D.O.I. http://dx.doi.org/10.1590/2176-4573e55885

ARTICLES

Discourse Diversity Database (3D) for Clinical Linguistics Research: Design, Development, and Analysis / Discourse Diversity Database (3D) *para pesquisa em linguística clínica: projeto, construção e análise*

Khudyakova Mariya[*]
Antonova Natalia[**]
Nelubina Maria[***]
Surova Anastasia[****]
Vorobyova Anna[*****]
Minnigulova Alina[******]
Gronskaya Natalia[*******]
Yashin Konstantin[********]
Medyanik Igor[*********]
Shishkovskaya Tatiana[**********]
Ryazanskaya Galina[**********]
Zuev Andrey[**********]
Dragoy Olga[**********]

ABSTRACT

Discourse Diversity Database (3D) is a *corpus* designed for clinical linguistics research. It consists of oral speech samples of three different genres: picture-elicited narratives, personal stories, and picture-based instructions. The sub-sections of 3D include recordings by Russian speakers from three independent groups: people with brain tumors

[*] HSE University, Center for Language and Brain, Moscow, Russia; https://orcid.org/0000-0002-5293-3991; mariya.kh@gmail.com
[**] HSE University, Faculty of Humanities, Center for Language and Brain, Nizhny Novgorod, Russia; https://orcid.org/0000-0003-4844-7218; natalie.eskadron@gmail.com
[***] HSE University, Faculty of Humanities, Center for Language and Brain, Nizhny Novgorod, Russia; https://orcid.org/0000-0001-6040-9180; marnelyubina@gmail.com
[****] HSE University, Faculty of Humanities, Center for Language and Brain, Nizhny Novgorod, Russia; https://orcid.org/0000-0003-2800-0929; asurova909@gmail.com
[*****] HSE University, Faculty of Humanities, Center for Language and Brain, Nizhny Novgorod, Russia; https://orcid.org/0000-0003-0043-2244; vorobyovaaa2015@gmail.com
[******] HSE University, Center for Language and Brain, Moscow, Russia; https://orcid.org/0000-0002-5568-8311; alinaminnigulovahouse@gmail.com
[*******] HSE University, Faculty of Humanities, Center for Language and Brain, Nizhny Novgorod, Russia; https://orcid.org/0000-0003-0593-2395; ngronskaya@hse.ru
[********] Privolzhsky Research Medical University, Nizhny Novgorod, Russia; https://orcid.org/0000-0002-5723-7389; jashinmed@gmail.com
[*********] Privolzhsky Research Medical University, Nizhny Novgorod, Russia; https://orcid.org/0000-0002-7519-0959; med_neuro@inbox.ru
[**********] Mental Health Research Center, Moscow, Russia; tszyszkowska@gmail.com
[**********] University of Potsdam, Cognitive Systems Program, Postdam, Germany; galka1999@gmail.com
[**********] National Medical and Surgical Center named after N. I. Pirogov, Moscow, Russia; https://orcid.org/0000-0003-2974-1462; mosbrain@gmail.com
[**********] HSE University, Center for Language and Brain, Moscow, Russia; https://orcid.org/0000-0002-6777-5164; olgadragoy@gmail.com

before and after tumor removal, people with schizophrenia, and neurologically healthy individuals. This article is devoted to the description of the data collection, the annotation scheme, and the specific characteristics of each sub-section of the *corpus*.
KEYWORDS: *Corpus* linguistics; Clinical linguistics; Brain tumors; Schizophrenia; Spoken discourse; Discourse Diversity Database

RESUMO
O Discourse Diversity Database (3D) *é um* corpus *desenvolvido para a pesquisa em linguística clínica. Ele consiste de amostras de fala oral de três gêneros diferentes: narrativas induzidas por imagens, histórias pessoais e instruções baseadas em imagens. As subdivisões do 3D incluem gravações de falantes de russo de três grupos independentes: pessoas com tumores cerebrais antes e depois da remoção do tumor, pessoas com esquizofrenia e indivíduos neurologicamente saudáveis. O presente artigo é dedicado à descrição do procedimento de coleta de dados, do esquema de anotação e das características específicas de cada subdivisão do* corpus.
PALAVRAS-CHAVE: Linguística de corpus*; Linguística clínica; Tumores cerebrais; Esquizofrenia; Discurso oral;* Discourse Diversity Database

Corpus analysis of spoken discourse allows conducting multidimensional assessment of speech in people with various language impairments and neurological and psychiatric disorders, as well as in healthy speakers. Annotated *corpora* are an important source for fundamental research in neuro- and psycholinguistics, automated analysis of language in clinical populations, and speech-language pathology. In this paper, we present the Discourse Diversity Database (3D), which is a collection of audio recordings by Russian speakers with brain tumors before and after tumor removal, people with schizophrenia spectrum disorders, and neurologically and mentally healthy individuals. The structure of the paper is as follows: in Section 1, we provide an overview of some of the existing clinical *corpora*; in Section 2, we describe the specific characteristics of speech in people with brain tumors and schizophrenia, as well as in healthy people depending on their age and condition; in Section 3, we describe the motivation for collecting various genres and the stimuli used for speech elicitation in 3D; in Section 4, we describe the sub*corpora* of 3D , including the participant data and data collection procedure; in Section 5, we provide an overview of the annotation scheme.

1 Oral Speech *Corpora* in Clinical Linguistics: An Overview

In clinical linguistics, *corpus* analysis of spoken discourse mainly follows two aims. The first one is the investigation of specific language deficits on various linguistic levels and how they depend on the specific characteristics of the patients and the diagnoses. This aim is mostly a part of fundamental research, although the results can be further used for improvement of assessment criteria and speech therapy. Such *corpora* are generally annotated on various levels by humans. The second aim is training models for the automated analysis of speech which could be used to detect early symptoms of different disorders; such *corpora* do not always have manual annotation.

