммуникации, в распоряжении исследователя должен иметься некоторый корпус — коллекция аннотированных записей мультиканальной коммуникации. Корпус, в свою очередь, должен быть снабжен содержательно обоснованной системой поиска и извлечения информации. При всей очевидности этой задачи, для мультиканальных корпусов она пока решена в меньшей степени, чем для корпусов письменной или устной речи. В качестве примера можно привести, пожалуй, самый известный ресурс такого рода на русском языке — мультимедийный корпус в составе Национального корпуса русского языка (МУРКО; http://ruscorpora.ru/search-murco.html). В МУРКО используется дополнительная жестовая разметка, в основном поиск, как и в других частях НКРЯ, ориентирован на вербальные единицы. В настоящем докладе представлено краткое описание того, какая поисковая система разрабатывается для корпуса «Рассказы и разговоры о грушах».

Формулировка задачи

Корпус «Рассказы и разговоры о грушах» (https://multidiscourse.ru/) состоит из однотипных записей, в каждой из которых принимают участие четыре коммуниканта с фиксированными ролями: Рассказчик, Комментатор, Пересказчик и Слушатель (подробнее см. Кибрик, 2018). Во время записи фиксируется вокальное и кинетическое поведение трех основных участников (Рассказчика, Комментатора и Пересказчика), которое затем анализируется и размечается в программах Praat (http://www.fon.hum.uva.nl/praat/) и ELAN (https://tla.mpi.nl/tools/tla-tools/elan/). С содержательной точки зрения, имеющаяся разметка позволяет оценить, как соотносятся между собой единицы различных каналов и их свойства — и в поведении одного участника, и при взаимодействии нескольких участников. Соответственно, разрабатываемая система поиска должна в первую очередь извлекать данные как одного, так и нескольких каналов, ориентируясь на задаваемые пользователем свойства. Дополнительным требованием является реализация этой системы в браузере, без необходимости скачивать «тяжелые» медиафайлы и/или осваивать дополнительные программные средства. Разобравшись в интуитивно понятном интерфейсе, пользователь должен получить возможность формулировать простые и составные запросы, просматривать результаты поиска и узнавать количество найденных контекстов.

Методы

Содержательной основой поиска является единая схема мультиканальной аннотации. В схеме реализованы общие принципы разметки для различных

коммуникативных каналов. Внутри каждого канала для каждого участника, во-первых, выделяются единицы поведения (элементарные дискурсивные единицы, слова и паузы в речи; жесты и минимальные движения в мануальной жестикуляции; фиксации глаз в окуломоторном канале; и проч.); во-вторых, для каждой выделенной единицы указываются ее свойства (подробнее см. Кибрик и др., 2019). Выполненная разметка хранится в формате eaf, который используется при работе в программной среде ELAN. В этой программе существует и встроенная поисковая система, которая, однако, не отвечает описанным выше требованиям: она не функционирует в браузере, а для работы с ней пользователю нужно освоить достаточно сложный формат построения запроса и досконально разобраться в используемой аннотационной схеме.

В то же время в ELAN уже реализован готовый поисковый движок по аннотациям в формате eaf — и этот движок можно использовать во внутренней (серверной) части онлайн-поиска. В текущей версии разрабатываемой нами поисковой системы принято именно это решение. Серверная часть выполнена на языке Java. Запросы от клиентской части (создаваемые в формате JSON) обрабатываются с использованием технологии Java Servlet, далее в качестве механизма поиска используются классы ELAN: запрос преобразуется в поисковые объекты и условия этой программы. Полученные результаты конвертируются в JSON и возвращаются клиенту. Клиентская часть, в свою очередь, организована как одностраничное приложение на языке JavaScript; это приложение было разработано «с нуля».

Текущие результаты

Ниже описаны характеристики первой стабильной версии поисковой системы (v0.8.29), доступной по адресу http://multidiscourse.ru/search/. Пользователю доступны основные возможности, имеющиеся во встроенном поиске программы ELAN, однако процесс составления запроса сделан более интуитивным и дополнительно адаптирован под особенности проекта. Используя графический интерфейс, пользователь системы может: (а) ограничивать область поиска конкретными записями или этапами записей (на данный момент доступны разметки трех записей суммарной продолжительностью около 60 минут; в дальнейшем планируется как минимум удвоить объем аннотированного материала); (б) выбирать единицы поиска (пока выбор производится между единицами вокального, мануального и окуломоторного каналов) и указывать их дополнительные свойства; (в) составлять запросы, включающие в себя одну или несколько единиц — в последнем случае можно также задавать различные ограничения на характер взаимного расположения единиц на единой временно́й шкале.

На рис. 1 показана структура запроса, включающая в себя три единицы поиска: (1) элементарную дискурсивную единицу (ЭДЕ), произносимую Пересказчиком, продолжающуюся не менее 200 мс и имеющую иллокутивное значение вопроса или полуутверждения; (2) фиксацию взгляда Пересказчика на Рассказчике или Комментаторе, левая граница которой находится на расстоянии от 0 до 400 мс от левой границы ЭДЕ 1; (3) прагматический мануальный жест

