A «Portrait» Approach to Multichannel Discourse

Andrej A. Kibrik, Olga V. Fedorova
Institute of Linguistics RAS & Lomonosov Moscow State University
B. Kislovskij per., 1 & Leninskie Gory, 1, Moscow, Russia
aakibrik@gmail.com, olga.fedorova@msu.ru

Abstract

This paper contributes to the research field of multichannel discourse analysis. Multimodal discourse analysis explores numerous channels involved in natural communication, such as verbal structure, prosody, gesticulation, facial expression, eye gaze, etc., and treats them as parts of an integral process. Among the key issues in multichannel studies is the question of the individual variation in multichannel behavior. We address this issue with the help of a multichannel resource "Russian Pear Chats and Stories" that is currently under construction (multidiscourse.ru). This corpus is based on a novel methodology of data collection and is produced with the help of state of the art technology including eyetracking. To address the issue of individual variation, we introduce the notion of a speaker's individual portrait. In particular, we consider the Prosodic Portrait, the Oculomotor Portrait, and the Gesticulation Portrait. The proposed methodology is crucially important for fine-grained annotation procedures as well as for accurate statistic analyses of multichannel data.

Keywords: multichannel discourse, prosody, gesticulation, eye gaze, a speaker's portrait

1. Introduction

People communicate with each other, using words, intonation, gestures, gaze, facial expression (Kress, 2002; Loehr, 2012; McNeill, 2005; Goldin-Meadow, 2014; Church et al. eds., 2017, inter alia), see Fig. 1.

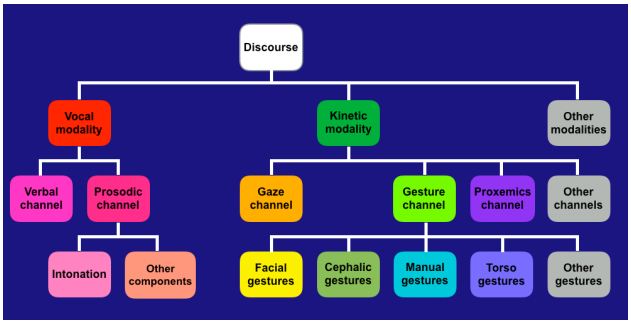


Figure 1: Spoken multichannel discourse.

All of these communication channels are employed simultaneously and in conjunction with each other. Therefore, everyday human communication is a multichannel process. We are immersed in this communication throughout our lives, but still it remains an underexplored phenomenon. There are at least two causes leading to this situation. First, the process of multichannel communication is ephemeral and passes by without leaving material traces. Second, the content of various communication channels traditionally belongs to the competence of different disciplines, weakly connected with each other. In particular, the verbal component is studied by linguists, while gestures and eye movements are explored, primarily, by psychologists. Our research group set as its goal a comprehensive study of multichannel discourse, based on an integrated theoretical and methodological approach.

A "multi-modal corpus" is defined as "an annotated collection of coordinated content on communication channels including speech, gaze, hand gesture and body language, and is generally based on recorded human behavior" (Foster, Oberlander, 2007: 307–308). As compared to monomodal corpora that already have a substantial history and tradition, multimodal corpora are still at their incipient stage. The most natural data have been assembled in the Fruit Carts Corpus that contains 240 videorecordings of 12 participants, each four to eight minutes long (Aist et al., 2012), the corpus D64, created for studies of everyday communication (Campbell, 2009), the InSight Interaction Corpus consisting of 15 recorded face-to-face conversations 20 min long each (Brône, Oben, 2015), as well as corpora created in the tradition of Conversation Analysis (Mondada, 2014; 2016, inter alia).

When exploring any linguistic phenomenon, one addresses two opposite (and complementary) issues: general trends, on the one hand, and individual variation, on the other. In the domain of multichannel discourse, exploration of either of the two issues requires a good quality, representative corpus. At this time we have created the resource "Russian Pear Chats and Stories" that includes a number of recorded sessions of natural communication between several participants, as well as vocal and kinetic annotations of these sessions. The sessions were recorded with the help of original technical solutions, including high quality audio and video recording, as well as eye tracking methods. The vocal (verbal and prosodic) annotation used in this project follows the principles previously developed for spoken Russian discourse, for more detail see Kibrik, Podlesskaja, 2009 and the website spokencorpora.ru. Within the framework of the present project, we have developed principles of kinetic and oculomotor annotation (see Section 2).

