***Н. А. Коротаев***

**Поиск в мультиканальном корпусе: содержательные задачи и техническая реализация**

**Аннотация.** В докладе представлены основные возможности поисковой системы по корпусу «Рассказы и разговоры о грушах», доступной на сайте <https://multidiscourse.ru/search/>. Корпус состоит из аудио- и видеозаписей однотипных коммуникативных сессий, снабженных разметкой вокального и кинетического поведения. Обсуждаются конкретные исследовательские вопросы, решение которых может быть облегчено благодаря наличию разрабатываемой поисковой системы.

**Ключевые слова.** Мультиканальный дискурс, корпуса устной речи, разработка поисковых движков, диалог и монолог

## 1. Вводные замечания

Одна из насущных задач современной корпусной лингвистики – это создание и использование ресурсов, которые позволяют изучать язык в ситуациях естественной коммуникации; см., в частности [Kotov, Budyanskaya 2012; Богданова-Бегларян (ред.) 2013; Joo et al. 2017]. Базовой формой такого использования языка является устная речь, при этом только речью естественная коммуникация не ограничивается: общаясь, мы также используем жесты, мимику, движения глаз, положения тела и проч. Подход к анализу языковой деятельности, в котором наряду с речью учитывается и вклад других коммуникативных ресурсов, принято называть мультимодальным, или мультиканальным [Müller et al. (eds) 2013]. Создание мультиканальных корпусов связано с несколькими проблемами. Приведем две из них, наиболее существенные в контексте данного доклада. Во-первых, крайне трудоемкой является содержательная разметка мультиканального дискурса: здесь (почти) нет успешных автоматизированных решений, поэтому приходится во многом полагаться на ручное аннотирование. Во-вторых, нетривиален вопрос о способах извлечения и представления размеченных данных – особенно в веб-интерфейсе. В этом докладе мы кратко рассмотрим эти вопросы на примере корпуса «Рассказы и разговоры о грушах». Дальнейшее изложение построено следующим образом. В разделе 2 приводятся основные данные о корпусе и обсуждается принципы разметки. В разделе 3 описывается веб-интерфейс поиска по корпусу. В разделе 4 перечисляются некоторые исследовательские вопросы, решение которых, на наш взгляд, существенно облегчается благодаря наличию основанной на размеченных данных поисковой системы.

## 2. «Рассказы и разговоры о грушах»: общие принципы разметки

Корпус «Рассказы и разговоры о грушах» (RUPEX; <https://multidiscourse.ru/>) был записан в 2015 и 2017 годах. Всего было получено 40 записей, имеющих единую структуру: участники всех записей в заранее определенном порядке рассказывают и обсуждают содержание «Фильма о грушах», созданного в 1970-е годы группой под руководством У. Чейфа [Chafe (ed.) 1980]. Поведение участников фиксировалось посредством петличных микрофонов, фронтальных видеокамер и видеокамеры общего плана, у двух участников в каждой записи при помощи портативных айтрекеров также регистрировалась глазодвигательная активность. Общий объем аудио- и видеоматериала составляет около 15 часов; подробнее о корпусе см. [Kibrik, Fedorova 2018].

Разметка записей корпуса выполняется независимо для отдельных коммуникативных каналов: вокального (вербальный и просодический компоненты), канала мануальной жестикуляции, канала глазодвигательной активности и проч. Ниже приведен фрагмент текстовой аннотации (транскрипта) вокального канала одной из записей. Транскрипт выполнен в соответствии с принципами, изложенными в [Kibrik et al. 2020]. Нумерованные строки транскрипта соответствуют элементарным дискурсивным единицам (ЭДЕ) и паузам между ними. Под ЭДЕ понимается минимальный шаг в развитии устного дискурса, соотносимый с единым когнитивным усилием и выделяемый в потоке речи на основании набора просодических характеристик. В скобках указана длительность абсолютных и заполненных пауз; при помощи слэшей и стрелок отображаются движения частоты основного тона в акцентированных словоформах; пунктуационные знаки в конце строк указывают на иллокутивно-фазовое значение ЭДЕ: дефолтная незавершенность (запятая), незавершенность в контексте неполноты информации (три запятых), завершение сообщения (точка) и др.