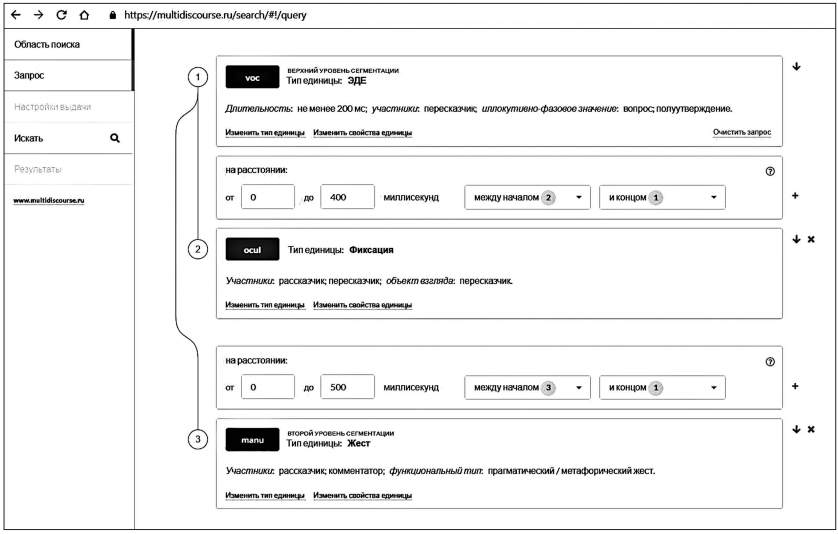


Рисунок 1. Общая структура сложного запроса на сайте http://multidiscourse.ru/search/

Рассказчика или Комментатора, начинающийся не позднее чем через 500 мс после завершения произнесения Пересказчиком ЭДЕ 1. Сложные контексты подобного рода могут встречаться на этапе разговора: на их примере можно наблюдать, как происходит координация мультиканальных действий участников при выяснении Пересказчиком тех или иных подробностей обсуждаемого стимульного материала (см. Коротаев, 2018).

Результаты поиска отображаются на отдельной вкладке. Образец выдачи представлен на рис. 2. Это результаты простого запроса, в котором требовалось найти все ЭДЕ со свойствами, указанными выше на рис. 1, произносимые Пересказчиками на этапе разговора в записях 04 и 22. Приведено общее количество найденных соответствий; для каждого соответствия указан кодовый номер найденной ЭДЕ и приведена ее транскрипция; также приводятся минимальный левый (верхний) и правый (нижний) контексты. В верхней области

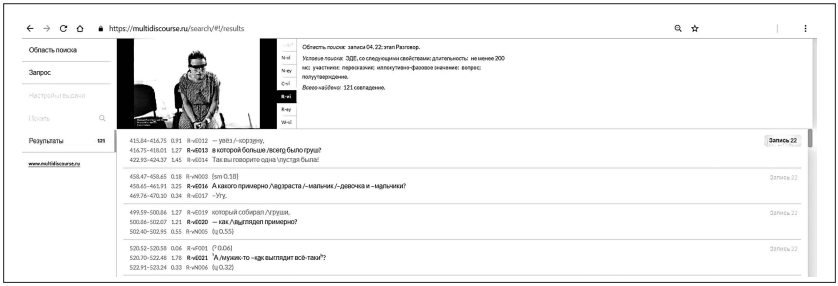


Рисунок 2. Образец выдачи результатов по запросу

окна отображается плеер с индивидуальным видеофайлом; при нажатии на каждый из найденных контекстов в плеере проигрывается соответствующий видеофрагмент.

Таким образом, система поиска по мультиканальному корпусу представляет собой полезный инструмент для изучения естественной коммуникации. С его помощью можно формулировать запросы, относящиеся к различным аспектам мультиканального поведения, как изолированно, так и в разнообразных комбинациях. При разработке системы мы стремимся отделить технические решения от конкретных аннотаций: потенциально это позволяет использовать мультиканальную разметку, выполненную другими научными коллективами.

Литература

Гришина Е. А. Русская жестикуляция с лингвистической точки зрения (корпусные исследования). М: Языки славянских культур, 2017. doi:10.31912/rusgest-2017-301-8

Кибрик А. А. Русский мультиканальный дискурс. Часть I. Постановка проблемы // Психологический журнал. 2018. Т. 39. № 1. С. 70 – 80.

Кибрик А. А., Коротаев Н. А., Федорова О. В., Евдокимова А. А. Единая мультиканальная аннотация как инструмент анализа естественной коммуникации // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог – 2019». М: 2019. http://www.dialog-21.ru/media/4603/kibrikaaplusetal-046.pdf

Коротаев Н. А. Вопрос и полуутверждение в структуре мультиканального дискурса // Восьмая Международная конференция по когнитивной науке, 18 – 21 октября 2018 г., Светлогорск, Россия. Тезисы докладов. 2018. С. 1311 – 1313.

Adolphs S., Carter R. Spoken corpus linguistics: From monomodal to multimodal. NY: Routledge, 2013. doi:10.4324/9780203526149

McNeill D. Hand and mind: What gestures reveal about thought. Chicago: University of Chicago Press, 1992.

Body — Language — Communication: An international handbook on multimodality in human interaction / C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, S. Teßendorf (Eds.). Berlin: De Gruyter Mouton, 2014.

AN ONLINE SEARCH ENGINE FOR THE "RUSSIAN PEARS CHATS AND STO RIES" MULTICHANNEL CORPUS

N. A. Korotaev* (1, 2), G. B. Dobrov (2), A. N. Khitrov (3)
n_korotaev@hotmail.com
1 — RSUH, Moscow; 2 — Institute of Linguistics RAS, Moscow; 3 — Russian Language Institute RAS, Moscow

Abstract. In this talk, we briefly characterize the online search engine that is being developed for the "Russian Pears Chats and Stories" multichannel corpus (http://multidiscourse.ru/search/). The system under development uses multichannel annotations stored in .eaf format, which is compatible with ELAN software. Since ELAN provides an internal search system that allows for an integration via Java Servlet, we use this solution on the server side. On the client side, however, we have developed a single-page application from scratch. Users of the online interface can define search domains; select units of vocal, manual, or

oculomotor behavior; specify their formal and substantial properties; and combine them in complex queries. The results are shown in a text format together with the corresponding video fragments.

Keywords: corpus, search, multichannel communication, online interface, corpora annotation, linguistic software, multimodality