Psychology. Journal of the Higher School of Economics.
2018. Vol. 15. N 2. P. 191–200. DOI: 10.17323/1813-8918-2018-2-191-200

AN EMPIRICAL STUDY OF MULTICHANNEL COMMUNICATION: RUSSIAN PEAR CHATS AND STORIES

A.A. KIBRIK[a,b], O.V. FEDOROVA[b,a]

[a] *The Institute of Linguistics of the Russian Academy of Sciences, 1/12 Bolshoy Kislovsky Lane, Moscow, 125009, Russian Federation*

[b] *Lomonosov Moscow State University, 1 Leninskie Gory, Moscow, 119991, Russian Federation*

Abstract

This paper addresses language in its most natural form — in the form of spoken multichannel discourse. It includes the verbal component, prosody, eye gaze, as well as the different kinetic aspects of communication — facial, head, hand and torso gestures. To explore natural multichannel discourse as is, we created a resource "Russian Pear Chats and Stories". The resource includes 40 sessions with 160 Russian native speakers aged 18–36, 60 men and 100 women; it consists of 15 hours of recording and about 170,000 words. This paper details how the corpus is created and how it can be used. First, we provide an overview of the methodology of multimodality and multichannel corpora. Then we describe the properties of our resource — the data collection set up, the recording software, types of annotation, as well as some avenues of (future) research, including: prosody as an interface between the vocal and gestural channels, specific nature and degree of coordination between manual gestures and elementary discourse units, individual variation and the "portrait" methodology, language production and comprehension in face-to-face communication, and visual attention in natural communication. In its current version, the corpus is available to the scientific community at the project website multidiscourse.ru (in Russian).

Keywords: multimodality, multichannel discourse, corpus creation, prosody, gestures, eye gaze, annotation.

Introduction. Multichannel communication and multichannel corpora

In face-to-face communication, interlocutors combine verbal structure, prosody, eye gaze, as well as facial, head, hand and torso gestures to produce integrated discourse. All of these communication channels are employed simultaneously and in conjunction with each other. Therefore, everyday human communication is a multichannel (multimodal)[1] process (Kress, 2002; McNeill, 2005; Kibrik, 2010;

This study is supported by Russian Science Foundation (grant No 14-18-03819 "Language as is: Russian multimodal discourse").

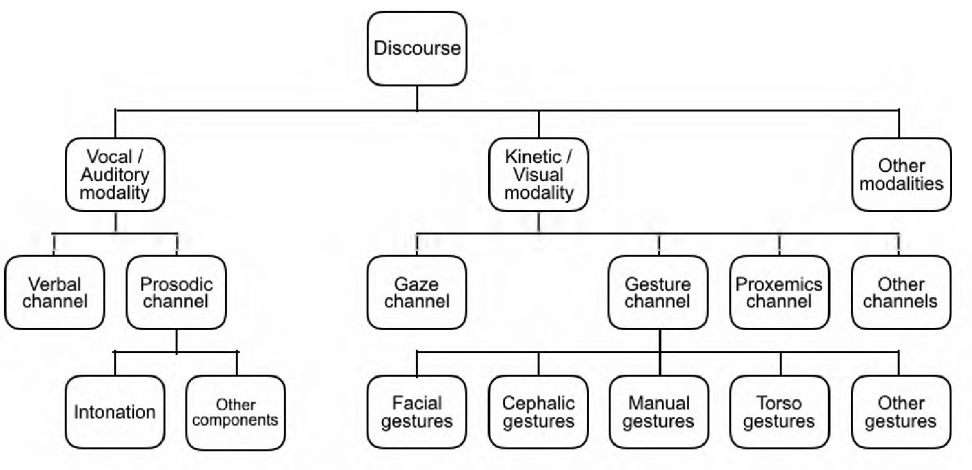
[1] In psychology and neurophysiology, modality is defined as affiliation of a signal with a particular sensory system. At present, the term "multimodal" is more common, but it is more precise to use the terms "multichannel" or "bimodal", since only two modalities — vocal (auditory) and kinetic (visual) — are predominantly studied, while the remaining modalities, for example, smell or touch, remain outside of consideration; however, see Mondada 2016 on studies in the domain of touch modality.

