Russian CliPS:
a Corpus of Narratives by Brain-Damaged Individuals

Mariya Khudyakova, Mira Bergelson, Yulia Akinina,

Ekaterina Iskra, Svetlana Toldova, Olga Dragoy

National Research University Higher School of Economics
20 Myasnitskaya Ulitsa, Moscow 101000, Russia
E-mail: mariya.kh@gmail.com, mirabergelson@gmail.com, julia.akinina@gmail.com,
ekaterina.iskra@gmail.com, toldova@yandex.ru, olgadragoy@gmail.com

Abstract

In this paper we present a multimedia corpus of Pear film retellings by people with aphasia (PWA), right hemisphere damage (RHD), and healthy speakers of Russian. Discourse abilities of brain-damaged individuals are still under discussion, and Russian CliPS (Clinical Pear Stories) corpus was created for the thorough analysis of micro- and macro-linguistic levels of narratives by PWA and RHD. The current version of Russian CliPS contains 39 narratives by people with various forms of aphasia due to left hemisphere damage, 5 narratives by people with right hemisphere damage and no aphasia, and 22 narratives by neurologically healthy adults. The annotation scheme of Russian CliPS 1.0 includes the following tiers: quasiphonetic, lexical, lemma, part of speech tags, grammatical properties, errors, laughter, segmentation into clauses and utterances. Also analysis of such measures as informativeness, local and global coherence, anaphora, and macrostructure is planned as a next stage of the corpus development.

Keywords: aphasia, brain damage, discourse, Russian

1. Introduction

We present a corpus of Pear film (Chafe, 1980) retellings made by brain-damaged individuals and neurologically healthy speakers of Russian language. The primary aim of the Russian CliPS (Clinical Pear Stories) project is investigation of discourse abilities of people with aphasia (PWA) and right hemisphere damage (RHD) in comparison with neurologically healthy speakers. In the recent years there has been development in the studies of discourse in aphasia and other neurological conditions (Armstrong, 2000; Linnik et al., 2015), however, the effect of lesions in language-dominant and non-language-dominant hemispheres on discourse production and comprehension is still discussed. Aphasia is an acquired language impairment resulting from brain damage to the language-dominant hemisphere (usually left; Dronkers and Baldo, 2010). Aphasia of different types can manifest in disturbances in both production and comprehension of language on different levels: phonetic, lexical and syntactic. However, the research shows the discrepancy between the language competence of PWA on micro- and macro-linguistic levels (Armstrong, 2000; Linnik et al., 2015; Wright, 2011). Though early studies show that discourse structure is not impaired in aphasia (Ulatowska et al., 1983a, 1983b), some of the recent research demonstrates that it is not necessarily true. While several studies report that such discourse properties as informativeness and coherence are significantly different in discourse of PWA and healthy speakers (Van Leer and Turkstra, 1999; Nicholas and Brookshire, 1993; Wright et al., 2010), other studies report the opposite results (Glosser and Deser, 1990; Marini et al., 2005).

The damage to the non-language-dominant hemisphere (usually right) is not directly linked to any problems on micro-linguistic level; however there is evidence that people with RHD experience difficulties in language comprehension and production at discourse level (Brookshire and Nicholas, 1984; Tompkins et al., 1997).

At the present moment the only large corpus of aphasic speech is AphasiaBank (MacWhinney et al., 2011). Discourse elicitation stimuli for AphasiaBank are several pictorial stimuli as well as a picturebook with a Cinderella story. Though texts from AphasiaBank are used for discourse research (for example Richardson et al., 2016), any additional annotation does not become part of the corpus.

The goal of the Russian CliPS project is to create a corpus that could be used as a research tool with its existent annotation, and on the other hand, would be constantly developed by new research and additional information.

2. Corpus compilation

2.1 Speakers

Brain-damaged individuals were recruited in the inpatient departments of Moscow rehabilitation centers. The individuals with aphasia had been admitted to the centers with reported language problems after stroke in the left hemisphere and were diagnosed with chronic aphasia (not less than 6 months post-stroke). Aphasia types were diagnosed using Luria's classification (Akhutina, 2015; Luria, 1972), and the corpus contains stories by people with efferent motor, dynamic, acoustic-mnestic and sensory aphasia.

Aphasia can be generally divided in two different types: with fluent and non-fluent speech output. The non-fluent aphasia types include efferent motor aphasia and dynamic aphasia. Russian CliPS corpus contains 10 stories by individuals with efferent motor aphasia and 9 stories by individuals with dynamic aphasia. Aphasia types with fluent speech output include sensory and acoustic-mnestic aphasia. Russian CliPS corpus contains 10 stories by people with sensory aphasia and 10 stories by people with acoustic-mnestic aphasia.

