Channels of multimodal communication: Relative contributions to discourse understanding

Andrej A. Kibrik (aakibrik@gmail.com)
Institute of Linguistics of the Russian Academy of Sciences and Lomonosov Moscow State University
B. Kislovskij per. 1, Institute of Linguistics RAN, Moscow 125009 Russia
Natalia B. Molchanova (natascha.molchanova@gmail.com)
BearingPoint
B. Ordynka 40-4, Moscow 119017 Russia

Keywords: discourse; communication channels; understanding; multimodality

1. Introduction: Communication channels

The mainstream view of linguistic form, characteristic of modern linguistics, can be formulated as follows: language consists of hierarchically organized segmental units, such as phonemes, morphemes, words, phrases, and sentences. Mainstream linguistics thus equates linguistic form with *verbal* form, that is, the segmental vocal material. However, as we all know, apart from sound, there are other channels (or components) of communication, in the first place through vision. The visual channel is what is sometimes named with the cover term *body language*, including gesture, mimic, gaze, posture, etc. (see e.g. McNeill, 1992; Kendon, 1994; Goldin-Meadow, 1999; Krejdlin, 2002; Butovskaja, 2004; Andersen, 2007; Burgoon et al., 2011).

Furthermore, the vocal material is not exhausted by verbal elements. There is also *prosody*, that is, non-verbal (= non-segmental) aspects to sound, including intonation, tempo, pausing, loudness, discourse accents, tonal registers, etc. (see e.g. Cruttenden, 1986; Kodzasov, 2009).

An unbiased view should probably be the following: all of these components must be taken into account in a realistic model of communication. For example, imagine that you are staying in a hotel room with thin walls and can hear people next door talking. You cannot hear words (the verbal component) but you can hear prosody, and you get something about the conversation, for example you may know that the people are quarreling. On the other hand, prosody-free talk, as sometimes heard from TV simultaneous interpreters on the Euronews channel, is unnatural and hinders comprehension. In this study we address the question of the *relative contribution* of the various communication channels or components to the overall comprehension of spoken discourse.

2. Views on the importance of various communication channels

The traditional approach of mainstream linguistics has been to consider the verbal channel so central that prosody and the visual channel have often been downgraded as "paralinguistics". Many contemporary textbooks in linguistics barely mention prosody and do not mention gesture and body language at all (see e.g. Hall, 2005).

The other extreme is represented by the view common in applied psychology that words matter less than prosody and especially than body language. It is very often that the following figures are quoted, going back to Mehrabian (1971): body language conveys 55% of information, prosody conveys 38% of information, and the verbal component only 7% of information[1], see e.g. http://jobsearch.about.com/od/interviewsnetworking/a/nonverbalcomm.htm. According to this view, "words may be what men use when all else fails" (Krejdlin, 2002: 6).

Most likely, the truth lies between these two extremes. All of the communication channels must be valuable and none can be negligible. This kind of balanced approach is characteristic of the modern multimodal paradigm (see e.g. Granström et al. eds., 2002; Norris, 2004; Ventola et al. eds., 2004; Bengio & Bourlard eds., 2005; Royce & Bowcher, 2007; Jewitt ed., 2011). According to Kress (2002), "A multimodal approach assumes that the message is 'spread across' all the modes of communication. If this is so, then each mode is a partial bearer of the overall meaning of the message." To use a quotation from the computational domain, "within biology, experimental psychology, and cognitive neuroscience, a separate rapidly growing literature has clarified that multisensory perception and integration cannot be predicted by studying the senses in isolation" (Cohen & Oviatt, 2006). Kibrik (2010) described the research program of multimodal linguistics, taking into account all of the communication channels in an integrated approach.

Taking up the challenge of Mehrabian (1971), in this study we try to numerically estimate the contribution of each communication channel into the overall process of message understanding. (Cf. two early psychological studies Walker, 1977 and Hollandsworth et al., 1979, arriving at rather opposite conclusions, and also Cutica & Bucciarelli, 2006.)

[1] In fact, Mehrabian originally investigated just the contributions of the channels to a listener's attitude towards a message in emotional settings, but his figures have often been misinterpreted as accounting for any kind of communication, see e.g. http://www.speakingaboutpresenting.com/presentation-myths/mehrabian-nonverbal-communication-research/.