Loehr, 2012; Adolphs & Carter, 2013; Goldin-Meadow, 2014; Müller, Fricke, Cienki, & McNeill, 2014; Church, Alibali, & Kelly, 2017, inter alia), see Figure 1.

The terms "multimodal communication"[2] and "multimodal corpus" first appeared in the 1980s, cf. Taylor, 1989. A multimodal corpus is "an annotated collection of coordinated content on communication channels including speech, gaze, hand gesture and body language, and is generally based on recorded human behaviour" (Foster & Oberlander, 2007, p. 307–308). As compared to monomodal corpora that already have a substantial history and tradition, multimodal corpora are still in their incipient stage. There are at least three criteria that help to characterize a corpus: 1) its size; 2) naturalness of the data; 3) the goals set for the corpus. Judging by the published metadata, the largest multimodal corpus is the AMI Meeting Corpus, 100 hours long (Carletta, 2006); however, most of the information of this corpus is presented in the form of non-annotated video files. The naturalness of corpus data can be conveniently represented as a scale from strictly controlled experiments on the left side to unrestricted free communication on the right. The left side of the scale can be exemplified by the Czech Audio-Visual Speech corpus (Železný, Krňoul, Císař, & Matoušek, 2006) created for testing the system of speech recognition and including 25 hours of recordings of 65 participants who were instructed to read 200 sentences aloud. More natural data has been assembled in the Fruit Carts Corpus that contains 240 video recordings of 12 participants, each four to eight minutes long (Aist, Campana, Allen, Swift, &

Figure 1

Model of multichannel discourse



[2] To our knowledge, the term "multichannel" was first used in Cosnier & Brossard (1984): "c'est à l'époque contemporaine que la conception de ***la communication multicanale*** a été élargie, précisée et étayée par les réflexions et les travaux des éthologues, des anthropologues, des sociologues et des 'psy' (psychologues et psychiatres)." (p. 2–3).

Tanenhaus, 2012). Along this scale of naturalness, still more to the right is the English-language corpus D64, created for studies of everyday communication (Campbell, 2009), and the InSight Interaction Corpus consisting of 15 recorded face-to-face interactions 20 min long each (Brône & Oben, 2015). On the rightmost side of the scale are found corpora created in the tradition of Conversation Analysis, e.g. the corpus described in Mondada, 2014; see also the recent paper Mondada, 2016. As for the final criterion mentioned above, in accordance with Knight, 2011, p. 403, all the existing corpora are created with a particular research goal in mind and only address specific research questions, whereas no standard procedures of data collection, annotation, and exploration have yet been established.

The structure of the paper is as follows. In section 2 we describe the properties of our resource — the stimulus material, the data collection setup, participants and corpus size, and the recording software. Section 3 addresses different types of annotations. In section 4 we consider some avenues of (future) research.