1. Фрагмент вокального транскрипта записи pears04

|  |  |
| --- | --- |
| R-vE409 | и /тут мимо проезжает \↑девочка, |
| pR-305 | (0.27) |
| R-vE410 | тоже на /→велосипеде,,, |
| pR-306 | (0.06) |
| R-vE411 | –ну уже на \таком˗м || (0.17) \женском /→велосипедике,,, |
| pR-307 | (0.60) |
| R-vE412 | (ə 0.48) она с \↑косичками, |
| R-vE413 | с \двумя \↑косичками, |
| pR-308 | (0.27) |
| R-vN065 | {sm 0.05} |
| R-vE414 | (\–во˗от,) |
| R-vE415 | и /проезжает \мимо. |

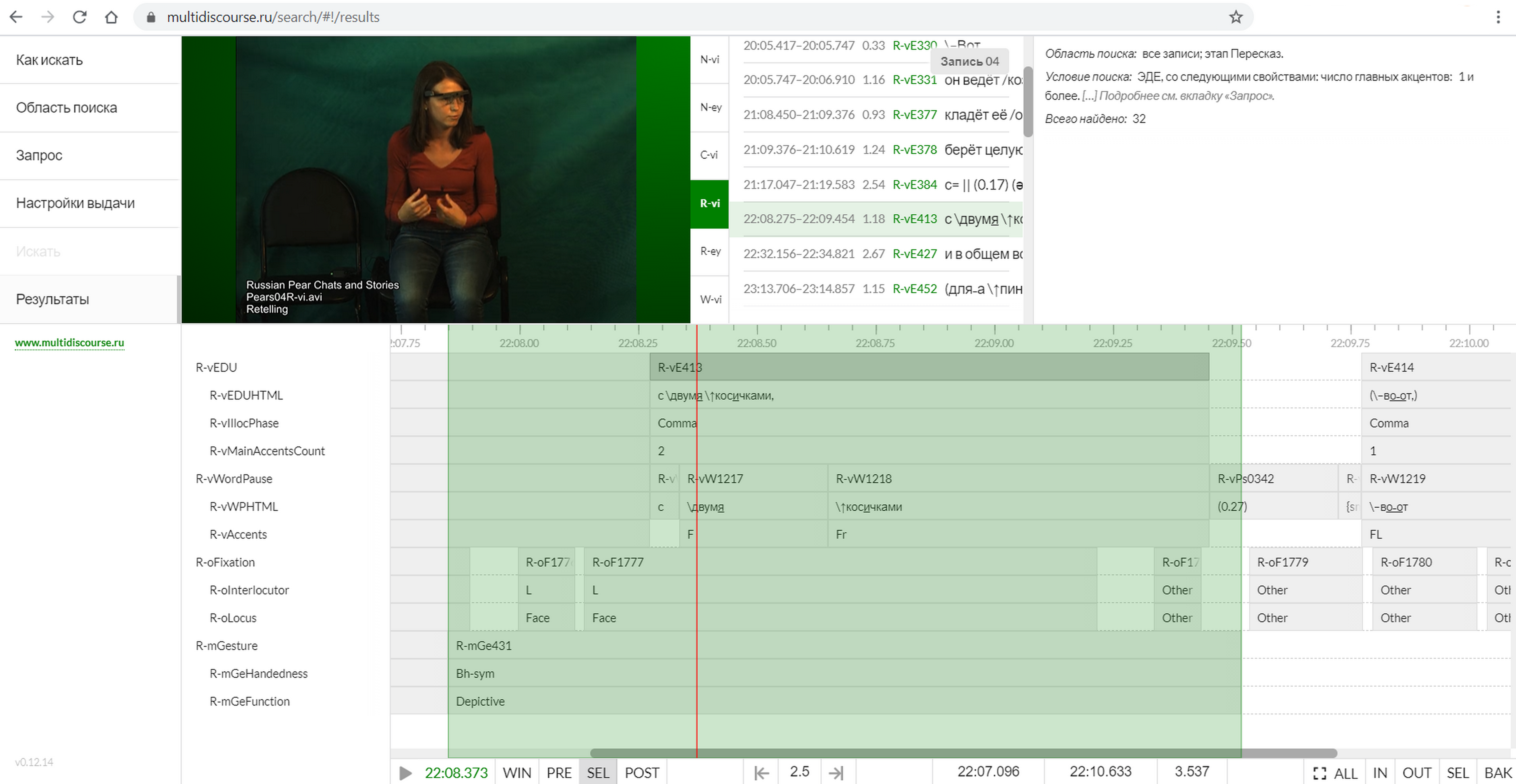
Разумеется, текстовый формат аннотации пригоден только для вокального канала. Прочие каналы размечаются в многоуровневом формате, к которому – для получения единой мультиканальной аннотации – в дальнейшем приводятся и речевые данные. Для хранения единой разметки используется программа ELAN [Hellwig et al. 2018], подробнее о принципах мультиканальной аннотации в корпусе см. [Korotaev et al. to appear].