3. Experimental design

The experimental design, first developed by Andrej Kibrik in 2006 at the Deparment of Theoretical and Applied Linguistics, Philological Faculty, Moscow State University (see Kibrik & Èl'bert, 2008, Kibrik, 2010), consists of several elements. For the purposes of this study, we differentiate between three communication channels, or components, including two vocal channels, the verbal and the prosodic ones, and the visual channel comprising all elements of body language; see Figure 1.

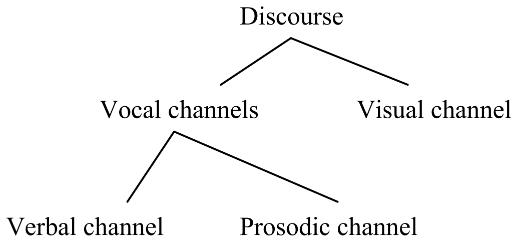

Figure 1: Three communication channels.

If we take a sample of natural discourse, we can isolate three communication channels. For example, if we have a recording of communication, a video without sound is equivalent to the visual channel alone. We also need to isolate the verbal channel and the prosodic channels; specific technical ways of how that can be made possible are explained in sections 4 and 5 below. Assuming that the three channels have been isolated, we can produce eight (2^3=8) variants of the sample discourse and present them to separate groups of experimental participants. These eight variants include three in which only one channel is represented, three in which two channels are represented, one with the three channels (the original material), and the null variant in which nothing has been shown to participants. We will thus need eight groups of participants, each presented with one of the eight kinds of experimental material.

The null variant of the experimental discourse and the corresponding group is necessary in order to evaluate which part of the overall content can be inferred on the basis of background knowledge and common sense.

At the next stage the degree of the participants' understanding of the discourse can be assessed with the help of a questionnaire, and such assessment may be used as an estimate of a communication channel's contribution to the overall discourse understanding.

4. Experiment A: movie-based material

The first line of studies in this paradigm was implemented in a series of experiments by Ekaterina Èl'bert, particularly in her diploma thesis (2007), and further reinterpreted and refined in Kibrik and Èl'bert (2008). In this line of studies the decision was made to use an excerpt of a movie as experimental discourse. Specifically, the Russian TV serial "Tajny sledstvija" ("Mysteries of the investigation") was used. The experimental excerpt ran for 3 minutes and 20 seconds, and it was preceded by a 8 minutes context excerpt, starting from the beginning of a series. The experimental excerpt fully consisted of a conversation, to ensure that we are testing the understanding of discourse rather than of the film in general.

The two vocal channels were separated from each other through the following procedures. The verbal channel was presented in the written mode, by means of temporally aligned running subtitles. The prosodic channel was obtained from the original sound by superimposing a filter creating the "behind a wall" effect. Figure 2 illustrates a snapshot from the experimental type "visual plus verbal", in other words, video plus running subtitles.

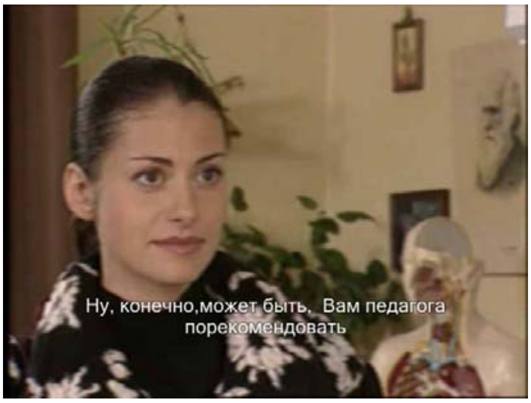

Figure 2: Frame from the experimental material "visual plus verbal".

99 participants took part in the study, divided into eight groups, each group comprising 10 to 17 persons. All eight groups watched the identical context excerpt. As for the experimental excerpt, each of the eight groups had access to different material. The null group did not see anything apart from the context excerpt, three groups only had access to one communication channel of the experimental excerpt (either verbal or prosodic or visual), the other three groups to two communication channels (verbal+prosodic = original sound; verbal+visual = video and subtitles, see Figure 2; prosodic+visual = video and filtered sound), and the eighth group watched the original version of the experimental excerpt.