Collecting the data

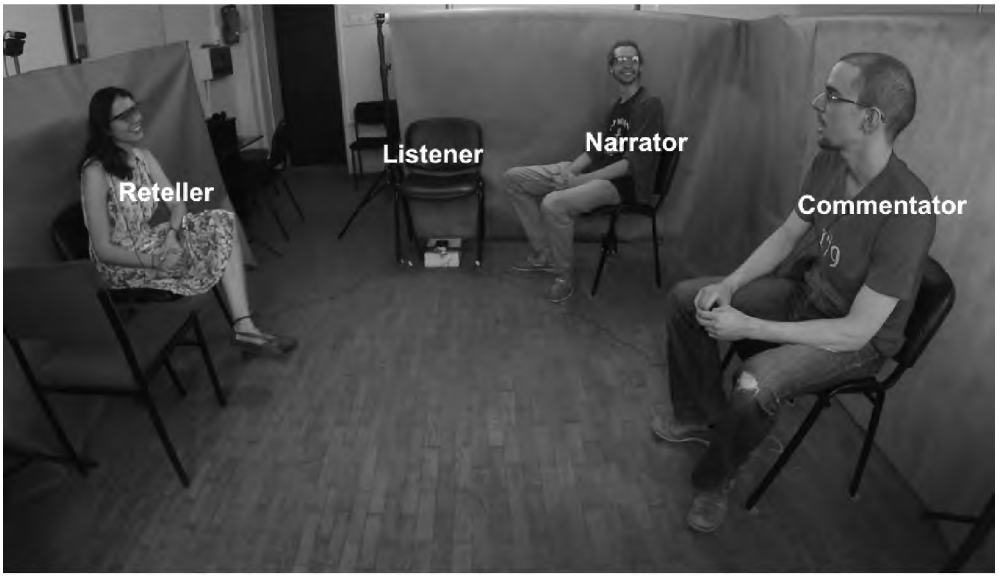
Stimulus material. We have used the well-known Pear Film (Chafe, 1980, pearstories.org) that has proved its efficiency in a variety of linguistic and cultural tasks. This six-minute film produced at the University of California at Berkeley was designed to elicit stories from speakers around the world. The film was constructed so that the scenes incline participants to describe landscapes, explain cause-effect relations, account for the characters' thoughts and emotions, and resolve ambiguities.

Data collection setup. We have developed a new experimental procedure. Each session lasted for about one hour and involved four participants with fixed roles: three main interlocutors — the Narrator, the Commentator, and the Reteller — and the Listener. At the very beginning the Narrator and the Commentator each watched the film on a personal computer trying to memorize the plot as precisely as possible. Then the main stages began. First, the Narrator told the Reteller about the plot of the film; this is a monologic stage — *first telling*. During the subsequent, interactive, stage —*conversation* — the Commentator added details and corrected the Narrator's story where necessary, and the Reteller checked her/his understanding of the plot, asking questions to both interlocutors. Then the Listener joined the group and another monologic stage — *retelling* — followed, during which the Reteller was retelling the plot of the film to the Listener. Finally, the Listener wrote down the content of the film (see Kibrik, 2018 for more detail). The data collection set up is depicted in Figure 2.

Participants and Corpus Size. The resource "Russian Pear Chats and Stories" consists of two parts. The first part collected in the summer of 2015 includes 24 sessions with 96 Russian native speakers aged 18–36, 34 men and 62 women; the overall duration is nine hours (the average length of a recording was 24 min) and the recordings contain 110,000 words. The second part collected in the summer 2017 includes 16 sessions with 64 Russian native speakers aged 18–36, 16 men and 48 women; the overall duration is six hours of recording (the average length of a

Figure 2

Data collection setup



recording was 21 min) and the recordings contain 60,000 words. Each session consists of ten synchronized media files: four audio files (three individual files of Narrator's, Commentator's, and Reteller's voices and one file of all vocal events recorded), three individual video files of Narrator's, Commentator's, and Reteller's kinetic activities, one video file from the cover shot camera, and two eye-tracker video files, from the Narrator's and the Reteller's viewpoints. Each set also includes an eye-tracker video file recorded while the Narrator was viewing the stimulus film (see Figure 3a).

Figure 3

Video scene, as recorded by a camera built into the eye-trackers, with superimposed marker of visual attention

a. Viewing the Pear film b. From the N's eye-tracker c. From the R's eye-tracker

Recording Software. We have used the following state of the art equipment:

(1) a professional ZOOMH6 Handy Recorder, which ensured automatic synchronization. The speech of each of the three main interlocutors was recorded at 96kHz and 24 bit with a lapel mic SONY ECM-88B, in the mono mode; the fourth recording was done with an inbuilt recorder, in the stereo mode;

(2) three individual industrial cameras JAI-GO-5000M, 100 fps, resolution 1392x1000, that made a frontal recording of each of the three participants; for further analysis of kinetic behavior it is important that these cameras create files in the mjpeg format that is free of interframe compression; the 100 fps frame rate allows the analysis with the precision of up to 10 msec, which is a prerequisite for accurate annotation of kinetic behavior;