## 3. Интерфейс онлайн-поиска[[1]](#footnote-1)

При наличии ELAN, разметок и медифайлов поиск можно выполнять на локальной машине непосредственно по единой мультиканальной аннотации. Однако в этом решении есть ряд очевидных ограничений, обойти которые позволяет браузерная реализация. При ее разработке мы исходили из следующих положений. Во-первых, поиск должен работать в браузере, без необходимости устанавливать дополнительное программное обеспечивание и скачивать «тяжелые» медиафайлы. Во-вторых, он должен обладать достаточно понятным интерфейсом – доступным в том числе и для тех пользователей, которые не изучили сложную схему мультиканальной аннотации. В-третьих, при отображении результатов поиска необходимо отображать видеофайлы и ассоциированную с ними содержательную разметку – причем именно те ее аспекты, которые релевантны для конкретного запроса и / или конкретного пользователя.

Разрабатываемая поисковая система доступна по адресу <https://multidiscourse.ru/search/>. В ней используется SQL-реплика мультиканальных аннотаций формата ELAN; клиентская часть представляет собой одностраничное приложение на языке JavaScript, спроектированное с использованием шаблона MVVM; серверная часть выполнена на языке Java. На апрель 2021 года поиск производится по трем размеченным записям корпуса; пользователям доступны следующие функции (краткое описание основных возможностей также содержится во вкладке «Как искать»):

* Ограничение области поиска отдельными записями и / или этапами записей (вкладка «Область поиска»).
* Составление поисковых запросов: как простых, т.е. содержащих в себе одну поисковую единицу, так и сложных (вкладка «Запрос»). В простых запросах пользователю достаточно указать тип искомой единицы и при необходимости указать ее свойства, выбрав требуемые значения в именованных полях: общих для единиц всех типов (длительность, принадлежность участнику с конкретной ролью в записи) или специфических для конкретного типа (наличие и тип акцента для слов; рукость для мануальных жестов; направления взгляда для глазных фиксаций и др.). Каждое поле снабжено краткой справкой. Сложные запросы включают в себе несколько поисковых единиц, связанных между собой ограничениями на расстояние и (опционально) на (не)совпадение участников. При формировании сложного запроса происходит динамическое перестроение «карты запроса».
* Просмотр результатов поиска (вкладка «Результаты»). По умолчанию результаты представлены в виде простого текстового списка; при нажатии на элемент выдачи отображается фрагмент многоуровневой разметки и плеер с выбором из нескольких синхронизированных видеофайлов.
* Настройка слоев для отображения в расширенном режиме просмотра результатов (вкладка «Настройки выдачи»).



*Рис. 1.* **Расширенный режим просмотра результатов по поисковому запросу на сайте https://multidiscourse.ru/search**/

На рис. 1 показан расширенный режим просмотра результатов по сложному поисковому запросу, включающему в себя три единицы:

1. ЭДЕ участника X, содержащие в себе не менее одного акцентированного слова;
2. Фиксации взгляда того же участника, направленные на лицо кого-либо из других участников записи и пересекающиеся с ЭДЕ (i);
3. Изобразительные жесты того же участника, пересекающиеся с ЭДЕ (i).

Область поиска ограничена этапом пересказа. Для просмотра выбран фрагмент, соответствующей ЭДЕ R-vE413 из транскрипта (1). Видно, что говорящая смотрит на лицо слушателя и сопровождает произнесение ЭДЕ *с двумя косичками* изобразительным жестом, выполняемым двумя руками с симметричной траекторией.

