ПОИСК В МУЛЬТИКАНАЛЬНОМ КОРПУСЕ: РАЗРАБОТКА ОНЛАЙНОВОЙ СИСТЕМЫ*

Коротаев Н. А. (n_korotaev@hotmail.com)
РГГУ, Институт языкознания РАН, Москва, Россия

Добров Г. Б. (wslcdg@gmail.com)
Институт языкознания РАН, Москва, Россия

Хитров А. Н. (a.n.khitrov@gmail.com)
Институт русского языка РАН имени В. В. Виноградова РАН, Москва, Россия

Ключевые слова: мультиканальная коммуникация, корпус, онлайн-поиск, веб-разработка.

SEARCH IN A MULTICHANNEL CORPUS: DEVELOPPING AN ONLINE VERSION

Korotaev N. A. (n_korotaev@hotmail.com)
RSUH, Institute of Linguistics RAS, Moscow, Russia

Dobrov G. B. (wslcdg@gmail.com)
Institute of Linguistics RAS, Moscow, Russia

Khitrov A. N. (a.n.khitrov@gmail.com)
Russian Language Institute RAS, Moscow, Russia

In this talk, we present preliminary results of developing an online search engine for the multichannel corpus “Russian Pear Chats and Stories” (http://multidiscourse.ru/search/). The engine operates on about 200 000 ELAN annotations that register vocal, oculomotor, and manual behavior of the participants of three communication sessions (approximately one hour long). On the server side, we rely on the internal ELAN search engine that we extend to implement additional features. We use the Java Servlet technology to transform user-generated queries into ELAN classes. On the client side, we provide a new friendly graphic user interface. It is implemented as a single page JavaScript application based on the Model-view-viewmodel pattern. Users can define a search domain, select units of multichannel behavior and specify their properties, create simple and complex queries, and play relevant video fragments in the Results section.

Keywords: multichannel communication, corpus, online search, web development

1. Предварительные замечания

В докладе описывается опыт создания поисковой системы на сайте мультиканального корпуса «Рассказы и разговоры о грушах» (https://multidiscourse.ru/). Корпус состоит из аудио- и видеозаписей, в которых регистрируется вокальное и кинетическое поведение коммуникантов, рассказывающих и обсуждающих между собой содержание «Фильма о грушах» (Chafe (ed.) 1980). В каждом коммуникативном эпизоде (записи) задействованы три активных участника с фиксированными ролями: Рассказчик, Комментатор и Пересказчик. На первом этапе записи Рассказчик, до начала записи просмотревший «Фильм о грушах», в монологическом режиме сообщает его содержание Пересказчику, который фильма ранее не видел. Во время второго этапа Комментатор, также знакомый со стимульным фильмом, уточняет рассказ Рассказчика, а Пересказчик выясняет у обоих

* Исследование выполнено при поддержке РФФИ, грант №18-00-01598 КОМФИ «Речевые сбои и жестикуляция: лингвистический и нейрофизиологический аспекты».

своих собеседников дополнительные подробности. На заключительном этапе в комнате появляется Слушатель, которому Пересказчик рассказывает содержание фильма. Подробнее о корпусе см. Кибрик 2018.

При разметке корпуса используется единая схема мультиканальной аннотации (см. Коротаев и др. 2018), включающая разметку речи и просодии (вокальная составляющая), разметку мануальных и цефалических жестов, а также разметку движения глаз. Результаты разметки хранятся в аннотационных файлах формата eaf, используемых в программной среде ELAN (https://tla.mpi.nl/tools/tla-tools/elan/; Hellwig et al. 2018). Текущая версия поисковой системы основана на данных вокальной (Коротаев 2019), мануальной (Литвиненко и др. 2017) и окуломоторной (Федорова 2017) разметки трех записей корпуса. Суммарная длительность этих записей составляет около 1 часа, общее число непустых интервалов в слоях ELAN (т.н. «аннотаций») — около 200 000.

2. Компоненты поисковой системы

В соответствии со стандартными принципами веб-разработки поисковая система состоит из серверной и клиентской частей; связь между ними в нашем случае обеспечивается посредством интерфейса Java Servlet[1]. Клиентская часть выполнена в виде одностраничного приложения на языке JavaScript с использованием шаблона проектирования Model-view-viewmodel. Поисковые запросы составляются пользователем в графическом интерфейсе (подробнее см. раздел 3), после чего в формате JSON передаются на сервер.