(3) a wide angle camera GoPro Hero 4 used for cover shot, 50 fps (100 fps in 2017), resolution 2700×1500;

(4) two eye-trackers Tobii Glasses II Eye Tracker, sampling rate 50 Hz, video camera resolution 1920x1080. Tobii Glasses II have been in production since 2014. The eye-tracker provides two types of data: (i) video files produced by an inbuilt scene camera and (ii) data files representing eye movements. The screenshots in Figure 3 result from an overlay of video files from the scene camera and the gaze coordinates from the data files; the circles are generated by the eye-trackers and indicate the targets of interlocutors' gaze.

Annotations

Vocal Annotation. The vocal annotation follows the principles previously developed for spoken Russian discourse (see Kibrik & Podlesskaya, 2009 and spokencorpora.ru). The data was annotated using the Praat program (fon.hum.uva.nl/praat), in accordance with a vocal annotation scheme including temporal dynamics, absolute and filled pauses, segmentation into elementary discourse units (EDUs), accents, accelerated tempo, reduced pronunciation, lowered tonal register, etc.

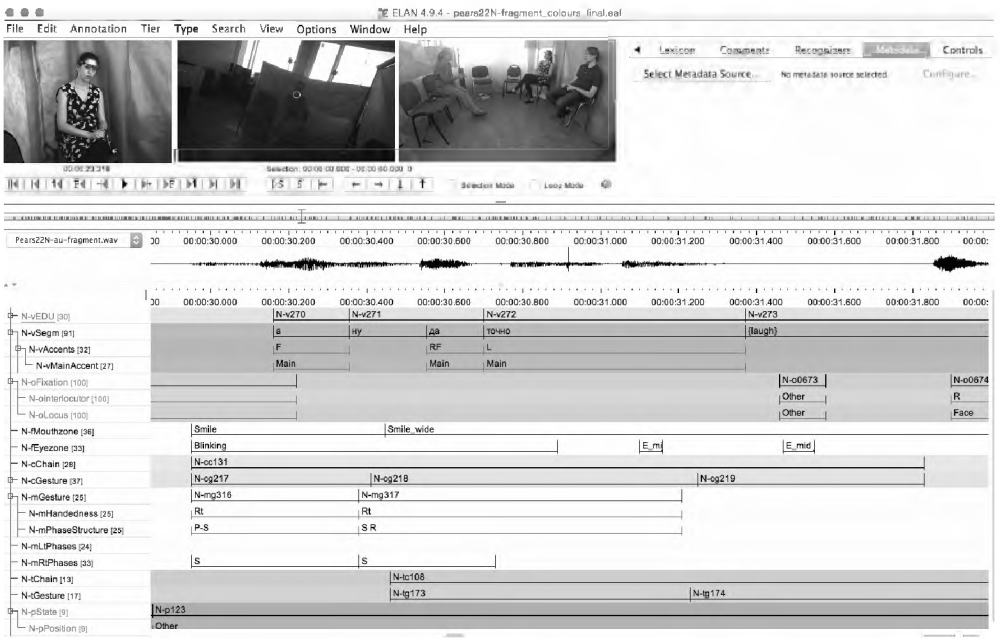
Annotation of Manual Gestures. For the transcription of the video data we used the annotation software ELAN (lat-mpi.eu/tools/elan/) and followed the annotation system developed in (Litvinenko, Nikolaeva, & Kibrik, 2017). We annotate gestural forms independently of speech (Bressem, 2013). We define the following layers for manual gestures, independently for each hand: gesture chains, gestures, gesture phases, handedness, self-adaptors (Ekman & Friesen, 1969), postures, and posture changes.

Annotation of Gaze. Gaze is coded for gaze target: (i) an Interlocutor (Narrator/Reteller, Commentator or Listener), further subdivided into face, hands, torso, and other; (ii) the Surroundings. The minimal fixation duration is 100 ms, i.e. a particpant' fixation on a target must last for at least 100 ms to be recognized as a gaze event (see Fedorova, 2017 for more detail).