One of the largest and most well-known databases of discourse speech samples from different populations is TalkBank (https://talkbank.org, MacWhinney, 2007) containing five clinical *corpora*: Aphasiabank (MacWhinney *et al.*, 2011), DementiaBank (Forbes; Fromm; MacWhinney, 2012), RHDBank for language in right hemisphere damage (Minga *et al.*, 2021), TBIBank for language in traumatic brain injury, and ASD Bank for language in autism. The TalkBank *corpora* contain a set of discourse tasks such as free speech samples, picture descriptions, storytelling tasks, and procedural discourse, all collected according to one protocol. The recordings are annotated according to the Codes for the Human Analysis of Transcripts (CHAT) format (MacWhinney, 2010) and coded for analysis with the Computerized Language Analysis (CLAN) program (MacWhinney, 2017). The annotation provides information about fluency, information content, lexical devices, disfluencies, and lexical and grammatical errors. Although TalkBank *corpora* contain discourse samples in over 41 different languages, most of the samples are in English.

There is a variety of *corpora* in different languages with a focus on annotation on different language levels and the collection of different discourse types. For example, the Cambridge Cookie-Theft *Corpus* (Williams *et al.*, 2010) contains picture descriptions and spontaneous speech samples from people with brain damage and healthy controls which were subsequently annotated using the orthographic transcription in the Praat program (Boersma; Weenink, 2005). The Greek *Corpus* of Aphasic Discourse (Varlokosta, 2016) was manually annotated in ELAN (Wittenburg *et al.*, 2006) but according to a different annotation scheme which included speech and non-speech events, micro-linguistic

features (words, POS, grammatical, semantic, phonological errors, clause types, etc.) and discourse features such as narrative structure units, main events, and evaluation devices. Similarly, the Russian Clinical Pear Stories *corpus* (Russian CliPS; Khudyakova *et al.*, 2016), contains retellings of the Pear film (Chafe, 1980) by patients with brain damage and neurologically healthy Russian speakers annotated in ELAN on micro- and macro-linguistic levels (cf. Bergelson; Khudyakova, 2017). The Night dream stories *corpus* (Kibrik; Podlesskaya, 2009) was created with the focus on phonetic and prosodic features of speech; stories about night dreams by Russian-speaking children with and without neurotic disorders are annotated with attention to pause types and functions, discursive accents, illocutionary and internal phases with a distinction between their canonical and non-canonical realizations.

Not all clinical *corpora* provide extensive linguistic annotation; some contain only basic transcription and are mostly used for automated analysis. For example, in the Carolina Conversations Collection, a database of conversations with people with Alzheimer's disease, the utterances are orthographically transcribed, and additional information, such as speech rate is calculated automatically (Davis; Pope, 2011). With the development of automatic spontaneous speech analysis, considerable attention has been given to speech biomarkers. Nevler and colleagues (2019) conducted research where they aimed to retrieve specific biomarkers of prosody from the acoustic characteristics of speech in patients suffering from primary progressive aphasia and in a control group of healthy participants. They also used automatic speech analysis protocol to retrieve and subsequently analyze such measures as: fundamental frequency, speech, and silent pause durations (Nevler *et al.*, 2019).

Some of the clinical *corpora* include spoken language samples of various sizes, ranging from single phonemes to small discourse passages. For example, the PRAUTOCAL *corpus* of speech in Down syndrome (Escudero-Mancebo *et al.*, 2021) contains sentences obtained from speakers with Down syndrome during a video game and qualitatively assessed by several experts. The EasyCall is a dysarthric speech dataset of commands most likely to be used in a voice-controlled contact application (Turrisi *et al.*, 2021). In the Atlanta Motor Speech Disorders *Corpus* (Laures-Gore *et al.*, 2016) the data include single vowels, single words, sentences and discourse passages from people with motor speech disorders who speak different dialects of English. Similarly, the

Carcinologic Speech Severity Index *corpus* (C2SI; Woisard *et al.*, 2021) consists of audio samples of various length: single sustained vowels, pseudowords, sentences, read-aloud passages, and spontaneous speech samples by patients after cancer treatment: surgery, radiotherapy, and chemotherapy. It includes acoustics, prosody qualitative assessment, and automatic analysis.

2 Speech in Different Populations

2.1 Language Before and After Brain Tumor Removal

Damage to brain structures critical for language production results in speech difficulties and disabilities. For example, brain tumors located in cortical areas and white matter tracts associated with language might lead to permanent speech impairments. Thus, persisting aphasia is common for patients with brain tumors who develop pathological brain conditions over a prolonged period of time. However, the gradual change leaves time for functional reorganization of language function due to neuroplasticity that is to some extent possible at any age (Brodtmann *et al.*, 2012; CAI *et al.*, 2016). Very often people with brain tumors preserve generally intact language processing and do not induce the appearance of neurological deficit, even though the pathological brain tissue may be large in volume and located in eloquent areas (Anderson; Damasio; Tranel, 1990; Duffau, 2005). Furthermore, even after tumor removal, language deficits are transient and can be observed directly after the surgery with the subsequent disappearance after several weeks or months according to results of standard clinical tests (Duffau, 2005; Wilson *et al.*, 2015).

Still, after neurosurgical treatment many patients encounter problems with daily communication that significantly affect quality of life (Papagno *et al.*, 2012). Currently the nature of these communication difficulties in early and late postsurgical periods is still understudied. It is not reliably established on which level the language impairment occurs and what the dynamics of changes of communication status after surgery are. Standard clinical tools for language assessment preclude a detailed characterization of the peak of human language ability – connected speech. The neurosurgical sub-section of the 3D *corpus* contains discourse samples by people before and after brain tumor removal, as

well as information about their scores on a standardized language assessment test and neuroimaging data.

2.2 Language in Schizophrenia Spectrum Disorders

Schizophrenia is a severe mental condition characterized by distorted perception of reality and disorganized behavior on one side and significant cognitive and emotional decline on the other (Owen; Sawa; Mortensen, 2016). One of the basic symptoms of schizophrenia since the introduction of the term has been formal thought disorder (FTD; Peralta; Cuesta, 2011). FTD refers to aberrations in the thought process, which usually present as speech and language disturbance. The most comprehensive classification so far divides FTD into two groups: positive, e.g., derailment, tangentiality, loose associations, and negative, e.g. alogia, and thought blocking (Cavelti *et al.*, 2018).