Серверная часть написана на языке Java. После обработки полученного от клиента файла JSON запрос преобразуется в поисковые объекты и условия программы ELAN. Таким образом, мы используем встроенный поисковый движок этой программы, оптимизированный для поиска по eaf-аннотациям. С известной долей условности можно сказать, что основным результатом нашей работы стало создание нового браузерного интерфейса для поиска ELAN. Впрочем, доступных для интеграции классов ELAN оказалось недостаточно для того, чтобы формировать и обрабатывать запросы всех требуемых типов. Поэтому исходный поисковый движок был доработан для обеспечения необходимой функциональности. В частности, были добавлены следующие возможности:

- ограничение результатов поиска по длительности запрашиваемых единиц;
- обработка числовых значений (в ELAN все аннотации хранятся и обрабатываются как строки);
- указание интервалов расстояния между левыми / правыми границами единиц в сложном запросе (см. ниже раздел 3).

После того как запрос обрабатывается при помощи модифицированного поискового движка ELAN, полученные результаты преобразуются в формат JSON и возвращаются на сторону клиента. В выдаче по запросу пользователь получает список найденных контекстов, ассоциированных с фрагментами соответствующих видеофайлов.

Общая схема взаимодействия клиентской и серверной частей представлена на рис. 1.

[1] Далее описываются характеристики поисковой системы, актуальные для версии v0.8.29, выгруженной на сайт проекта в феврале 2019 года.

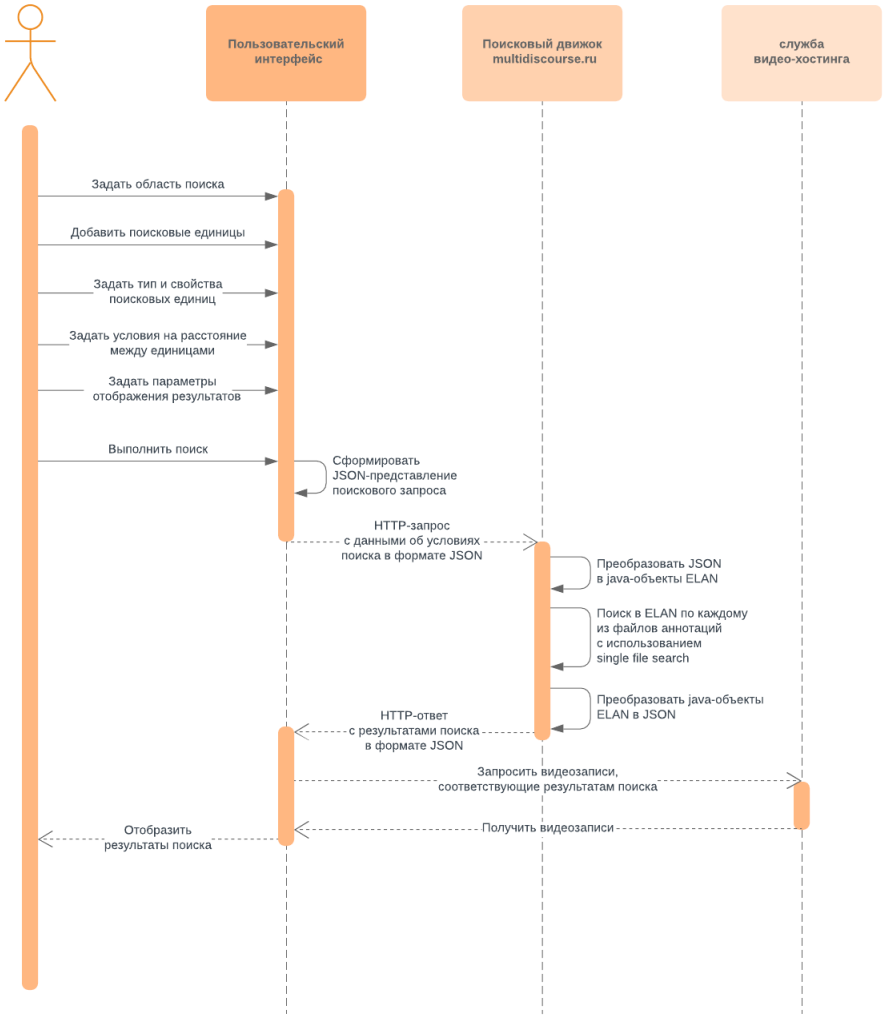

Рис. 1. Схема взаимодействия клиента и сервера в поисковой системе на сайте http://multidiscourse.ru/search/

3. Возможности пользовательского интерфейса

Ниже кратко представлены основные возможности, доступные пользователям в веб-интерфейсе поисковой системы. Приложение организовано как набор вкладок, для навигации по которым используется левое меню. Переходя по вкладкам, пользователь может:

- ограничить поиск конкретными записями и / или этапами записей (вкладка «Область поиска»);
- выбрать единицы поиска, указать их свойства и связать единицы в структуре запроса (вкладка «Запрос»);
- просмотреть результаты запроса (вкладка «Результаты»).