The context and the experimental excerpts were shown to the whole group of participants on a large screen. Each participant was instructed to attend the context and the experimental excerpt and then answer a set of questions concerned with the experimental excerpt alone. The questionnaire was constructed in accordance with the received principles of test tasks (Panchenko, 2000). There were 23 multiple-choice questions in the questionnaire; a participant was supposed to choose only one answer out of four listed variants. Here is an example of a question, along with the offered answers (translated from the Russian original):

What Tamara Stepanovna offers Masha before the beginning of the conversation:
a. to take off her coat
b. to have a cup of tea
c. to have a seat
d. to have a drink

One of the available answers (in this particular case, c) was correct, two were plausible but wrong (a, b) and one implausible (d); the latter was aimed at filtering out incompetent participants.

Results of Experiment A

Percentage of correct answers was used as a way to assess a participant's degree of discourse understanding. The summarized results are shown in a diagram in Figure 3.

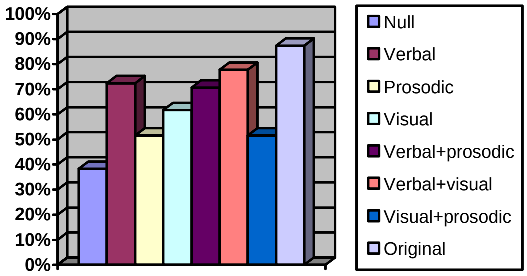

Figure 3: Degrees of discourse understanding in Experiment A.

We see from the second, third, and fourth columns in Figure 3 that each individual communication channel is substantially informative: The verbal channel is leading in this respect (72.4% correct answers), but the two other (prosodic: 51.5%, visual: 61.7%) also significantly (Mann-Whitney test, $p<0.05$) prevail over the null condition (leftmost column, 38.3%). The hierarchy of the individual channels turns out verbal>visual>prosodic (significant according to Kruskal–Wallis test (H (2, 69) = 24.2, $p<0.01$)). In spite of the prevalence of the verbal channel, the difference in the contributions of individual channels is not dramatic, and second, the degree of understanding in the "verbal alone" condition is significantly (Mann-Whitney test, $p<0.05$) lower than in the original material condition (three channels in conjunction, the rightmost column in Figure 3, 87.4%).

Another conclusion from the results of Experiment A concerns the comparison of the three groups that had access to two communication channels; see columns fifth to seventh from the left. There is a very noticeable (but not reaching the level of statistical significance) dip in the condition "visual+prosodic" (51.6%), compared to two other pairwise combinations (verbal+prosodic: 70.7%; verbal+visual: 77.8%). Apparently, that dip means that language users have difficulties integrating information from the visual and prosodic channels, in the absence of verbal material. In a natural setting, this condition can be compared to observing communication via a glass that is penetrable for prosody but blocks the verbal material. Most likely, the dip in the "visual+prosodic" condition is due to the unusual character of such situations in real life, as well as to the participants' inability to integrate information from the visual and prosodic channels in the absence of verbal material.

5. Experiment B: conversation-based material

At the following stage of the project, we modified and/or improved a number of the methodological decisions made in Experiment A, including the kind of stimulus material, the technical methods of isolating the prosodic channel and the verbal channel, the questionnaire, and the interviewing procedure. The below description of Experiment B is organized as follows. Each of the mentioned methodological decisions made in Experiment A is assessed, and a modification/improvement realized in Experiment B is presented.

Several problems of the movie-based stimulus material, used in Experiment A, were detected, including the following. First, the plot of the movie in certain instances facilitated guessing by the experiment participants. Second, it was not possible to exclude the familiarity of the movie to some of the participants. Third, the quasi-natural behavior of the actors could affect the results. Fourth, all speakers were of the same gender (women) which made it difficult for the participants to distinguish between voices, especially in the "prosodic alone" condition.

The solution realized in Experiment B was to employ a recording of natural dialogue between two speakers. In order to make the dialogue structured and predictable, a guessing game "Little garages" was recorded. One of the speakers, a woman, was laying a number of toothpicks on the table and was asking the guesser, a man: "How many little garages?" The guesser was trying to figure out how to provide a correct answer, which was difficult (because the intended amount of little garages was in fact the number of the the first speaker's fingers kept on the table at the moment). The guessing process lasted for 19 minutes, out of which the stimulus material of 5 minutes and 55 seconds was produced. The stimulus material consisted of a dialogue between the two speakers, culminating in the guesser's ultimate success. A frame from the guessing game recording appears in Figure 4.

The acoustic filter used in Experiment A produced the material in the "prosodic alone" condition that was excessively noisy. The solution used in Experiment B was to radically decrease the signal at all frequencies except for the speaker's average F0 frequency. This led to a more satisfactory "behind the wall" effect.