Group	Number of speakers	Gender	Mean age	Age range	SD
Acoustic-mnestic aphasia	10	5 female, 5 male	51.3	40-68	9.4
Dynamic aphasia	9	5 female, 4 male	51.8	41-68	8.1
Efferent motor aphasia	10	3 female, 7 male	48.6	30-57	8.0
Sensory aphasia	10	4 female, 6 male	59.3	33-81	8.1
RHD	5	2 female, 3 male	50	41-56	12.3
Brain damage (total)	44	19 female, 25 male	52.8	30-81	10.4
Healthy	22	11 female, 11 male	58	25-84	13.9

Table 1. Demographic information on Russian CliPS 1.0 speakers

Individuals with RHD all were at least 6 months post-stroke and were right-handed.

Speakers from the neurologically healthy group did not report any history of neurological disease or head traumas. All the participants were native speakers of Russian language. The information about all speakers is summarized in Table 1.

2.2 Material and procedure

The elicitation stimulus, the Pear film, was made at the University of California in Berkeley in 1975 specifically for elicitation and collection of narratives by people from various cultures and languages (Chafe, 1980). It is a color film, and though it is not silent, characters do not produce any language. The film has a unique plot that was written to not resemble any other film, or book, or tale. Some characters of the film are important for the plot, and some just appear for a short moment and do not participate in the story. The film motivates the retellers to provide some moral judgement or interpretation of the story.

For the Russian CliPS corpus all speakers were asked to watch the film and then retell it in detail to the person who had not seen it before (the listener could be present at the time of the retelling, or the experimenter told the speaker that a person would listen to the recording afterwards). Both the experimenter and the listener did not ask any specific questions about the story, but could encourage the speaker with the general questions such as "And what happened next?" or "Would you like to add anything else?" The retelling of the story was audio recorded, also 20 brain-damaged speakers and all healthy speakers gave permission to be recorded on video.

3. Corpus Annotation

The annotation of the corpus was performed in ELAN (Wittenburg et al., 2006). The annotation scheme 1.0 includes the following tiers: quasiphonetic, lexical, lemma, POS, grammatical properties, errors, laughter, segmentation into clauses and utterances.

The quasiphonetic tier (Transcript) is aligned with the media files, and contains orthographic transcript of the speech recorded. Most words in this tier appear in their regular spelling, however in the cases of a phonetic error or a specific pronunciation, the transcript reflects these declinations from regular language. For example, in Russian the word *сейчас 'seychas' – now* in oral speech can appear in its full form or as a reduced variant *щас 'tschas'*. In writing, however, only the full variant is acceptable in standard language. In this case the quasiphonetic transcript should follow pronunciation rather than standard language rules. Phonemic paraphasias (errors) that happen in speech of PWA are also reflected in this tier, for example *велосипел 'velosipel'* (the correct word is *велосипед 'velosiped' – bike*). At the quasiphonetic level all the pauses that are longer than 70ms are annotated, both absolute and filled pauses. If some segment of speech is not comprehensible, the note "incomprehensible" is used.

The quasiphonetic transcript makes it possible to capture some features of oral speech as well as phonemic paraphasias, but it would cause problems for analysis of lexical diversity and lexical density. The lexical transcript tier (Transcript_lex) contains the same information as the quasiphonetic tier, although all the spellings are brought to standard. Lexical transcript is used for calculating lexical richness, because different pronunciations of one lexeme are not counted as different words.

Lemma tier (Lemma) contains initial forms of the words, and in the English lemma tier (Lemma_eng) all the words are translated into English, which, in combination with information from grammatical tiers, makes the data from Russian CliPS available for non-Russian speaking researchers.

The part of speech tagging scheme and the annotation of grammatical categories is based on the manual of Russian National Corpus (http://www.ruscorpora.ru/en/corpora-morph.html).

Laughter is annotated on a separate tier (Laughter) and is aligned with the sound wave. This annotation enables the analysis of laughter as a marker of failure to produce a correct word or interpretation of an event in the stimulus film, as well as dissatisfaction with the whole narrative (Khudyakova and Bergelson, 2015).

Grammatical, semantic and phonetic errors are annotated in the special tier (Error). Phonetic errors include replacement of one sound with another, for example *сапка 'sapka'* (the correct word is *шапка 'shapka'* – hat), omission of a phoneme or inclusion of an extra one, for example *поропал 'poropal'* (the correct word is *пропал 'propal'* – got lost), and use of a word that is phonetically similar with the intended one, but is distant semantically, for example *грустные 'grustnye' – sad*, instead of *груши*

'grushi' – pears. Semantic errors include use of a member of the same semantic category as the target word, for example *apples* instead of *pears*, or *sheep* instead of *goat*. In several cases the distinction between the two types of errors is impossible, for example use of *сановник 'sanovnik' – dignitiary* instead of *садовник 'sadovnik' – gardener* can be interpreted both as a phonetic error (replacement of /d/ with a /n/) or as a semantic one (use of a wrong word from 'professions' category), and in this case both types of error are annotated. Grammatical errors include errors in agreement and number.