По умолчанию при открытии страницы http://multidiscourse.ru/search/ активна вкладка «Запрос»; на этой вкладке динамически создается формируемый пользователем запрос. На рис. 2 представлен процесс выбора единицы поиска. При помощи нажатия на одну из

верхних кнопок выбирается коммуникативный канал (вокальный, окуломоторный или мануальный), после чего отображается список доступных для данного канала единиц. На скриншоте показано состояние вкладки «Запрос» после выбора мануального канала: соответствующая кнопка выделена цветовой заливкой, ниже выведен список единиц мануальной жестикуляции, сгруппированный по уровням сегментации.

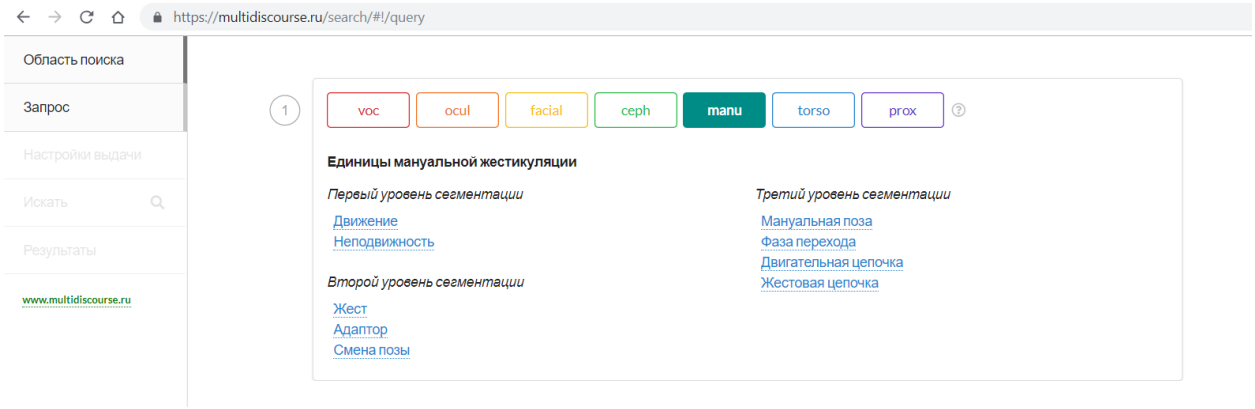

Рис. 2. Выбор типа единицы поиска во вкладке «Запрос»

После выбора единицы конкретного типа пользователь может указать ее дополнительные свойства. На рис. 3 и 4 показана форма редактирования свойств мануального жеста.