Incoherent or disordered speech is one of the key characteristics of FTD and an important diagnostic criterion. It is believed to be reflective of disruptions in normal thought processes (such as the ones that arise in FTD, see Hart; Lewine, 2017). There are two main types of theoretical frameworks explaining the origins of discourse incoherence observed in schizophrenia: executive dysfunction theories (also known as impaired cognition theories) and loose association theories (see Ditman; Kuperberg, 2010 for a review). The former theories state that the lack of control over the process of thinking is typical of negative thought disorder. The latter, on the other hand, explain the incoherences in terms of tangentiality and loose associations that are characteristic of positive thought disorder.

Depending on the type of FTD and the severity of the disease, speech can be affected on various linguistic levels (see Kuperberg, 2010 for review). People with schizophrenia can produce less predictable words (Salzinger *et al.*, 1970; Hart; Payne, 1973; Salzinger; Portnoy; Feldman, 1979) and more neologisms in discourse, non-normal pausation patterns (Spitzer *et al.*, 1994), and more grammatical and lexical errors than healthy speakers (Marini *et al.*, 2008). Discourse by people with schizophrenia is also characterized by lower lexical diversity. Studying language and speech abnormalities in schizophrenia patients with FTD is of high practical significance, because such studies

can provide more objective and easy-to-use diagnostic tools (like automated speech analysis) and support evidence-based classifications.

One of the important issues in mental disorders research is the definition of the norm. According to the International Classification of Diseases 10th Revision (ICD-10; World Health Organization (WHO), 1993) and Diagnostic and Statistical Manual of mental disorders, fifth edition (DSM-5; American Psychiatric Association, 2013), the most used contemporary clinical classifications of mental and behavioral disorders, a condition can only be qualified as a mental disorder if it is associated with either perceived distress or functional disability (Üstün; Kennedy, 2009). In view of this definition, researchers frequently use a self-reported psychiatric norm as a control group. But there is a growing awareness of the limited usefulness of clinical classifications for research purposes, and new classifications are arising. For example, the Hierarchical Taxonomy of Psychopathology (HiTOP; Kotov; Krueger; Watson, 2018) uses clusters of covariating symptoms for diagnosis formulation unlike the traditional clinical classifications (ICD-10 and DSM-5) which use a categorical approach. The categorical approach states the presence or absence of a pathological condition, while the dimensional approach, as in HiTOP, views separate symptoms on a continuum with different degrees of gravity. The dimensional approach is much in demand in psychiatric research nowadays, but it does not allow the separation of "normal" and "ill" people as clearly as the categorical approach. Therefore, we should keep in mind that the same symptoms can be present in different psychiatric conditions and even in the non-clinical population in different grades of severity. Because of that, in research conditions, careful examination of self-reported healthy participants is as important as in the case of diagnosed mental disorders. In the 3D *corpus* we created a Psychiatric norm sub-section containing discourse samples by speakers evaluated by a psychiatrist and not showing any symptoms of mental disorders.

2.3 Variability of Language in Healthy Adults

In clinical linguistics, the analysis of spoken discourse includes evaluation of such characteristics as phrase length, lexical diversity, speech rate, information content, grammatical complexity, paraphasias, informativeness, coherence (Prins; Bastiaanse, 2004; Bryant; Ferguson; Spencer, 2016). For each clinical population, it is important to

have the norms for comparison. Thus, it is crucial to have balanced normative data with the inclusion of all possible variability of the data. Below we overview some of the factors that can affect speech characteristics in healthy speakers.

Age is one the influential factors that can affect speech characteristics in healthy adults. Several studies claim that healthy adults demonstrate changes in spoken language with declining performance involving different language domains (Nadeau, 2019). Unlike young people, older adults demonstrate an increase in the total number of words (BORTFELD *et al.*, 2001), lower information content (Saling; Laroo; Saling, 2012), more tip-of-the-tongue states (e.g., Burke; Shafto, 2004; Gollan; Brown, 2006; Abrams; Farrell, 2011), difficulties in producing and comprehending syntactically complex or ambiguous sentences (e.g., Kemtes; Kemper, 1997; Kemper; Herman; Lian, 2003; Kemper; Crow; Kemtes, 2004) and larger vocabularies (Verhaeghen, 2003). The trend of producing more words by older people co-occurred with an increase in the number of disfluencies such as lexical fillers, non-lexical fillers, word repetitions, lengthy silent pauses, and empty words (Kemper *et al.*, 1990; Heller; Dobbs, 1993; Bortfeld *et al.*, 2001). These age-related increases in disfluencies are thought to result from older adults having more word retrieval problems (Lovelace; Twohig, 1990; Bortfeld *et al.*, 2001), as disfluencies could serve the purpose of giving them more time to locate the intended word.

3 Discourse Diversity Database

3.1 Discourse Types and Their Speech Characteristics

Discourse types, or genres, differ in their speech characteristics. In clinical linguistics, quite short speech samples are analyzed, and the type of discourse depends on the elicitation task. The most widely used methods of discourse elicitation are tasks containing a picture or picture sequence description (Williams *et al.*, 2010; Bryant; Ferguson; Spencer, 2016), discourse narrative which implies telling a personal story or retelling well-known stories or plots (Behrns *et al.*, 2009; Olness; Ulatowska, 2011), procedural discourse (Ulatowska; North; Macaluso-Haynes, 1981; Stark, 2019), or conversations (Webster; Morris, 2019).