Figure 4: Frame from the recording of the guessing game.

The main problem associated with the "verbal alone" condition in Experiment A was that the subtitles operated in the visual, rather than the vocal, mode. This had created a substantial deviation from the situation of spoken discourse, also leading to the undesired interaction and/or competition between the written verbal material and the visually perceived video material. In addition, some participants experienced difficulties in following the subtitles appearing and disappearing at the same pace as spoken words in the original material. The solution introduced in Experiment B was to produce an artificial spoken prosody-free signal. Both speakers participating in the recording were requested to individually pronounce each word that occurred in their conversation. All thus elicited words were then glued together in the right order, thus providing prosody-free discourse, devoid of intonation, reduction, differences in tempo, etc.

As far as the questionnaire is concerned, the imperfection of Experiment A is seen through the insufficient gap between the results of the null group and the original material group: 38.3% vs. 87.4%. These numbers indicate that the participants were able to reconstruct the correct answer quite often and, on the other hand, even the full original material did not provide reliable access to a correct answer. In order to improve the questionnaire, a testing stage was introduced in Experiment B, in which trivial questions were identified (high null group results), as well as unfortunate questions (low original material group results). Trivial and unfortunate questions were filtered out, and the number of questions was reduced from 30 to 17. The improved results in the two contrastive groups turned out 24.7% and 91.2% of correct answers, see below.

The interviewing procedure was improved in Experiment B. In Experiment A the participants were of various and uncontrolled age and life experience. The presence of multiple participants in the room could have led to undesirable and uncontrolled interference. Finally, the need for a large room, loud speakers, and a big screen is an unnecessary technical complication to the procedure. In Experiment B the participants were controlled for age, geographical origin, and social status: only students of Moscow origin were recruited, which provided a homogeneous sample. They were also balanced in terms of gender. The experiment was implemented in a remote fashion: the stimulus material was posted on youtube.com, and the questionnaire at Googledocs. The guidelines closely directed the participants' sequence of actions, from one experimental part to another and from one group of questions to another, so there are reasons to believe that the procedures were very similar in all participants. All participants worked in comparable, independent, and comfortable conditions, and there was no need for technical excessiveness such as a big screen and loud speakers. 92 participants altogether took part in the experiment, out of which 20 were employed at the testing stage and 72 at the main stage (from 10 to 15 in each experimental group).

Results of Experiment B

The quantitative results of Experiment B are shown in Figure 5.

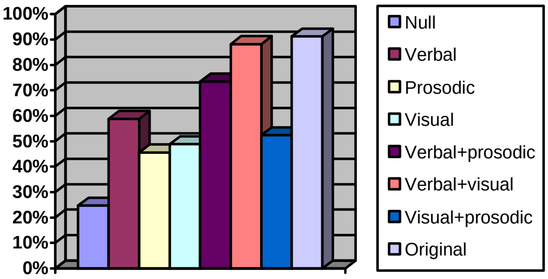

Figure 5: Degrees of discourse understanding in Experiment B.

The main findings of Experiment B are similar to those obtained in Experiment A. All three communication channels, taken in isolation (columns two to four from the left) are substantially and comparably informative: they lead to 58.8%, 45.6%, and 48.8% of correct answers, compare that to the 24.7% in the null group. The hierarchy of informativeness is again verbal>visual>prosodic. The conditions with two channels available (columns five to seven from the left) demonstrate the following results: 73.5%, 88.2%, and 52.4%. Compared to Experiment A, we here get a much cleaner picture as concerns the better participants' performance in the two channels conditions as contrasted with the one channel conditions. Finally, we see again a dramatic dip in the "visual+prosodic" condition: the second last column counting from the left.

6. Discussion

The main conclusion of Experiment B is the following: in spite of the substantial differences in the methodology from Experiment A, the results are remarkably similar. With minor differences the overall picture in Figures 3 and 5 is very similar. This makes us believe that our conclusions about the relative contributions of various communication channels to the overall discourse understanding are fairly robust.

Now, the picture in Figure 5 is cleaner and crisper in two respects: the more obvious advantage of the two channel conditions over the one channel conditions and the better contrast between the null group and the original material group.