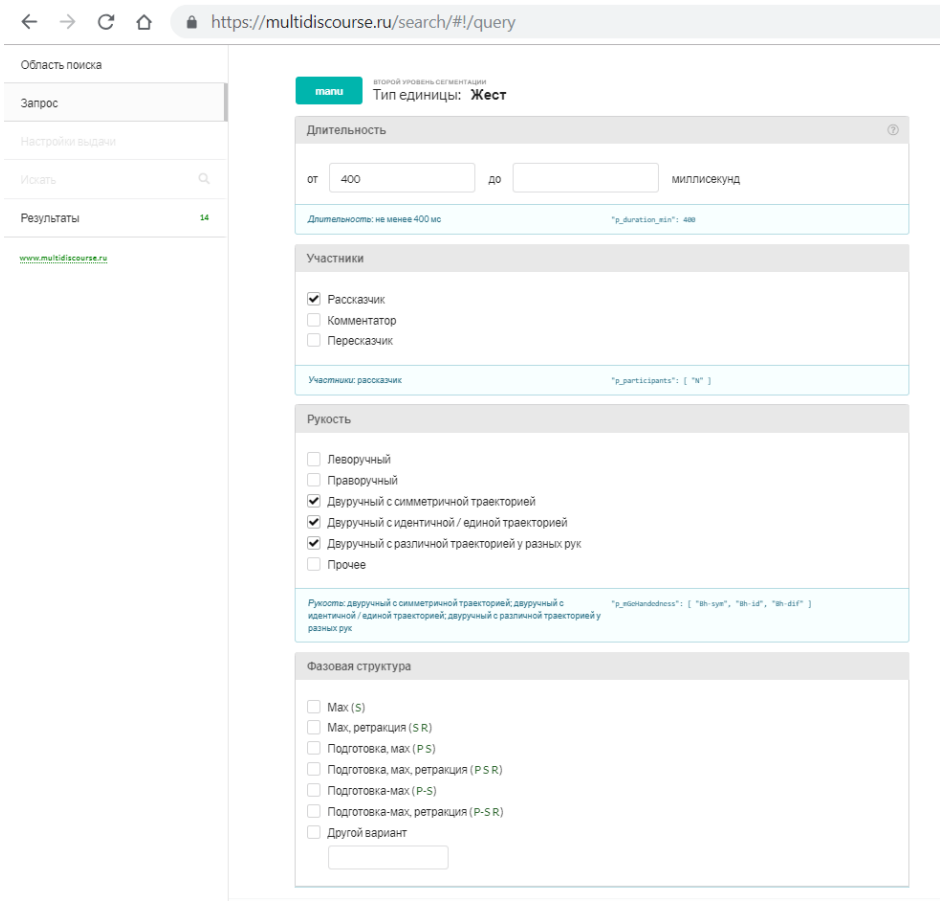

Рис. 3. Редактирование свойств мануального жеста (верхняя часть формы)

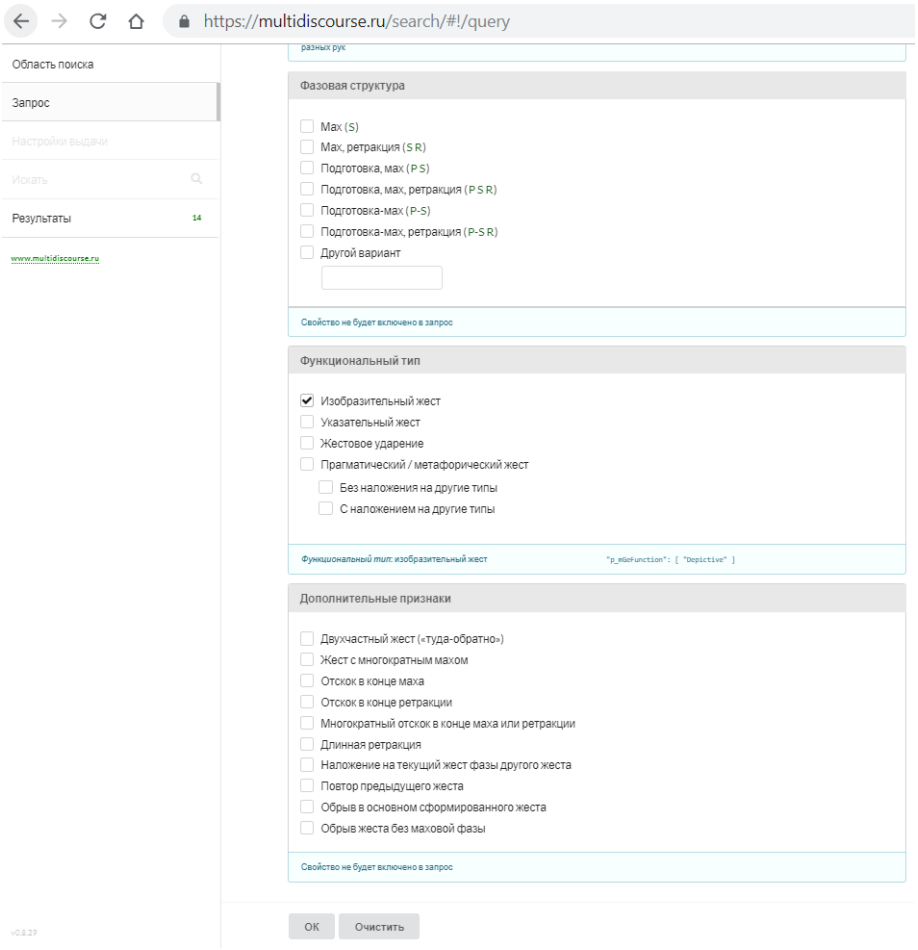

Рис. 4. Редактирование свойств мануального жеста (нижняя часть формы)

Для жеста указаны следующие свойства: его продолжительность не должна быть менее 400 миллисекунд, он должен быть реализован Рассказчиком посредством обеих рук и выполнять изобразительную функцию.