Different discourse elicitation tasks involve different cognitive processes (Olness, 2006; Fergadiotis; Wright, 2011; Gorno-Tempini *et al.*, 2011; Stark, 2019). For example, there is a difference in language performance in narrative discourse and picture description tasks, with the former expressed in more complex language while telling a personal story or retelling something (Fergadiotis; Wright, 2011; MacWhinney *et al.*, 2011). Narrative discourse tasks without visual stimuli elicit more variable speech and higher lexical diversity (Fergadiotis; Wright, 2011; Stark, 2019) than picture description tasks where plenty of descriptive words are observed (Olness *et al.*, 2002). Procedural discourse tasks, on the other hand, presuppose a strict staged scheme which leads to more frequent use of action words (Pritchard *et al.*, 2015).

We presume that for the proper assessment of individual linguistic ability in all aspects of real speech different elicitation tasks should be used. We chose three different elicitation tasks, with and without pictorial stimuli, of two different genres - narrative and procedural discourse.

3.2 3D Elicitation Tasks

For the 3D *corpus*, we collected and analyzed discourse samples across three elicitation tasks: picture-elicited narratives, personal stories, and picture-based instructions (procedural discourse). Each of the types of tasks contained three variants of stimuli. For picture-elicited narratives task, we used one of the three comics by Herluf Bidstrup ("Superman," "Discovery of the World," "Wonderful Day") for discourse elicitation based on sequential pictures. To elicit personal stories, we used one of three questions about notable occasions in the participant's life: (1) Please tell me about the best or the most memorable gift you have received; (2) Please tell me about the best or the most memorable trip you have gone on; (3) Please tell me about the best or the most memorable party you have had. As stimuli for the picture-based instructions, we used IKEA's self-assembly furniture manuals for a chair, a table, or a bench. In the subsections, different procedures were used for balance and randomization (see Table 1 for details). The order of the tasks was fixed.

4 Subsections

4.1 General Overview

In the 3D *corpus* there are two collections: adults with neurologic and psychiatric diagnoses and neurotypical adults. The collection consists of discourse samples from two clinical groups: (1) patients with brain tumors (N=45), and (2) people with schizophrenia spectrum disorders (N=26), and three normative sub-sections: 3) self-reported healthy adults, ages 18-80 (N =84), (4) young adults without mental disorders assessed by a psychiatrist (N=22), and (5) self-reported neurologically healthy young adults recorded at two time points: in an active state and in the state of fatigue (N=10). Data collection for 3 is completed, and for 1, 2, 4, 5 is ongoing. The summary of the subsections is presented in Table 1.[1]

Table 1. Sub-sections of the 3D *corpus*

	Sub-sections				
	Clinical				
	Neurosurgery	Schizophrenia	Age-balanced self-reported norm	Psychiatric norm	Norm in active and tired state
N participants	74	26	84	22	10
N time points	3	1	1	1	2
Stimuli distribution	Balanced across experimental lists, randomized	Two picture description tasks (Sportsman and Adventure)	Balanced across experimental lists, randomized	Two picture description tasks (Sportsman and Adventure)	Balanced across experimental lists, pseudorandomized
Age	M=49.7, SD=14.6	M=28.8, SD=4.3	18-29 y.o. (M = 21.2, SD = 2.6); 30-49 y.o. (M = 38.1, SD = 6.6); 50-64 y.o. (M= 57, SD = 3.8); 65+ y.o. (M = 72, SD = 7.0)	M=23.9, SD=4.3	M=28.80, SD=2.86
Diagnoses	Brain tumors	Schizophrenia, schizoaffective disorder	no	no	no

[1] Detailed information about the participants, stimuli distribution across lists and additional data can be found at https://osf.io/2wvdz/?view_only=a38147409b5042bbabf1fc7560b32805. At the present moment, the 3D *corpus* is not publicly available.

Meta-data	MRI RAT (Russian Aphasia Test) CETI (Communicative Effectiveness Index)	ICD-10 PANSS		SCL-90-R PANSS	Test of differential self-evaluation of one's functional state

4.2 Neurosurgery Subcorpus

4.2.1 Participants

Discourse samples from patients who underwent surgery for tumor resection were collected in the Privolzhsky Research Medical University (PIMU) in Nizhny Novgorod and National Medical and Surgical Center named after N. I. Pirogov in Moscow. The eligibility criteria for the clinical group required individuals to have a tumor in the left hemisphere in areas critical for language production as evaluated based on an MRI scan.

To date, we have tested 74 native Russian speakers (31 women, mean age=49.7, SD=14.6, age range – 19-72; mean N of years of education 14, SD=2.7). Three patients refused to perform discourse tasks 1-2 days before the surgery due to severe speech deficit or due to refusal. Six patients were not tested within 3-7 days after the surgery with discourse tasks and Russian Aphasia Test (RAT; Ivanova et al., 2019) because of test organization issues, medical reasons or due to patient refusal. Five patients did not perform all three discourse tasks 3-7 days after the surgery due to patient refusal, severe speech deficit or medical reasons. 28 patients did not complete language testing with discourse tasks and/or RAT three months after the surgery due to medical reasons or impossibility of testing and seven patients by cause of death. Other eight patients will be tested three months after the surgery by the end of the year. One patient was excluded from further analysis due to data collection error. We plan to collect discourse samples and RAT at three time points from more than 60 patients. All participants signed the consent form before the experiment.

4.2.2 Procedure

According to a specific mechanism of brain reorganization, samples of connected speech were collected at three time points: 1–2 days before the surgery, within 3–7 days after the surgery, and 3 months after the surgery. Each participant completed three tasks at each time point, and each participant completed different versions of one task at each of the time points. The distribution of the task versions was balanced across the elicitation tasks (see OSF[2] for detailed information). The stimuli were presented on a 10.4 inch tablet. All patients were tested in a quiet place.

At all the time points, the patients also completed the Russian Aphasia Test (RAT; (Ivanova *et al.*, 2019), a comprehensive standardized test for assessment of language production and comprehension on different language levels, and the Token Test (De Renzi; Vignolo, 1962), a standardized tool test for quick assessment of language comprehension in aphasia. Before the operation and three months after the operation, the patients' relatives completed the Communicative Effectiveness Index questionnaire (CETI; Lomas *et al.*, 1989).