In this paper we address ***individual differences*** and propose a "portrait" approach to multichannel discourse. We demonstrate that individual portraits of interlocutors is a necessary prerequisite for fine-grained annotation and for in-depth analysis of multichannel discourse (see Sections 3–5).

2. Russian Pear Chats and Stories Corpus

2.1 Stimulus Material

We use the well known Pear Film (Chafe ed., 1980) as the stimulus material for collecting the data. The film contains no speech, and the events shown are relatively clear to any viewer. The film was constructed so that the shown scenes incline participants to describe landscape, explain cause-effect relations, account for the characters' thoughts and emotions, and resolve ambiguities.

2.2 Data Collection Setup

We have developed a new ***method*** of collecting data on the basis of the Pear Film. Each session, including the instructions we provided and participants' filling out the written consent, lasted for about one hour. Each session involved four participants with fixed roles, see Fig. 2.

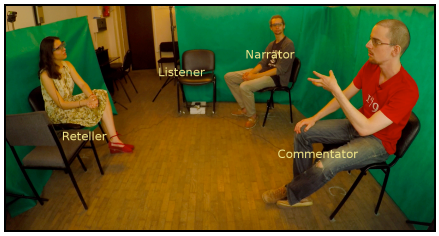


Figure 2: Data collection setup. Cover shot.

Three of them – the Narrator (N), the Commentator (C), and the Reteller (R) – took part in the main part of the recording, while the fourth participant, the Listener (L), joined them towards the end. The Narrator and the Commentator each watched the film on a personal computer and tried to memorize the plot and the details as precisely as possible. Then an interaction, also involving the Reteller, began. The Narrator told the Reteller about the plot of the film; this is a monologic stage – *first telling*. During an interactive stage (*conversation*), the Commentator supplied additional details and corrected the Narrator's story where necessary; the Reteller checked his/her understanding of the plot, asking questions to the Narrator and the Commentator. Then another monologic stage, *retelling*, followed, during which the Reteller was retelling the film to the Listener. Finally, the Listener wrote down the content of the film, as s/he had understood it from the Reteller's account.

2.3 Recording Devices

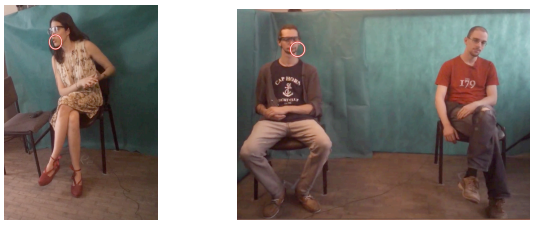
The participants' talk was recorded with the help of a six-channel recorder ZOOM H6 Handy Recorder (96 kHz / 24 bit). Three channels were used to record three speakers individually, with the lapel microphones SONY ECM-88B. Two more channels were used to produce a general recording of the whole speech signal, by using a stereo mic provided with the recorder.

Three industrial video cameras JAI GO-5000M-USB (100 frames per second and 1392x1000 pixels) recorded three participants, shooting individually from a frontal perspective (see Fig. 3). These cameras use the mjpeg format of recording, which is free of interframe compression; this is a crucial prerequisite for subsequent frame-by-frame annotation. In addition, the camera GoPro Hero 4 (50 frames per second and 2700x1500 pixels) was used to record the whole scene (see Fig. 2 above).

Narrator Commentator Reteller

Figure 3: Individual frontal shots.

In order to record eye gaze, two head-mounted eyetrackers were used (Tobii Glasses II, 50 Hz and 1920x1080 pixels). The Narrator and the Reteller were wearing eyetrackers, see Fig. 4. The eyetrackers provide two types of data: (i) videofiles produced by an inbuilt scene camera and (ii) data files representing eye movements. The screenshots in Fig. 4 result from an overlay of videofiles from the scene camera and the gaze coordinates from the data files; the circles are generated by the eyetrackers and indicate the targets of interlocutors' gaze.



From the N's tracker From the R's tracker

Figure 4: Video scene, as recorded by a camera built into the eyetrackers, with a superimposed marker of visual attention.

2.4 Participants and Corpus Size

The corpus includes 24 conversations among 96 Russian native speakers aged 18-36. It consists of 9 hours of recording (the average length of a recording was 24 min). We work with the data coming from 10 different sources: three video files from the individual industrial cameras, one video file from the cover shot camera, two files from the eyetrackers, and four audio files.