Multilayer Annotation. Figure 4 provides an example of a full multichannel annotation, including the above discussed channels, as well as additional components of phonetic realization, facial expressions, torso gestures, and proxemics; see also multidiscourse.ru/annotation.

Figure 4

Multilayer annotation



Some avenues of research

In the course of the project we have created a multichannel resource of natural Russian communication that does not have direct analogs among the contemporary resources. It is created for a wide range of research goals. Some avenues of research that is being (or can be) conducted on the basis of the resource include, inter alia:

Prosody as an interface between the vocal and gestural channels (Kodzasov, 2009; Kibrik & Podlesskaya, 2009). There are many specific similarities between prosodic and gestural phenomena — tempo, acceleration/deceleration, intensity, emphasis on most prominent semantic elements, etc.

Specific nature and degree of coordination between manual gestures and EDUs. It has been shown in a number of studies that gesture onset usually precedes speech onset (McNeill, 1992; Loehr, 2012; Karpiński, Jarmołowicz-Nowikow, & Malisz, 2009). In order to verify this claim through our material, we developed an analytic method that allowed a more detailed study. According to our results, it is only less than a half of all gestures that are produced before the corresponding fragment of talk (Fedorova, Kibrik, Korotaev, Litvinenko, & Nikolaeva, 2016).

Individual variation and the "portrait" methodology. It is created for fine-grained annotation procedures, as well as for accurate statistical analyses of multichannel data:

a. ***Prosodic Portrait***, i.e. a range of speaker's prosodic characteristics: minimal and maximal F0 value, standard level of EDU onsets, target level of final and non-final fallings, target level of rises in a canonical comma intonation, etc. (see Kibrik & Podlesskaya, 2009);

b. ***Oculomotor Portrait*** involving the data of a summary quantity of fixations throughout the duration of a session; a summary duration of the fixations; mean, minimal, and maximal durations, as well as 25%, 50%, and 75% quantiles (Fedorova, 2017);

c. ***Gesticulation Portrait*** including (dis)inclination to stillness; (dis)inclination to self-adaptors; typical amplitude; typical velocity; preferences in gesture handedness; a summary number of manual gestures throughout a session; their summary duration; their mean, minimal, and maximal durations, as well as 25%, 50%, and 75% quantiles (Kibrik & Fedorova, 2018).

Language production and comprehension in face-to-face communication. In language comprehension research, description is usually done in terms of either auditory or visual modality; in case of language production research, either vocal or kinetic modality. During the process of face-to-face communication, however, each interlocutor performs the roles of speaker and listener simultaneously. For example, a speaker, while producing speech at a given moment, simultaneously monitors the listener's kinetic behavior (nods, gaze, and manual gestures).

Visual attention in natural communication. Most eye-tracking studies were accomplished in experimental settings (but cf. Kendon, 1967). In accordance with the evidence we have collected and analysed, eye gaze strategies in natural communication fall into three types: general (for example, longer fixations on face (1 to 2 s) compared to fixations on hands (100 to 250 ms)); context-dependent (in interaction, the speaker's fixations on surroundings are rarer than in monologue); and individual (see Fedorova, 2017 for more detail).

References

Adolphs, S., & Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal.* New York: Routledge.

Aist, G., Campana, E., Allen, J., Swift, M., & Tanenhaus, M. K. (2012). Fruit carts: A domain and corpus for research in dialogue systems and psycholinguistics. *Computational Linguistics, 38*(3), 469–478.

Bressem, J. (2013). A linguistic perspective on the notation of form features in gestures. In C. Müller et al. (Eds.), *Body — Language — Communication: An international handbook on multimodality in human interaction* (Vol. 1, p. 1079–1098). Berlin/Boston: De Gruyter Mouton.

Brône, G., & Oben, B. (2015). InSight Interaction: a multimodal and multifocal dialogue corpus. *Language Resources and Evaluation, 49*(1), 195–214.

Campbell, N. (2009). Tools and resources for visualising conversational-speech Interaction. In M. Kipp et al. (Eds.), *Multimodal corpora: From models of natural interaction to systems and applications* (pp. 176–188). Heidelberg: Springer.