4.3 Schizophrenia Spectrum Disorders

4.3.1 Participants

26 patients of a Mental Health Research Center in Moscow, Russia participated in the study (21 women, mean age=28.8, SD=4.3, age range=18–42; mean N of years of education–13.6, SD=2.3). All patients were admitted to the clinic with acute schizophrenia spectrum psychosis (ICD-10 diagnoses: F20 Schizophrenia or F25 Schizoaffective disorder) and had no history of neurologic disorders or substance abuse. All the participants signed the informed consent. The research was approved by the Ethical Committee of the Mental Health Research Center.

[2] https://osf.io/2wvdz/?view_only=a38147409b5042bbabf1fc7560b32805.

4.3.2 Procedure

The discourse samples were collected when the patients were in partial remission, ranking 3—minimally improved or 2—much improved on the Clinical Global Impression Scale (Haro *et al.*, 2003). Each participant completed two versions of the picture-based narrative and one version of the other two tasks. The versions were distributed across experimental lists (see OSF[3] for detailed information). The stimuli were presented on paper in a fixed order.

In addition to standard clinical diagnosis, every patient was evaluated with the standard psychiatric Positive and Negative Symptoms Scale (PANSS; Kay; Fiszbein; Opler, 1987) which estimates positive, negative and general schizophrenia symptoms, as well as the overall severity.

4.4 Age-Balanced Self-Reported Norm

4.4.1 Participants

Normative discourse samples were collected from 84 native Russian speakers without neurological or psychiatric disorders. The participants were from four age groups: 18–29 y.o. (N=21; 16 women; mean age=21.2, SD=2.6; mean N of years of education 14, SD=2.0), 30–49 y.o. (N=23, 15 women; mean age – 38.1, SD=6.6; mean N of years of education – 16.5, SD=2.9), 50–64 y.o. (N=20, 16 women; mean age – 57, SD=3.8; mean N of years of education – 16.4, SD=2.1), and 65+ y.o. (N=20, 15 women; mean age – 72, SD=7.0; mean N of years of education – 16, SD=3.1). All participants signed the consent form before the experiment. The research was approved by the HSE Committee on Interuniversity Surveys and Ethical Assessment of Empirical Research.

[3] https://osf.io/2wvdz/?view_only=a38147409b5042bbabf1fc7560b32805.

4.4.2 Procedure

Normative discourse samples for 18–29, 30–49 and 50–64 age groups were collected online on the Finding Five platform (FindingFive, 2019). Discourse samples in the normative 65+ age group were obtained using printed versions of stimuli or while presenting the stimuli on a laptop or a 10.4 inch tablet in a quiet place. Each participant completed one version of each task. The distribution of the task versions was balanced across elicitation tasks (see OSF[4] for detailed information).

4.5 Psychiatric Norm

4.5.1 Participants

48 participants with no history of psychiatric or neurological disorders, alcohol or substance abuse took part in the study. After completing the questionnaires and the psychiatric exam, 22 participants (19 women, mean age=23.9, SD=4.3, age range – 20–36; mean N years of education 15.7, SD=1.7) participated in the study. All participants signed the informed consent. The research was approved by the Ethical Committee of the Mental Health Research Center.

4.5.2 Procedure

All participants completed the online version of the Symptom Checklist-90-Revised (SCL-90-R; Derogatis; Savitz, 1999), a list of the most prominent psychiatric symptoms, commonly used for screening purposes. SCL-90-R assesses nine symptom dimensions, such as somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation and psychoticism. To select the participants with a lower chance of undiagnosed psychiatric illness we set the threshold of the General Symptom Index (GSI) = 0.55, based on the previously established threshold for Russian students 17–20 years without any diagnosed psychiatric

[4] https://osf.io/2wvdz/?view_only=a38147409b5042bbabf1fc7560b32805.

disorders (Kioseva, 2016). Then, 27 participants were invited for a psychiatric interview to exclude participants having schizotypal traits, and 22 of them were qualified as a psychiatric norm for the current research.

All speech samples were collected online via Skype. Also, all participants were evaluated using psychiatric PANSS scale for further comparison with the patients. Each participant completed two versions of the picture-based narrative and one version of the other two tasks. The versions were distributed across experimental lists (see OSF[5] for detailed information). The stimuli were presented on paper in a fixed order.

4.6 Norm in Active and Tired State

4.6.1 Participants

Ten participants (8 women, mean age=28.80, SD = 2,86, age range=23–33, mean N of years of education 16.90; SD=1.91) with no history of psychiatric or neurological disorders, alcohol or substance abuse took part in the study. All participants signed informed consent.

4.6.2 Procedure

The discourse samples were collected online. Each participant took part in two sessions of the study: in the active state and in the tired state. The participants completed the Test of differential self-evaluation of one's functional state (Doskin *et al.*, 1973) in each session. The versions of the tasks were distributed across experimental lists (see OSF[6] for detailed information).

[5] https://osf.io/2wvdz/?view_only=a38147409b5042bbabf1fc7560b32805.
[6] https://osf.io/2wvdz/?view_only=a38147409b5042bbabf1fc7560b32805.

5 Annotation Scheme

Annotation of the narratives was performed in ELAN (WITTENBURG *et al.*, 2006) on multiple tiers.

5.1 Transcription and Segmentation

The *Transcript* tier is aligned with the media files and contains an orthographic transcript of the recorded speech. Most words in this tier appear in their regular spelling, however in the cases of a phonetic error or a specific pronunciation, the transcription reflects these departures from the linguistic norm. On the Transcript tier all the pauses that are longer than 70ms are annotated, both silent and filled (such as *ah*, *um*) pauses.

The main unit for discourse segmentation is an *elementary discourse unit* (EDU), and it roughly equals a clause. An EDU is a unit containing one predicate, or an omitted predicate that can be semantically restored; in the case of repetition of the predicate resulting from word-finding difficulties, all the repeated lexemes are included in the same EDU (for examples see Bergelson; Khudyakova, 2020). Utterances include a main clause with all its subordinate clauses; the ratio of clauses and utterances can be interpreted as a measure of syntactic complexity (Marini, 2012).