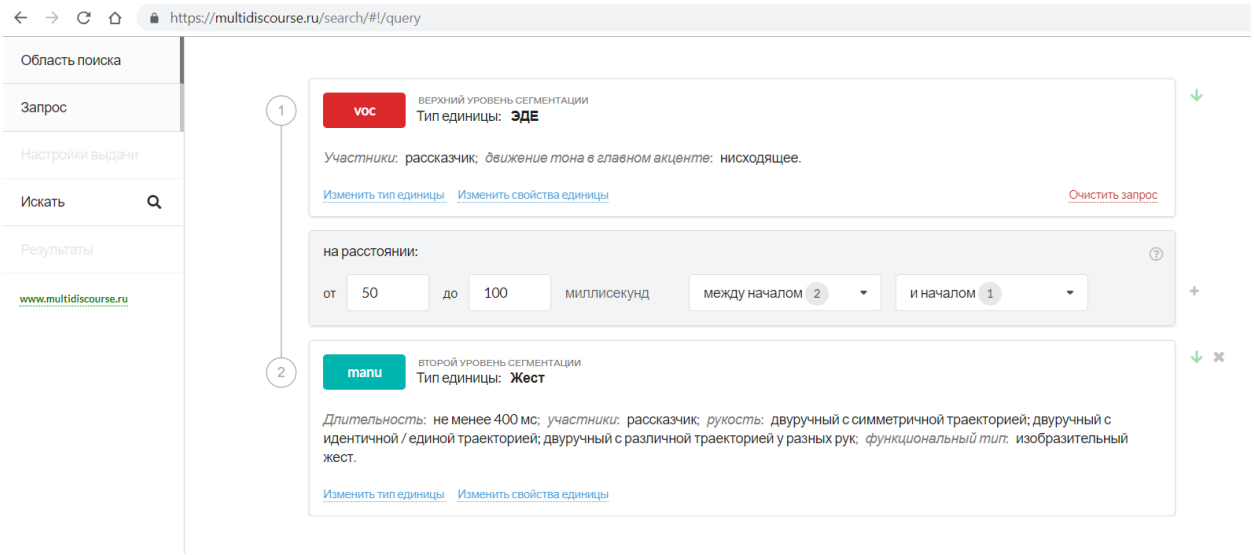

Рис. 5. Общая структура сложного запроса

Если в запрос входит более одной единицы, необходимо указать, какими временны́ми отношениями связаны эти единицы. На рис. 5 показана структура запроса, состоящего из двух единиц: жеста со свойствами, указанными на рис. 3 и 4, и элементарной

дискурсивной единицы (ЭДЕ; см. Кибрик, Подлесская (ред.) 2009), произносимой Рассказчиком и имеющей нисходящее движение тона в главном акценте. При этом начало реализации жеста должно запаздывать по отношению к началу произнесения ЭДЕ и находиться от него на расстоянии от 50 до 100 миллисекунд. Система позволяет формировать и более разветвленные запросы, в которых единицы связаны между собой не только последовательно, но и параллельно.

Поиск по запросу запускается по нажатии кнопки «Искать» в левом меню. Во вкладке «Результаты» отображается список найденных контекстов. При нажатии на каждый контекст проигрывается соответствующий фрагмент нужного видеофайла. На рис. 6 представлен скриншот выдачи результатов по запросу, рассмотренному на рис. 5. В видеоплеере зафиксирован жест Рассказчицы, начинающийся после начала произнесения ЭДЕ *и трёт себе ногу* (третий сверху контекст).