In order to provide a response to Mehrabian's (1971) famous (or infamous) numbers, the following method can be applied. Suppose the three communication channels are independent (this is a strong assumption, but it is necessary for calculating the relative contributions of the channels). We can sum up all percentages in the one-channel conditions and then normalize them to 100%. Let us perform this operation on the results of both experiments, looking at the numbers in columns two to four from the left in Figures 3 and 5 (percentages are rounded to 1 per cent). The outcome of this procedure is shown in Table 1.

Table 1: Normalized contributions of the three communication channels.

		Experiment A	Experiment B
Summed percentages		72+52+62=186	59+46+49=154
Normalized contributions	Verbal	72%:1.86≈39%	59%:1.54≈38%
	Prosodic	52%:1.86≈28%	46%:1.54≈30%
	Visual	62%:1.86≈33%	49%:1.54≈32%

Once again, we see the striking similarity in the results of the two experiments: the numerically evaluated contributions of the three channels never differ from each other by more than 2%. So the contributions of the channels are stable irrespective of the specifics of methodology.

Also, the gender differences between the participants were explored in Experiment B. Two particularly interesting results were obtained for the conditions "verbal alone" and "visual+prosodic"; they are shown in Table 2.

Table 2: Performance of men and women in two conditions in Experiment B (percentages of correct answers indicated)

Condition	Men	Women	Advantage
Verbal alone	59.1	69.9	Women: +10.7
Visual+prosodic	66.1	51.6	Men: +14.5

As is clear from Table 2, in the condition "verbal alone" the women have demonstrated a striking advantage, providing correct answers much more frequently than the men. In contrast, the men demonstrated a strong advantage in the condition "visual+prosodic" that, as was discussed above, corresponds to an unusual situation and generally creates a difficulty in comparison with other two-channel conditions. These results conform to certain generalizations about gender intelligence, such as the women's better performance in verbal tasks and men's better performance in novel situations (see e.g. Bendas, 2006).

7. Conclusions

This study is the first linguistically-informed demonstration of the importance of several communication channels for understanding natural discourse. The following conclusions can be drawn from the reported study.

First, all communication channels are highly significant in encoding content and understanding of discourse. Therefore, the attitude common in mainstream linguistics, according to which linguistic communication is performed mostly by the verbal component, whereas other channels are negligible, is incorrect.

Second, among the communication channels the verbal channel is the leading one. Therefore, the viewpoint popular in applied psychology, according to which the contribution of the verbal component is negligible, is erroneous as well.

Third, the specific normalized contributions of the verbal, prosodic, and visual channels are in the vicinity of 38%, 30%, and 32%, respectively.

Fourth, participants have difficulties integrating the information from the visual and prosodic channels, in the absence of the verbal channel. This suggests that in normal communication the verbal channel plays the role of an anchor to which the information from other channels is attached.

Fifth, men and women perform differently in the conditions of isolated communication channels, women having advantage in the "verbal alone" condition and men having advantage in the novel and unusual "visual+prosodic" condition.

As was pointed out in section 5, many questions from the original questionnaire were filtered out for certain substantial reasons, which has reduced the number of questions from 30 to 17. In combination with the large number of conditions (eight), this has led to the fact that the quantitative tendencies observed in Experiment B do not quite reach the level of statistical significance. In April 2013 we collected additional data, bringing the number of subjects in each group to at least fifteen (total=132). We expect that, when the statistical analysis is completed, full significance of the results will be attained, as well as a formal comparison of the results of the two experiments.

A number of methodological issues remain for further research. In particular, we would like to pinpoint two of those. First, we are planning to experiment with monologic discourse addressed to public audience, such as presentations of travel agents in front of a group of people. This would complement the already attained results from our studies of dialogic communication. Second, we will keep working on refining the methods allowing to isolate the verbal channel. Both of the so far employed methods have their shortcomings, the subtitles switching from the auditory to the visual modality and the prosody-free talk being the unnatural kind of input. We will keep searching for additional methods helping to present the "verbal alone" condition in a more ecologically valid way.

A major problem in the studies of human communication and discourse is associated with the fact that different disciplinary traditions and paradigms address communication from different angles, not consulting each other's results. Linguists usually only pay attention to the verbal component, while non-verbal communication is mostly explored by social psychologists. In this study we propose an approach that is hopefully relevant for each of the fields studying human communication and bridging the gap between them.

We would like to conclude with a quotation from Ron Scollon (2006): "Any use of language is inescapably multimodal".