5.2 Microlinguistic Annotation

The *lexical transcript* is a technical tier and contains the same information as the transcript tier segmented into words and non-word elements.

Lemma and *POS* (part-of-speech) tiers contain initial forms and POS labels. The POS tagging scheme and lemmatization is based on the manual of the Russian National *Corpus* (http://www.rus*corpora*.ru/en/*corpora*-morph.html).

Grammatical, semantic and phonetic errors are annotated in the *Error* tier.

The *Non-word* tier contains annotations of silent pauses, filled pauses, false-starts, repetitions, semantically empty words and automatized expressions. Silent pauses are silent speech segments longer than 70 ms, filled pauses are segments longer than 70ms filled with a non-word sound (for example, a*h* or *um*) that is not the beginning of a new

word. False-starts are non-finished word segments (usually one syllable-long), see (1), false-start is marked by =. Repetitions include words and phrases repeated without change (cf. MacWhinney, 2010, p.77), in the *Non-word* tier words are marked as repetitions starting from the second mention, see (1), repetitions are marked in italics.

(1) it was *it was it was* a bi= big dog

5.3 Macrocomponent Annotation

Each EDU is annotated as one of five macrocomponents. *Mainline* EDUs describe either events in the story or the actions in the instructions. *Background* EDUs contain general information about the characters of the story, the setting where the story takes place, or a description of elements and surroundings in the instructions. *Comment* EDUs contain the speaker's thoughts and opinions about the events, characters, and details of the story. Unlike comments, *meta-comments* contain the speaker's attitudes and thoughts about the process of telling a story or giving an instruction, such as word-finding problems. *Regulator* EDUs organize the flow of discourse and do not have any information content.

Final Remarks

The composition and annotation of the 3D *corpus* allows for the investigation of various speech characteristics on several linguistic levels, across discourse genres, and in different populations.

Manual annotations allow us to extract the following measures commonly used in speech assessment: fluency measures (such as pause ratio, speech rate, and articulation rate), mean length of EDUs and utterances in words and milliseconds, lexical diversity, and the number of errors, false starts, and repetitions. Moreover, the macrocomponent annotation allows to extract the data only from the EDUs related to the story line, and exclude comments and regulators from the analysis since tangential EDUs can affect some of the measures such as lexical diversity (Kintz; Fergadiotis; Wright, 2016).

The Neurosurgery sub*corpus* includes speech samples from three time-points, as well as the neuroimaging data, allowing us to create full language profiles of the patients

pre- and post-operatively and investigate the way that brain tumor growth and damage to white matter tracts affects language. In addition, due to the availability of the standardized language assessment data in the Neurosurgery sub*corpus*, we can analyze how deficits on various language levels (phonetic, lexical, and syntactic) manifest in spoken discourse of various genres.

The analysis of discourse from the 3D *corpus* is not limited to extracting data from the available manual annotation. For example, we are currently running an automated analysis of global and local coherence in schizophrenia based on the transcripts from the Schizophrenia and the Psychiatric control sub*corpora* using word2vec models (see method description and assessment in Ryazanskaya; Khudyakova, 2020).

Creating a large database of speech samples from clinical populations and healthy control groups is a time- and labor-intensive process. However, such resources provide a great number of possibilities for fine-grained linguistic research, as well as automated analysis, comparisons of different speaker groups, and the study of the neural substrate of speech.

REFERENCES

ABRAMS, L.; FARRELL, M. T. Language Processing in Normal Aging. *The Handbook of Psycholinguistic and Cognitive Processes: Perspectives in Communication Disorders*, n. 352, pp.49-73, 2011.

AMERICAN PSYCHIATRIC ASSOCIATION. *Diagnostic and Statistical Manual of Mental Disorders*. 5. ed., 2013.

ANDERSON, S. W.; DAMASIO, H.; TRANEL, D. Neuropsychological Impairments Associated with Lesions Caused by Tumor or Stroke. *Archives of neurology*, v. 47, n. 4, pp.397-405, 1990.

BEHRNS, I. *et al*. A Comparison Between Written and Spoken Narratives in Aphasia. *Clinical Linguistics & Phonetics*, v. 23, n. 7, pp.507-528. 13 Jan. 2009. Available on: http://www.ncbi.nlm.nih.gov/pubmed/19585311.

BERGELSON, M.; KHUDYAKOVA, M. Interaction and Empathy as Elements of Narrative Strategies in the Russian CliPS *Corpus*. *In*: *Computational Linguistics and Intellectual Technologies*. Moscow: -RSUH, 2017. pp.55-67.

BORTFELD, H. *et al*. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Lang Speech*, v. 43, n. 2, pp.123-147, 2001.

BRODTMANN, A. *et al*. Changes in Regional Brain Volume Three Months after Stroke. *Journal of the Neurological Sciences*, v. 322, n. 1-2, pp.122-128, 2012.

BRYANT, L.; FERGUSON, A.; SPENCER, E. Linguistic Analysis of Discourse in

Aphasia: A Review of the Literature. *Clinical Linguistics and Phonetics*, v. 30, n. 7, pp.489-518, 2016.

BURKE, D. M.; SHAFTO, M. A. Aging and Language Production. *Current Directions in Psychological Science*, v. 13, n. 1, pp.21-24, 2004.

CAI, J. *et al*. Contralesional Cortical Structural Reorganization Contributes to Motor Recovery after Sub-Cortical Stroke: A Longitudinal Voxel-Based Morphometry Study. *Frontiers in Human Neuroscience*, v. 10, n. August, p.8, 2016

CAVELTI, M. et al. Is Formal Thought Disorder in Schizophrenia Related to Structural and Functional Aberrations in the Language Network? A Systematic Review of Neuroimaging Findings. *Schizophrenia Research* Elsevier B.V. 1, Sep. 2018.

CHAFE, L. Uspec.Rs of Rturrafit' E Produc-Iion. 1980.