Segmentation into clauses is based on grammatical rather than prosodic (Kibrik and Podlesskaya, 2009) principle. Utterances include a main clause with all its subordinate clauses (Glosser and Deser, 1990). A ratio of clauses and utterances can be interpreted as a measure of grammatical complexity (Marini, 2012).

4. Corpus data

The current version of Russian CliPS contains 66 narratives. The total length of the recorded material is 4 hours 33 minutes. The mean length of each recording is 4 minutes 7 seconds (min – 38seconds, max – 18 minutes 27 seconds, SD = 175 seconds).

The quantitative information on the current version of the Russian CliPS corpus is shown in Table 2.

At the present moment the Russian CliPS corpus is not publicly available.

5. Coreference Annotation

We started with coreference annotation. As an annotation tool, we have chosen the platform that was designed for the annotation of RuCoref – Russian Coreference corpus (Toldova et al., 2014; http://ant0.maimbava.net/). This platform was developed for the purpose of anaphora resolution systems evaluation campaign. It is based on MySQL database engine and has a convenient Web-interface that allows parallel annotation by several annotators and on-line tracking of discrepancies between annotators. It supports embedding of annotations, marking zero anaphora, annotation features enhancing by users and establishing links between markables. At start, this platform had built-in annotation scheme worked-out for coreference annotation of written texts (primarily, news). The scheme has a set of features concerning NP structure type, referential status type and coreferent NPs relation type.

We have completed pilot annotation of 10 narratives by healthy speakers and 5 narratives by PWA (efferent motor). This pilot stage revealed steps needed for adaptation of the coreference annotation scheme designed for written texts for a genre of oral narrations.

Firstly, while in written texts discontinuous noun phrases are a special type of coreference relations (e.g. a referent is a group of two other referents before mentioned in a text), in oral texts disrupted NPs are a very frequent phenomenon. The disruption is due to pauses, discourse markers and filler words. The NP disruptions are even more frequent in narratives by people with non-fluent types of aphasia. Thus, we have to create additional functionality for our platform, namely, the annotation of two disrupted text pieces as one markable.

Another problem is that the standard relation between two NPs denoting the same entity in written texts is a

Group		Narrative length (ms)	Pauses (%)	Narrative length (words)	Narrative length (clauses)	Narrative length (utterances)
Acoustic-mnestic aphasia	**Mean**	231 196	43	281,1	52,6	45
	Range	85 229 - 473 025	25-55	76-480	18-84	16-69
	SD	106 700	10	122,2	19,7	16,6
Dynamic aphasia	**Mean**	406 023	60	220,4	39,8	38,8
	Range	138 096 – 810 867	29-71	135-371	27-59	26-59
	SD	196 132	13	91,1	9,4	9,9
Sensory aphasia	**Mean**	275 765	40	346,4	66,3	58,9
	Range	148 023 – 549 223	24-56	170-631	28-110	25-94
	SD	117 912	9	174,7	29,6	25,6
Efferent motor aphasia	**Mean**	377 137	45	228,8	49,9	43,8
	Range	167 879 – 1 107 112	26-72	58-436	14-91	14-64
	SD	275 043	14	119,7	24,4	17,8
RHD	**Mean**	195 922	49	279	63	55,7
	Range	122 845 – 427 025	39-65	185-477	32-120	29-105
	SD	147 132	11	133,5	39,2	33,8
Healthy speakers	**Mean**	152 437	33	269,5	53,7	42,2
	Range	47 389 – 296 805	17-51	88-405	16-80	9-71
	SD	62 524	9	113,7	21,7	18,4

Table 2. Quantitative data on Russian CliPS 1.0

coreferential relation (the relation of referent identity), though other relations such as apposition (c.f. *a 10-year boy*, *the one with a basket*, …) or predicative relation (c.f. *this boy is a boy who* …) are taken into consideration in the built-in scheme. It oral texts, some NPs denoting the same referent as a previous NP are just mere NP repetitions (c.f. *a boy, this boy, went*…), or self-correcting (*a gardener, a farmer, went*…). Thus, in order to distinguish the latter from apposition we need additional labels for NPs relational types (e.g. repetition, renaming). We also need additional rules in the annotation instruction for the differentiation of appositions vs. different types of repetitions.

The third issue worth mentioning concerns the naming problem in speech-impaired people. These are the cases of semantic paraphasias (c.f. *apples* instead of *pears*). Sometimes the speakers make self-correction during the narration. These cases are also should be captured by our scheme. The annotation scheme should also allow marking potential coreference relations between an NP and more than one potential referent.