2.5 Annotations

2.5.1 Vocal Annotation

Verbal-prosodic (vocal) annotation consists of spoken discourse divided into relevant elements (elementary discourse units – EDUs, words, absolute and filled pauses, nonspeech sounds), as well as of attributes assigned to EDUs and their parts (pitch accents, accelerated tempo, reduced pronunciation, lowered tonal register, etc.); see Kibrik 2011 for details. As a result of vocal annotation (performed by Nikolay Korotaev and Vera Podlesskaya) transcripts were obtained, in the form of text documents, as well as textgrid files with multi-layer annotation prepared with the help of speech analyzing program Praat (fon.hum.uva.nl/praat) and reflecting temporal anchoring of all annotated phenomena.

2.5.2 Annotation of Manual Gestures

For the transcription of the videodata we used the annotation software ELAN (lat-mpi.eu/tools/elan/) and followed the annotation system developed in (Litvinenko et al., 2017); annotation was done by Alla Litvinenko and Julia Nikolaeva. The proposed annotation scheme proceeds from the basic level of the distinction between motion and stillness to more complex structures at the next level. At the first stage we annotate simplest motion units, or movements, for each hand separately. Each of these movements functions as a gesture phase (Kendon, 1980; 2004; Ladewig, Bressem, 2013), a self-adaptor phase, or a position change movement. All the stillness intervals are also annotated and classified. At the next stage, we annotate hand postures. Gesture and movement chains are the final component of the scheme. A gesture chain is an uninterrupted series of gestures; a movement chain is an uninterrupted series of movements.

The proposed annotation scheme was originally created for manual movements. At the same time, the scheme is broadly applicable to other gestural components, in particular head movements.

2.5.3 Annotation of Gaze

In the course of annotation of the eye gaze component we conducted the export of the eyetracking data onto the video scene, and used the Tobii Pro Glasses Analyzer program to extract the information on the temporal structure of all fixations longer than 100 ms (that is, a participant's fixation on a target must last for at least 100 ms to be recognized as a gaze event). The possible targets include "Interlocutor" ("Narrator / Reteller", "Commentator" or "Listener"), further subdivided into "face", "hands", "torso", and "other", and "Surroundings", see Fedorova 2017 for more detail.

2.5.4 Multilayer Annotation

Fig. 5 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; for more details see our website multidiscourse.ru/annotation.

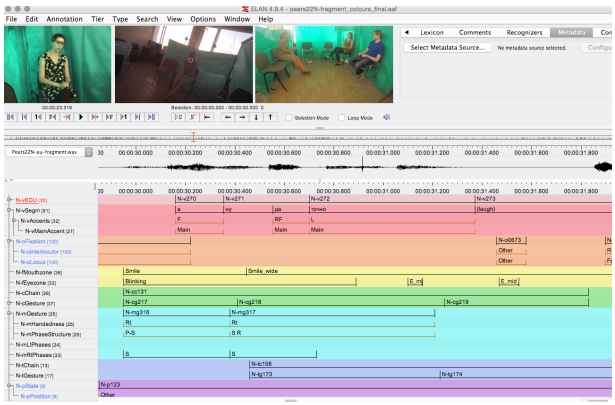


Figure 5: Multilayer annotation.

3. Prosodic Portrait

Many prosodic phenomena are of a relative, rather than absolute, nature: their specific realizations can be assessed and identified only with respect to neutral characteristics of a given speaker's voice. In particular, two kinds of the falling intonation must be distinguished: a final falling (so-called period intonation) and a non-final falling (a falling comma intonation). In order to posit punctuation marks in a transcript belonging to a certain speaker, one needs to explore and describe the speaker's prosodic system. That is, a complete account of a spoken discourse presupposes a speaker's Prosodic Portrait, i.e. a range of his or her prosodic characteristics. Compiling such a portrait is stage zero in the work on a transcript, preceding vocal annotation as such; see Kibrik, 2009, 2011 for further details.

Possible components of a Prosodic Portrait are shown in Table 1; for the annotation in Word and Praat see multidiscourse.ru/corpus. Minimal and maximal F0 values are point values, as they indicate the limits of a range in which F0 of a speaker's voice fluctuates. Other measurements are interval values (rare and extreme outliers are ignored). Where possible, e.g. in the case of a target level of falling, such interval values are centered around a median value (which also is a modal value). In other instances specific values are so widely scattered that no median or modal value is indicated; for example, in this particular speaker a wide interval is found in his target levels of rises in the canonical comma intonation.