DAVIS, B. H.; POPE, C. Finding a Balance: The Carolinas Conversation Collection. Corpus *Linguistics and Linguistic Theory*, v. 7, n. 1, pp.143-161, 2011.

DE RENZI, E.; VIGNOLO, L. A. The Token Test: A Sensitive Test to Detect Receptive Disturbances in Aphasics. *Brain: a Journal of Neurology*, v. 85, pp.665-678, 1962. Available on: http://www.ncbi.nlm.nih.gov/pubmed/14026018.

DEROGATIS, L. R.; SAVITZ, K. L. The SCL-90-R, Brief Symptom Inventory and Matching Clinical Rating Scales. *In*: *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment*, 1999.

DITMAN, T.; KUPERBERG, G. R. Building Coherence: A Framework for Exploring the Breakdown of Links across Clause Boundaries in Schizophrenia. *Journal of Neurolinguistics*, v. 23, n. 3, pp.254-269. 1 May 2010. Available on: https://linkinghub.elsevier.com/retrieve/pii/S0911604409000244. Access on: 29 Dec. 2020.

DOSKIN, V. A. *et al*. A Test of Differential Self-Evaluation of One's Functional State. *Voprosy Psychologii*, 1973.

DUFFAU, H. Lessons from Brain Mapping in Surgery for Low-Grade Glioma: Insights into Associations between Tumour and Brain Plasticity. *Lancet Neurology*, v. 4, n. 8, pp.476-486, 2005.

ESCUDERO-MANCEBO, D. *et al*. PRAUTOCAL *Corpus*: A *Corpus* for the Study of Down Syndrome Prosodic Aspects. *Language Resources and Evaluation*, 2021.

FERGADIOTIS, G.; WRIGHT, H. H. Lexical Diversity for Adults with and without Aphasia across Discourse Elicitation Tasks. *Aphasiology*, v. 25, n. 11, pp.1414-1430, 2011.

FINDINGFIVE. FindingFive: A Web Platform for Creating, Running, and Managing your Studies in one Place. NJ, USA. FindingFive Corporation (nonprofit), 2019. Available on: https://www.findingfive.com/ .

FORBES, M. M.; FROMM, D.; MACWHINNEY, B. AphasiaBank: A Resource for Clinicians. *Seminars in Speech and Language*, v. 33, n. 3, pp.217-222, 2012.

GOLLAN, T. H.; BROWN, A. S. From Tip-of-the-Tongue (TOT) Data to Theoretical Implications in Two Steps: When More TOTs Means Better Retrieval. *Journal of Experimental Psychology: General*, v. 135, n. 3, pp.462-483, 2006.

GORNO-TEMPINI, M. L. *et al*. Classification of Primary Progressive Aphasia and Its Variants. *Neurology*, v. 76, n. 11, pp.1006-1014, 2011. Available on: http://www.scopus.com/inward/record.url?eid=2-s2.0-79952823979&partnerID=40&md5=fc55b3557b983061aa3b1dadc242c006.

HARO, J. M. *et al*. The Clinical Global Impression-Schizophrenia Scale: A Simple Instrument to Measure the Diversity of Symptoms Present in Schizophrenia. *Acta Psychiatrica Scandinavica*, v. 107, n. 416, pp.16-23, 2003. Available on: https://onlinelibrary.wiley.com/doi/full/10.1034/j.1600-0447.107.s416.5.x. Access on: 20 Sep. 2021.

HART, D. S.; PAYNE, R. W. Language Structure and Predictability in Overinclusive Patients. *British Journal of Psychiatry*, v. 123, n. 577, pp.643-652, 1973.

HART, M.; LEWINE, R. R. J. Rethinking Thought Disorder. *Schizophrenia Bulletin*, v. 43, n. 3, pp.514-522, 2017.

HELLER, R. B.; DOBBS, A. R. Age Differences in Word Finding in Discourse and Nondiscourse Situations. *Psychology and Aging*, v. 8, n. 3, pp.443-450. Sep. 1993. Available on: https://psycnet.apa.org/journals/pag/8/3/443. Access on: 1 Apr. 2021

IVANOVA, M. *et al*. Standardizing the Russian Aphasia Test: Normative Data of Healthy Controls and Stroke Patients. *Frontiers in Human Neuroscience*, v. 13, 2019. Available on: http://www.frontiersin.org/Community/AbstractDetails.aspx?ABS_DOI=10.3389%2Fconf.fnhum.2019.01.00088.

KAY, S. R.; FISZBEIN, A.; OPLER, L. A. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bulletin*, v. 13, n. 2, pp.261-276, 1 Jan. 1987. Available on: https://academic.oup.com/schizophreniabulletin/article-lookup/doi/10.1093/schbul/13.2.261. Access on: 9 Aug. 2021.

KEMPER, S. *et al*. Telling Stories: The Structure of Adults' Narratives. *European Journal of Cognitive Psychology*, v. 2, n. 3, pp.205-228 1990.

KEMPER, S.; CROW, A.; KEMTES, K. Eye-Fixation Patterns of High- and Low-Span Young and Older Adults: Down the Garden Path and Back Again. *Psychology and Aging*, v. 19, n. 1, pp.157-170, 2004.

KEMPER, S.; HERMAN, R. E.; LIAN, C. H. T. The Costs of Doing Two Things at Once for Young and Older Adults: Talking While Walking, Finger Tapping, and Ignoring Speech or Noise. *Psychology and Aging*, v. 18, n. 2, pp.181-192, 2003.

KEMTES, K. A.; KEMPER, S. Younger and Older Adults' On-Line Processing of Syntactically Ambiguous Sentences. *Psychology and Aging*, v. 12, n. 2, pp.362-371, 1997.

KHUDYAKOVA, M, v. *et al*. Russian CliPS: a *Corpus* of Narratives by Brain-Damaged Individuals. In: LREC Proceedings, Portoroz, Slovenia. *Anais*... Portoroz, Slovenia: 2016.