Our pilot study has highlight some peculiarities of coreference chaining in oral discourse and in speech of different kinds of brain-damaged individuals and some special issues in annotation process for this type of discourse. Thus, this discourse level needs further investigation and the coreference annotation scheme needs further enhancing and adaptation.

6. Future Work

The Russian CliPS corpus at its present stage is annotated on micro-linguistic level. Much discourse annotation is still needed in order to evaluate the discourse abilities of brain-damaged speakers.

The next version of Russian CliPS will also have annotation of informativeness, global and local coherence, and macrostructure of discourse.

7. Bibliographical References

Akhutina, T. (2015). Luria's classification of aphasias and its theoretical basis. *Aphasiology*, 1–20. doi:10.1080/02687038.2015.1070950.

Armstrong, E. (2000). Aphasic discourse analysis: The story so far. *Aphasiology* 14, 875–892. doi:10.1080/02687030050127685.

Brookshire, R. H., and Nicholas, L. E. (1984). Comprehension of directly and indirectly stated main ideas and details in discourse by brain-damaged and non-brain-damaged listeners. *Brain and language* 21, 21–36.

Chafe, W. (1980). *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. , ed. W. Chafe Norwood, New Jersey: Ablex.

Dronkers, N. F., and Baldo, J. V. (2010). "Language: Aphasia," in *Encyclopedia of Neuroscience* (Elsevier Ltd), 343–348.

Glosser, G., and Deser, T. (1990). Patterns of Discourse Production among Neurological Patients with Fluent Language Disorders. *Brain and Language* 49, 67–88.

Khudyakova, M. V., and Bergelson, M. B. (2015). Interpretation of "Embarrassement" Laughter in Narratives by People with aphasia and Non-language-impaired speakers. in *Proceedings of the 4th Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech, 14-15 April 2015* (Enschede).

Kibrik, A. A., and Podlesskaya, V. I. eds. (2009). *Night Dream Stories: A corpus study of spoken Russian discourse*. Moscow: Languages of Slavonic Culture.

Van Leer, E., and Turkstra, L. (1999). The effect of elicitation task on discourse coherence and cohesion in adolescents with brain injury. *Journal of Communication Disorders* 32, 327–349. doi:10.1016/S0021-9924(99)00008-8.

Linnik, A., Bastiaanse, R., and Höhle, B. (2015). Discourse production in aphasia : a current review of theoretical and methodological challenges. 7038. doi:10.1080/02687038.2015.1113489.

Luria, A. R. (1972). Aphasia reconsidered. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior* 8, 34.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology* 25, 1286–1307. doi:10.1080/02687038.2011.589893.

Marini, A. (2012). Characteristics of narrative discourse processing after damage to the right hemisphere. *Seminars in Speech and Language* 33, 68–78. doi:10.1055/s-0031-1301164.

Marini, A., Carlomagno, S., Caltagirone, C., and Nocentini, U. (2005). The role played by the right hemisphere in the organization of complex textual structures. *Brain and Language* 93, 46–54. doi:10.1016/j.bandl.2004.08.002.

Nicholas, L. E., and Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of speech and hearing research* 36, 338–350.

Richardson, J. D., Dalton, S. G., Richardson, J. D., Grace, S., Main, D., Richardson, J. D., et al. (2016). Main concepts for three different discourse tasks in a large non-clinical sample. 7038. doi:10.1080/02687038.2015.1057891.

Toldova, S., Roytberg, A., Ladygina, A., Vasilyeva, M., Azerkovich, I., Kurzukov, M., et al. (2014). RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"* 13, 681–695.

Tompkins, C. A., Baumgaertner, A., Lehman, M. T., and Fossett, T. R. D. (1997). Suppression and discourse comprehension in right brain-damaged adults: A preliminary report. *Aphasiology* 11, 505–519.

Ulatowska, H. K., Doyel, A. W., Stern, R. F., Haynes, S.

M., and North, A. J. (1983a). Production of procedural discourse in aphasia. *Brain and language* 18, 315–341. doi:10.1016/0093-934X(83)90023-8.

Ulatowska, H. K., Freedman-Stern, R., Doyel, A. W., Macaluso-Haynes, S., and North, A. J. (1983b). Production of Narrative Discourse in Aphasia. *Brain and Language* 19, 317–334. doi:10.1016/0093-934X(83)90074-3.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. in *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.

Wright, H. H. (2011). Discourse in aphasia: An introduction to current research and future directions research. *Aphasiology* 25, 1283–1285.

Wright, H. H., Koutsoftas, A., Fergadiotis, G., and Capilouto, G. (2010). Coherence in Stories told by Adults with Aphasia. *Procedia - Social and Behavioral Sciences* 6, 111–112. doi:10.1016/j.sbspro.2010.08.056.