Minimal F0 value, Hz	91
Maximal F0 vaue, Hz	214
Standard level of EDU onsets, Hz	140±5
Target level of final fallings, Hz	105±3
Target level of non-final fallings, Hz	115±3
Target level of rises in the canonical comma intonation, Hz	140~185
Standard level of falling on post-accent syllables, typical of the canonical comma intonation, Hz	120±5
Target level of rises in the "three dots" intonation, Hz	118~132

Table 1: Prosodic Portrait for participant 04C.

4. Oculomotor Portrait

Whereas a Prosodic Portrait is an important element of the transcription process, an oculomotor portrait fulfils a different function: it is created after the procedure of semi-automatic annotation and serves at the stage of data analysis. Individual differences between speakers are so substantial that one cannot properly compare the data belonging to different participants without averaging them, or, more precisely, without normalizing them and reducing them to quantiles for the purposes of the subsequent analysis of variance. This is what is done with the help of oculomotor portraits. Table 2 illustrates a so-called Standard Oculomotor Portrait of Narrator 04, involving the data of the summary quantity of fixations throughout the duration of the session; the summary duration of the fixations; mean, minimal, and maximal durations; as well as 25%, 50%, and 75% quantiles. In

addition, the portrait also includes analogous distributions of fixations on R, C, L, and the Surroundings. An Extended Oculomotor Portrait includes analogous data on the participants' fixations distributions during the stages of *first telling*, *conversation* and *retelling*, as well as a more complex distribution table of the fixations on the participant's body parts: face, hands, torso, and other. (Application developer is Ivan Zherdev, see github.com/ivan866/readTobiiGlasses; for the annotation in Excel and ELAN see multidiscourse.ru/corpus).

Total duration	1170.667
Total count	2249
Mean duration & std	0.52 & 0.66
Min, Quantile 25, 50, 75, Max	0.06, 0.16, 0.26, 0.58, 10.5
R: count & ratio	1033 & 0.46
R: duration & ratio	848.55 & 0.72
R: mean & std	0.82 & 0.85
R: Min, Quantile 25, 50, 75, Max	0.06, 0.24, 0.5, 1.16, 10.5
C: count & ratio	129 & 0.057
C: duration & ratio	43.8 & 0.036
C: mean & std	0.33 & 0.28
C: Min, Quantile 25, 50, 75, Max	0.06, 0.16, 0.24, 0.4, 1.32
L: count & ratio	16 & 0.007
L: duration & ratio	3.28 & 0.003
L: mean duration & std	0.2 & 0.12
L: Min, Quantile 25, 50, 75, Max	0.06, 0.12, 0.19, 0.26, 0.44
Surroundings: count & ratio	1071 & 0.48
Surroundings: duration & ratio	275.02 & 0.23
Surroundings: mean & std	0.26 & 0.21
Surroundings: Min, Quantile 25, 50, 75, Max	0.06, 0.13, 0.2, 0.3, 2.2

Table 2: Standard Oculomotor Portrait for participant 04N (durations shown in seconds).

5. Gesticulation Portrait

Finally, a Gesticulation Portrait fulfils a double function. It is necessary both at the stage of manual gesture annotation and at the stage of gesture analysis. Table 3 contains a Standard Gesticulation Portrait for Reteller 04; for ELAN annotation see the website multidiscourse.ru/corpus. The data necessary at the stage of annotation include: (dis)inclination to stillness; (dis)inclination to self-adaptors; typical amplitude; typical velocity; and preferences in gesture handedness: predominance of two-handed or one-handed gestures, as well as a distribution of one-handed gestures in accordance with handedness. Data necessary for a subsequent comparison include: 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. In addtion, a portrait includes an analogous distribution of manual gestures in accordance with the stages of *first telling*, *conversation*, and *retelling*. An Extended Gesticulation Portrait contains the same data, listed separately for the right and left hands.