## 4. Примеры использования и дальнейшие перспективы

Приведем некоторые примеры исследовательских задач, решение которых становится возможным – или существенно облегчается – благодаря наличию мультиканальной аннотации и возможности поиска по аннотированным данным. (Список далеко не полный и носит сугубо иллюстративный характер.)

1) Составление «жестикуляционного портрета» лексических единиц. Например, если задать в поиске слова, которые указывают на регулярно встречающихся в корпусе референтов, и включить в настройках выдачи отображение мануальных жестов, можно получить список изобразительных кинетических техник, которые участники записей используют для сопровождения речевого описания этих объектов. Аналогичный вопрос можно поставить и для единиц, не обладающих предметным значением, – например, дискурсивных маркеров *вот* и *ну*. Так, согласно данным размеченного подкорпуса, из 372 вхождений самых частотных в корпусе существительных с предметным значением (*груша, мальчик, корзина, шляпа, велосипед*), произносимых с акцентом, 267 (72%) пересекаются с тем или иным мануальным жестом, реализуемым тем же участником. (В качестве критерия пересечения было задано составное условие вида «начало жеста – от -3000 до 200 от начала слова, конец жеста – от 200 до 5000 от начала слова».) В то же время для полноударных *вот* и *ну* аналогичное пересечение с жестами наблюдается лишь в 49% случаев (100 из 206). При этом из 100 жестов, сопровождающих *вот* и *ну*, 62 выполняют прагматическую функцию, т.е. преимущественно нацелены на установление контакта между участниками коммуникации; тогда как из 267 жестов, сопровождающих частотные слова с предметным значением, прагматическую функцию реализуют только 100 (37%). Таким образом, предварительный анализ подобного рода показывает, что различия в лексической семантике слов в определенной степени поддерживаются сопутствующей мануальной жестикуляцией.

2) Интегральное описание мультиканального поведения участников в заданной коммуникативной ситуации. Примером тут могут служить диалогические обмены, инициируемые вопросом или запросом на подтверждение («полуутверждением»). Указав соответствующие значения в поле «Иллокутивно-фазовое значение ЭДЕ», можно получить предварительный список релевантных контекстов. Впоследствии поисковый запрос можно усложнять, вводя условия, относящиеся к глазодвигательной активности участников (так, инициатор обмена регулярно поддерживает речевой запрос установлением зрительного контакта с тем участником, от которого он ожидает реакции), структуре коммуникативного обмена (инициатор обычно ожидает реакции на свой запрос, а получив ее от назначенного участника, осуществляет завершающую реплику подтверждения, которая при этом может сопровождаться или заменяться кивком головой), жестикуляции и проч. Как указано в разделе 3, все полученные результаты поиска можно просматривать непосредственно в браузере, выбирая нужную камеру и при необходимости редактируя набор отображаемых слоев разметки.

3) Сопоставительный анализ поведения участников на монологических vs. диалогических этапах записей. Поскольку одни и те же участники задействованы как при (пере)рассказах содержания фильма, так и при его обсуждении, материал корпуса располагает к подобного рода контрастивным исследованиям. Так, манипулируя значениями в поле «Этапы записей» на вкладе «Область поиска», можно, в частности, обнаружить, что у одних и тех же участников в диалоге по сравнению с монологом значительно падает частотность заполненных, а также внутренних (т.е. располагающихся внутри ЭДЕ) пауз. Это наблюдение подтверждает сложившееся в литературе представление о большей сложности в порождении монологического дискурса и позволяет дать ему количественную оценку.

Мы полагаем, что при увеличении объема размеченных данных разрабатываемая система поиска позволит получать новые результаты об устройстве устной и – шире – мультиканальной коммуникации. Кроме того, мы надеемся, что предложенные технические решения могут заинтересовать разработчиков других мультиканальных корпусов и поспособствовать созданию единого формата мультиканальной разметки.

## Литература

1. Богданова-Бегларян Н. В. (ред). (2013), Звуковой корпус как материал для анализа русской речи. Коллективная монография. Часть 1. Чтение. Пересказ. Описание. СПб, Филологический факультет СПбГУ.