Carletta, J. (2006). Announcing the AMI meeting corpus. *The ELRA Newsletter, 11*(1), 3–5.

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

Church, R. B., Alibali, M. W., & Kelly, S. D. (Eds.). (2017). *Why gesture? How the hands function in speaking, thinking and communicating*. Amsterdam: John Benjamins Publishing.

Cosnier, J., & Brossard, A. (1984). *La communication non verbale* [Non-verbal communication]. Neuchâtel: Delachaux et Niestlé. (in French)

Ekman, F., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*(1), 49–98.

Fedorova, O. V. (2017). Raspredelenie zritel'nogo vnimaniya sobesednikov v estestvennoi kommunikatsii: 50 let spustya [Distribution of the interlocutors' visual attention in natural communication: 50 years later]. In E. V. Pechenkova & M. V. Falikman (Eds.), *Kognitivnaya nauka v Moskve: novye issledovaniya. Materialy konferentsii* [Cognitive science in Moscow: new research. Proceedings of the conference] (pp. 370–375). Moscow: BukiVedi/IPPiP. (in Russian)

Fedorova, O. V., Kibrik, A. A., Korotaev, N. A., Litvinenko, A. O., & Nikolaeva, Yu. V. (2016). Temporal coordination between gestural and speech units in multimodal communication [Vremennaya koordinatsiya mezhdu zhestovymi i rechevymi edinitsami v mul'timodal'noy kommunikatsii]. In *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnoi konferentsii "Dialog 2016"* [Computational linguistics and intellectual technologies: Proceedings of the International conference "Dialogue 2016"] (pp. 159–170). Moscow: RGGU. (in Russian)

Foster, M. E., & Oberlander, J. (2007). Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation, 41*(3/4), 305–323.

Goldin-Meadow, S. (2014). Widening the lens: What the manual modality reveals about language, learning, and cognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 369*(1651), 20130295.

Karpiński, M., Jarmołowicz-Nowikow, E., & Malisz, Z. (2009). Aspects of gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogues. *Speech and Language Technology, 11*, 113–122.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, 22–63.

Kibrik, A. A. (2010). Mul'timodal'naya lingvistika [Multimodal linguistics]. In Yu. I. Aleksandrov & V. D. Solov'ev (Eds.), *Kognitivnye issledovaniya* [Cognitive studies] (Iss. IV, pp. 134–152). Moscow: Institute of Psychology of RAS. (in Russian)

Kibrik, A. A. (2018). Russkii mul'tikanal'nyi diskurs. Chast' 1. Postanovka problemy [Russian multichannel discourse. Part I. Setting up the problem]. *Psikhologicheskii Zhurnal, 39*(1), 70–80. (in Russian)

Kibrik, A. A., & Fedorova, O. V. (2018). A "Portrait" approach to multichannel discourse. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 7–12 May 2018, Miyazaki (Japan)*.

Kibrik, A. A., & Podlesskaya, V. I. (Eds.). (2009). *Rasskazy o snovideniyakh: Korpusnoe issledovanie ustnogo russkogo diskursa* [Night dream stories: A corpus study of spoken Russian discourse]. Moscow: Yazyki slavyanskikh kul'tur. (in Russian)

Knight, D. (2011). *Multimodality and active listenership: A corpus approach*. London: Bloomsbury.

Kodzasov, S. V. (2009). *Issledovaniya v oblasti psikhologii* [Studies in the field of Russian prosody]. Moscow: Yazyki slavyanskikh kul'tur. (in Russian)

Kress, G. (2002). The multimodal landscape of communication. *MedienJournal, 4*, 4–19.

Litvinenko, A. O., Nikolaeva, Yu. V., & Kibrik, A. A. (2017). Annotirovanie russkikh manual'nykh zhestov: teoreticheskie i prakticheskie voprosy [Annotation of Russian manual gestures: Theoretical and practical issues]. In *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnoi konferentsii "Dialog 2017"* [Computational linguistics and intellectual technologies: Proceedings of the International conference "Dialogue 2017"] (pp. 255–268). Moscow: RGGU. (in Russian)

Loehr, D. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology, 3*(1), 71–89.