Inclination to stillness	low
Inclination to adaptors	high
Amplitude of manual gesture	low
Velocity of manual gesture	high
Two-handed vs. one-handed	0.32 vs. 0.68
One-handed gestures in accordance with handedness: right vs. left	0.68 vs. 0.32
Total count	481
Total gesticulation duration	414.53
Mean duration & std	0.86
Min, Quantile 25, 50, 75, Max	0.09, 0.42, 0.64, 1, 8.7
Conversation: count & ratio	210 & 0.44
Conversation: duration & ratio	198.72 & 0.48
Conversation: Min, Quantile 25, 50, 75, Max	0.16, 0.45, 0.68, 1.15, 8.73
Retelling: count & ratio	271 & 0.56
Retelling: duration & ratio	215.81 & 0.52
Retelling: Min, Quantile 25, 50, 75, Max	0.09, 0.38, 0.62, 0.96, 3.73

Table 3: Standard Gesticulation Portrait for participant 04R (durations shown in seconds).

6. Conclusion

In modern linguistics, as well as in other domains of cognitive science, there is a growing understanding that human communication is inherently multimodal. A research program of multimodal linguistics is gradually evolving (Kibrik, 2010; Knight, 2011; Adolphs, Carter, 2013; Kibrik, Molchanova, 2013; Müller et al. eds., 2014) that treats verbal structure on a par with non-verbal devices. Among non-verbal devices, sometimes only kinetic-visual behaviors are considered. But we find it very important to include prosody (see e.g. Kodzasov, 2009), that is non-segmental aspects of the vocal signal, as a distinct communication channel.

In the course of the project we create a multimodal resource of natural Russian discourse that does not have direct analogs among the contemporary resources. It is created for a wide range of research goals (coordination between units belonging to different channels: clause, EDU, gesture; visual attention and units of communicative behavior; multimodal turn-taking; reinterpretation of "pause" in the multimodal perspective, inter alia). Our resource is based on a novel methodology of data collection and is produced with the help of state of the art technology. It is annotated on the basis of a multilayer discourse transcription system and is freely available online to anyone interested (website multidiscourse.ru). The corpus will serve as a data source for multichannel communication research by linguists, as well as by specialists in other domains of cognitive science.

In this paper we have considered prosodic, gesticulation, and oculomotor portraits of the participants and have demonstrated that such portraits constitute an essential part of both annotation process and data analysis.

7. Acknowledgements

This study is supported by Russian Science Foundation (grant #14-18-03819).

8. Bibliographical References

Adolphs, S., Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. N.-Y.: 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.

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*. Springer: Heidelberg.

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

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

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*, 369.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M.R. Key (Ed.) *The relationship of verbal and nonverbal communication* (pp. 207–227).

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. New York: Cambridge University Press.

Kibrik, A.A. (2009). A speaker's prosodic portrait. In A.A. Kibrik, V.I Podlesskaja (Eds.) *Night Dream Stories: A corpus study of spoken Russian discourse*. Moscow: Jazyki slavjanskix kul'tur.

Kibrik, A.A. (2010). Multimodal linguistics. In Yu.I. Aleksandrov, V.D. Solov'jev (Eds.) *Cognitive studies*, IV. Moscow: Institute of psychology.

Kibrik, A.A. (2011). Cognitive discourse analysis: local discourse structure. In: M. Grygiel and L.A. Janda (Eds.) *Slavic Linguistics in a Cognitive Framework*. (pp. 273-304). Frankfurt/New York: Peter Lang Publishing Company,

Kibrik, A.A., Podlesskaja, V.I. (Eds.) (2009). *Night Dream Stories: A corpus study of spoken Russian discourse*. Moscow: Jazyki slavjanskix kul'tur.

Kibrik, A.A., Molchanova, N.B. (2013). Channels of multimodal communication: Relative contributions to discourse understanding. In M. Knauff, M. Pauen, N. Sebanz, I. Wachsmuth (Eds.) *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2704–2709). Austin, TX: Cognitive Science Society.

Kodzasov, S.V. (2009). *Studies in the field of Russian prosody*. Moscow: Jazyki slavjanskix kul'tur.

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

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

Ladewig, S.H., Bressem, J. (2013). A linguistic perspective on the notation of gesture phases. In *C. Müller, A. Cienki, E. Fricke, S.H. Ladewig, D. McNeill, S Teßendorf (Eds) Body — Language — Communication: An International Handbook on Multimodality in Human Interaction* (Vol. 2, pp. 1060–1078). Berlin: Mouton.

Litvinenko, A.O., Nikolaeva Ju.V., and Kibrik, A.A. (2017). Annotation of Russian manual gestures: Theoretical and practical issues. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017"* (pp. 255–268). Moscow: RGGU.

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

McNeill, D. (2005). *Gesture and thought*. Chicago.

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: Mouton de Gruyter.