Acknowledgements

We are grateful to Olga Fedorova, Anna Laurinavičiute, and Andriy Myachykov who provided useful and important input during the preparation of this paper. The study was conducted with financial support from grant #11-04-00153 of the Russian Foundation for the Humanities.

References

Andersen, P. (2007). *Nonverbal Communication: Forms and Functions*. Long Grove: Waveland Press.

Bendas, T. V. (2006). *Gendernaja psixologija [Gender psychology]*. Saint-Petersburg: Piter.

Bengio, S., & Bourlard, H. (Eds.) (2005). *Machine learning for multimodal interaction*. Berlin: Springer.

Burgoon, J. K., Guerrero, L. K., & Floyd, K. (2011). *Nonverbal communication*. Boston: Allyn & Bacon.

Butovskaja, M. L. (2004). *Jazyk tela: priroda i kul'tura [Body language: nature and culture]*. Moscow: Nauchnyj mir.

Cohen, P. R., & Oviatt, S. L. (2006). Multimodal interaction with computers. In K. Brown (Ed.), *Encyclopedia of language and linguistics*, 2nd ed. Oxford: Elsevier.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Cutica, I., & Bucciarelli, M. (2006). Why gestures matter in learning from a discourse. In B. M. Velichkovsky, T. V. Chernigovskaya, Yu. I. Alexandrov, & D. N. Akhapkin (Eds.), *The Second Biennial Conference on Cognitive Science. Vol.* 1. Saint-Petersburg: Saint-Petersburg State University, Philological Faculty, 40-41.

Èl'bert, E. M. (2007). *Vklad verbal'nogo, prosodicheskogo i vizual'nogo kanalov v ponimanie ustnogo diskursa [Contribution of the verbal, prosodic, and visual channels in the understanding of spoken discourse]*. Diploma thesis. Moscow State University.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Science, 3*, 419-429.

Granström, B., House, D., & Karlsson, I. (Eds.) (2002). *Multimodality in Language and Speech Systems*. Norwell: Kluwer.

Hall, C. J. (2005). *An Introduction to Language and Linguistics: Breaking the Language Spell*. London: Continuum.

Hollandsworth, J. G. Jr., Kazelskis, R., Stevens, J. & Dressel, M. E. (1979). Relative contributions of verbal, articulative, and nonverbal communication to employment decisions in the job interview setting. *Personnel Psychology, 32*, Issue 2, 359-367.

Jewitt, C. (Ed.) (2011). *The Routledge Handbook of Multimodal Analysis*. London: Routledge.

Kendon, A. (1994). Do gestures communicate? A review. *Research on Language and Social Interaction, 27*, 175-200.

Kibrik, A. A., & Èl'bert, E. M. (2008). Understanding spoken discourse: the contribution of three information channels. In Yu.I.Alexandrov et al. (Eds.), *Materials of the 3rd International Conference on Cognitive Science,* 82-84. Moscow: IP RAN.

Kibrik, A. A. (2010). Multimodal'naja lingvistika [Multimodal linguistics]. In Yu. I. Alexandrov & V. D. Solovyev (Eds.), *Kognitivnye issledovanija*, issue 4. Moscow: IP RAN.

Kodzasov, S. V. (2009). *Issledovanija v oblasti russkoj prosodii [Studies in the field of Russian prosody]*. Moscow: Jazyki slavjanskix kul'tur.

Krejdlin, G. E. (2002). *Neverbal'naja semiotika [Non-verbal semiotics]*. Moscow: NLO.

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

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

Mehrabian, A. (1971). *Silent messages*. Belmont, CA: Wadsworth Publishing Company.

Norris, S. (2004). *Analyzing multimodal interaction: A methodological framework*. London: Routledge.

Panchenko, A. A. (2000). *Razrabotka testov [Developing tests]*. Khabarovsk.

Royce, T. D., & Bowcher, W. L. (2007). *New directions in the analysis of multimodal discourse.* Mahwah: Lawrence Erlbaum.

Scollon, R. (2006). Multimodality and the language of politics. In K. Brown (Ed.), *Encyclopedia of language and linguistics*, 2nd ed. Oxford: Elsevier.

Ventola E., Charles, C., & Kaltenbacher, M. (Eds.) (2004). *Perspectives on multimodality*. Amsterdam: Benjamins.

Walker, M. B. (1977). The relative importance of verbal and nonverbal cues in the expression of confidence. *Australian Journal of Psychology, 29*, Issue 1, 45-57.