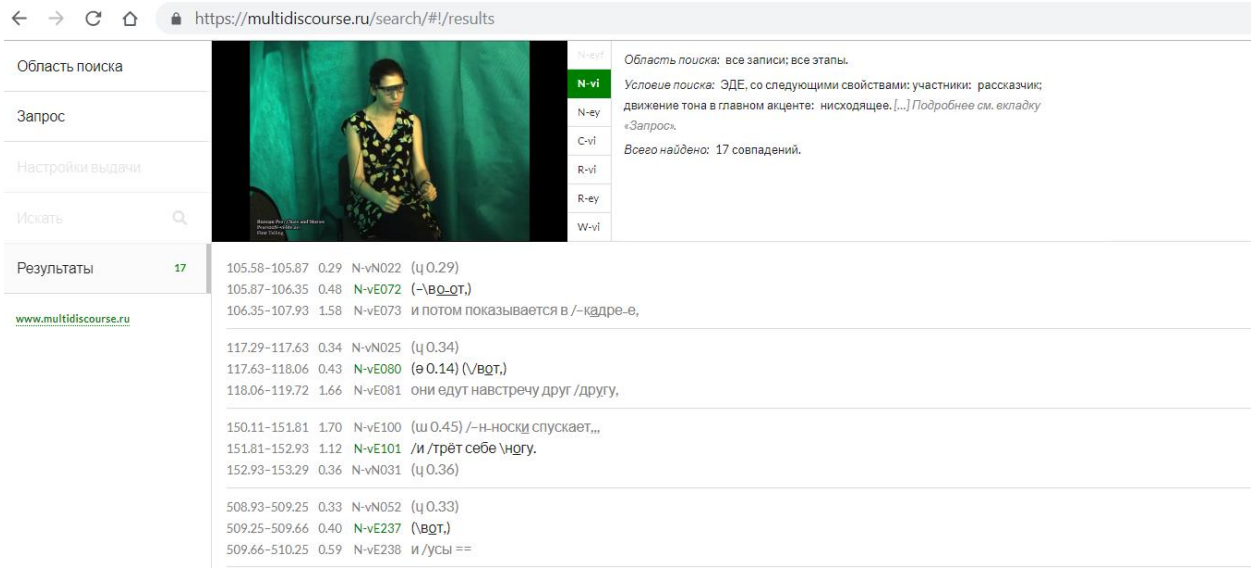

Рис. 6. Образец выдачи результатов по запросу

4. Заключение

В докладе представлены основные технические и содержательные характеристики системы поиска по мультиканальному корпусу, функционирующей на сайте https://multidiscourse.ru/. Описано соотношение клиентской и серверной частей системы. На конкретном сценарии продемонстрированы базовые возможности пользовательского интерфейса.

Литература

Chafe W. (ed.) (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Ablex, Norwood, NJ.

Fedorova O. V. (2017), Distribution of the interlocutors' visual attention in natural communication: 50 years later [Raspredelenie zritel'nogo vnimanija sobesednikov v estestvennoj kommunikacii: 50 let spustja], E. V. Pečenkova, M. V. Falikman (eds.), Cognitive science in Moscow: New studies [Kognitivnaja nauka v Moskve: novye issledovanija. Materialy konferencii 15 ijunja 2017 g], BukiVedi, Moscow, pp. 370–375.

Hellwig B., Hulsbosch M., Somasundaram A., Tacchetti M., Geerts J. (2018), ELAN — Linguistic Annotator: version 5.4. Manual updated on 2018-12-05, available at: https://www.mpi.nl/corpus/manuals/manual-elan.pdf

Kibrik A. A. (2018), Russian multichannel discourse. Part I. Setting up the problem [Russkij mul'tikanal'nyj diskurs. Čast' I. Postanovka problemy], Psixologiceskij žurnal, Vol. 39 (1), pp. 70–80.

Kibrik A.A., Podlesskaja V.I. (eds.), (2009), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovidenijax: korpusnoe issledovanie russkogo ustnogo diskursa]. Moscow: Jazyki slavjanskix kul'tur.

Korotaev N. A. (2019), "Russian Pear Chats and Stories": Vocal annotation guide. Version 10.01.2019, available at: http://multidiscourse.ru/data/ann/pears_vocal_annotation_en.pdf

Korotaev N. A., Evdokimova A. A., Litvinenko A. O., Nikolaeva Ju. V., Sukhova N. V. (2018). Multichannel annotation in ELAN: Vocal, oculomotor, cephalic, and manual channels. Version 14.12.2018, available at: http://multidiscourse.ru/data/ann/pears_multichannel_annotation_eng.pdf

Litvinenko A. O., Nikolaeva Ju. V., Kibrik A. A. (2017), Annotation of Russian manual gestures: Theoretical and practical issues [Annotirovanie russkix manual'nyx žestov: teoretičeskie i praktičeskie voprosy], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017" [Komp'juternaja Lingvistika i Intellektual'nye Texnologii: Trudy Meždunarodnoj Konferencii "Dialog 2017"], RGGU, Moscow, pp. 255–268.