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

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.

Mondada, L. (2014). Bodies in action. *Language and Dialogue, 4*(3), 357–403.

Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics, 20*, 336–366.

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

Taylor, M. (1989). *The structure of multimodal dialogue*. Amsterdam: Elsevier.

Železný, M., Krňoul, Z., Císař, P., & Matoušek, J. (2006). Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis. *Signal Processing, 83*(12), 3657–3673.

Andrej A. Kibrik — director, the Institute of Linguistics, Russian Academy of Sciences; professor, Lomonosov Moscow State University, D.Sc.
Research area: cognitive linguistics, multimodality, discourse analysis, semantics, grammar, linguistic typology, areal linguistics, linguistic diversity, field linguistics.
E-mail: aakibrik@gmail.com

Olga V. Fedorova — professor, leading research fellow, the Institute of Linguistics, Russian Academy of Sciences, Lomonosov Moscow State University, D.Sc.
Research area: psycholinguistics, cognitive linguistics, multimodality, first language acquisition, discourse analysis, neurolinguistics.
E-mail: olga.fedorova@msu.ru

Эмпирическое исследование мультиканальной коммуникации: русские рассказы и разговоры о грушах

А.А. Кибрик[a,b], **О.В. Федорова**[b,a]

[a] *Институт языкознания РАН, 125009, Россия, Москва, Б. Кисловский пер., 1*
[b] *МГУ имени М.В. Ломоносова,119991, Россия, Москва, Ленинские горы, 1*

Резюме

Статья описывает язык в его наиболее естественной форме — в форме разговорного мультиканального дискурса. Он включает в себя вербальный компонент, просодию, движения взора, а также различные кинетические аспекты коммуникации — мимику, жесты головы, рук и туловища. Для изучения естественного многоканального дискурса как он есть мы создали ресурс «Русские рассказы и разговоры о грушах». Ресурс включает 40 записей, проведенных с 160 носителями русского языка в возрасте 18–36 лет, среди которых было 60 мужчин и 100 женщин; он состоит из 15 часов записи и около 170 000 слов. В статье описываются методология создания корпуса и возможности его использования. Во-первых, мы предлагаем обзор методологии мультимодальности и мультиканальных корпусов. Затем мы описываем характеристики нашего ресурса — методику сбора данных, используемое оборудование, типы аннотаций, а также некоторые пути (будущих) исследований, в том числе: просодия в качестве интерфейса между вокальным и жестовым каналами, особенности и степень координации между мануальными жестами и элементарными дискурсивными единицами, индивидуальное варьирование и «портретная» методология, порождение и понимание речи в естественной коммуникации, а также зрительное внимание в естественной коммуникации. Текущая версия корпуса доступна для научного сообщества на веб-сайте проекта multidiscourse.ru (на русском языке).

Ключевые слова: мультимодальность, мультиканальный дискурс, создание корпусов, просодия, жесты, движение взора, аннотация.

Кибрик Андрей Александрович – директор, Институт языкознания РАН; профессор, кафедра теоретической и прикладной лингвистики, филологический факультет, МГУ имени М.В. Ломоносова, доктор филологических наук.
Сфера научных интересов: когнитивная лингвистика, мультимодальность, анализ дискурса, семантика, грамматика, лингвистическая типология, ареальная лингвистика, языковое разнообразие, полевая лингвистика.
Контакты: aakibrik@gmail.com

Федорова Ольга Викторовна — профессор, кафедра теоретической и прикладной лингвистики, филологический факультет, МГУ имени М.В. Ломоносова; ведущий научный сотрудник, Институт языкознания РАН, доктор филологических наук.
Сфера научных интересов: психолингвистика, когнитивная лингвистика, мультимодальность, усвоение языка, анализ дискурса, нейролингвистика.
Контакты: olga.fedorova@msu.ru