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

3. Joo J., Steen F., Turner M. (2017), Red Hen Lab: Dataset and Tools for Multimodal Human Communication Research, *KI - Künstliche Intelligenz*, 31, pp. 357-361.

4. Hellwig B. et al.(2018), ELAN — Linguistic Annotator: version 5.4. Manual updated on 2018-12-05, available at URL: <https://www.mpi.nl/corpus/manuals/manual-elan.pdf> (дата обращения 01.04.2021)

5. Kibrik A. A., Fedorova O. V. (2018), Language production and comprehension in face-to-face multichannel communication. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue*”, 24(17), pp. 305–316.

6. Kibrik A. A., Korotaev N. A., Podlesskaya V. I. (2020), Russian spoken discourse: Local structure and prosody, S. Izre’el et al. (eds.), In search of basic units of spoken language: A corpus-driven approach. John Benjamins, pp. 37–76.

7. Korotaev N. A. et al.(to appear), Multichannel annotation scheme, O. V. Fedorova, A. A. Kibrik (eds), The MCD handbook: A practical guide to annotating multichannel discourse, Institute of Linguistics RAS.

8. Kotov A. A., Budyanskaya E. M. (2012), The Russian emotional corpus: Communication in natural emotional situations, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue”*, 11(18), pp. 296–306.

9. Müller C. et al. (eds.) (2013), Body – Language – Communication: An international handbook on multimodality in human interaction (Vol. 1), De Gruyter Mouton.

## References

1. Bogdanova-Beglarian N. V. (ed.). (2013), Zvukovoj korpus kak material dlja analiza russkoj reči. Kollektivnaja monografija. Čast’ 1. Čtenie. Pereskaz. Opisanie. St. Petersburg, SPbU Faculty of Philology.

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

3. Joo J., Steen F., Turner M. (2017), Red Hen Lab: Dataset and Tools for Multimodal Human Communication Research, *KI - Künstliche Intelligenz*, 31, pp. 357-361.

4. Hellwig B. et al.(2018), ELAN — Linguistic Annotator: version 5.4. Manual updated on 2018-12-05, available at URL: <https://www.mpi.nl/corpus/manuals/manual-elan.pdf> (last access 01.04.2021)

5. Kibrik A. A., Fedorova O. V. (2018), Language production and comprehension in face-to-face multichannel communication. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue*”, 24(17), pp. 305–316.

6. Kibrik A. A., Korotaev N. A., Podlesskaya V. I. (2020), Russian spoken discourse: Local structure and prosody, S. Izre’el et al. (eds.), In search of basic units of spoken language: A corpus-driven approach. John Benjamins, pp. 37–76.

7. Korotaev N. A. et al.(to appear), Multichannel annotation scheme, O. V. Fedorova, A. A. Kibrik (eds), The MCD handbook: A practical guide to annotating multichannel discourse, Institute of Linguistics RAS.

8. Kotov A. A., Budyanskaya E. M. (2012), The Russian emotional corpus: Communication in natural emotional situations, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue”*, 11(18), pp. 296–306.

9. Müller C. et al. (eds.) (2013), Body – Language – Communication: An international handbook on multimodality in human interaction (Vol. 1), De Gruyter Mouton.

**Searching in a multichannel corpus:   
Research questions and technical solutions**

N. Korotaev

**Abstract.** The paper presents the search system developed for the “Russian Pear Chats & Stories” (RUPEX) corpus and discusses its main features. RUPEX is a multichannel corpus that includes audio and video recorded communication sessions and their annotations. The search system, which is based on the integrated multichannel annotation, is a one-page browser application that allows users to run queries and view results in a multi-layered format together with relevant video fragments. Several phenomena intrinsic to natural communication can be extensively studied using this research tool.

**Keywords.** Multichannel discourse, spoken corpora, search engines, dialogue and monologue.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Коротаев Николай Алексеевич**

Российский государственный гуманитарный университет (Россия).

**Korotaev Nikolay**

Russian State University for the Humanities (Russia).

***E-mail: n\_korotaev@hotmail.com***

1. Серверная часть поисковой системы разработана Г. Б. Добровым, клиентская – А. Н. Хитровым. [↑](#footnote-ref-1)