KIBRIK, A. A.; PODLESSKAYA, V. I. (ed.). *Night Dream Stories: A* Corpus *Study of Spoken Russian Discourse [Rasskazy o snovidenijah: korpusnoe issledovanie ustnogo russkogo diskursa]*. Moscow: Languages of Slavonic Culture, 2009.

KINTZ, S.; FERGADIOTIS, G.; WRIGHT, H. H. Aging Effects on Discourse

Production. *In*: *Cognition, Language and Aging*. Amsterdam: John Benjamins Publishing Company, 2016. pp.81-106.

KOTOV, R.; KRUEGER, R. F.; WATSON, D. A Paradigm Shift in Psychiatric Classification: The Hierarchical Taxonomy of Psychopathology (HiTOP). *World Psychiatry*, v. 17, n. 1, pp.24-25, 2018.

KUPERBERG, G. R. Language in Schizophrenia Part 1: An Introduction. *Linguistics and Language Compass*, v. 4, n. 8, pp.576-589, 2010.

LAURES-GORE, J. *et al*. The Atlanta Motor Speech Disorders *Corpus*: Motivation, Development, and Utility. *Folia Phoniatrica et Logopaedica*, v. 68, n. 2, pp.99-105, 1 Oct. 2016.

LINNIK, A. *et al*. Linguistic Mechanisms of Coherence in Aphasic and Non-Aphasic Discourse. *Aphasiology*, v. in press, 2021.

LOMAS, J. *et al*. The Communicative Effectiveness Index: Developmental and Psychometric Evaluation of a Functional Communication Measure for Adult Aphasia. *Journal of Speech and Hearing Disorders*, v. 54, n. 1. 1989.

LOVELACE, E. A.; TWOHIG, P. T. Healthy Older Adults' Perceptions of Their Memory Functioning and Use of Mnemonics. *Bulletin of the Psychonomic Society*, v. 28, n. 2, pp.115-118, 1990.

MACWHINNEY, B. The Talkbank Project. *Creating and Digitizing Language* Corpora. pp.163-180, 2007.

MACWHINNEY, B. Part 1: The CHAT Transcription Format. *The CHILDES Project: Tools for Analyzing Talk*, 2010.

MACWHINNEY, B. et al. AphasiaBank: Methods for Studying Discourse. *Aphasiology*, v. 25, n. 11, pp.1286-1307. Nov. 2011. Available on: http://www.tandfonline.com/doi/abs/10.1080/02687038.2011.589893.

MACWHINNEY, B. Tools for Analyzing Talk Part 2: The CLAN Program. *Talkbank.Org*. no. 2000, 2017.

MARINI, A. *et al*. The Language of Schizophrenia: An Analysis of Micro and Macrolinguistic Abilities and Their Neuropsychological Correlates. *Schizophrenia Research*, v. 105, nos. 1-3, pp.144-155. Oct. 2008. Available on: http://www.ncbi.nlm.nih.gov/pubmed/18768300. Access on: 4 Jun. 2014.

MARINI, A. Characteristics of Narrative Discourse Processing after Damage to the Right Hemisphere. *Seminars in Speech and Language*, v. 33, n. 1, pp.68-78, 2012. Available on: http://www.ncbi.nlm.nih.gov/pubmed/22362325.

MINGA, J. et al. Making Sense of Right Hemisphere Discourse Using RHDBank. *Topics in Language Disorders*, v. 41, n. 1, pp.99-122, Jan. 2021. Available on: https://journals.lww.com/10.1097/TLD.0000000000000244. Access on: 2 Feb. 2021.

NADEAU, S. E. Aging-Related Alterations in Language. *Cognitive Changes and the Aging Brain*. pp.106-126, 2019.

NEVLER, N. *et al*. Validated Automatic Speech Biomarkers in Primary Progressive Aphasia. *Annals of Clinical and Translational Neurology*, v. 6, n. 1, pp.4-14, 2019.

OLNESS, G. S. *et al*. Discourse Elicitation with Pictorial Stimuli in African Americans and Caucasians with and without Aphasia. *Aphasiology*, v. 16, nos. 4-6, pp.623-633, 2002.

OLNESS, G. S. Genre, Verb, and Coherence in Picture-Elicited Discourse of Adults with Aphasia. *Aphasiology*, v. 20, n. 2/3/4, pp.175-187, 2006. Available on: http://www.tandfonline.com/doi/abs/10.1080/02687030500472710.

OLNESS, G. S.; ULATOWSKA, H. K. Personal Narratives in Aphasia: Coherence in the Context of Use. *Aphasilogy*, v. 25, n. 11, pp.1393-1413, 2011. Available on: http://www.tandfonline.com/doi/abs/10.1080/02687038.2011.599365 . Access on: 10 Jun. 2014.

OWEN, M. J.; SAWA, A.; MORTENSEN, P. B. Schizophrenia. *The Lancet*, v. 388, n. 10039, pp.86-97, 2016.

PAPAGNO, C. *et al*. Measuring Clinical Outcomes in Neuro-Oncology. A Battery to Evaluate Low-Grade Gliomas (LGG). *Journal of Neuro-Oncology*, v. 108, n. 2, pp.269-275, 2012.

PERALTA, V.; CUESTA, M. J. Neuromotor Abnormalities in Neuroleptic-Naive Psychotic Patients: Antecedents, Clinical Correlates, and Prediction of Treatment Response. *Comprehensive Psychiatry*, v. 52, n. 2, pp.139-145, 2011.

PRINS, R.; BASTIAANSE, R. Analysing the Spontaneous Speech of Aphasic Speakers. *Aphasiology*, v. 18, n. 12, pp.1075-1091, 2004. Available on: http://www.tandfonline.com/doi/abs/10.1080/02687030444000534.

PRITCHARD, M. *et al*. Language and Iconic Gesture Use in Procedural Discourse by Speakers with Aphasia. *Aphasiology*, v. 29, n. 7, pp.37-41, 2015. Available on: http://www.tandfonline.com/doi/pdf/10.1080/02687038.2014.993912.

RYAZANSKAYA G.; KHUDYAKOVA M. Automated Analysis of Discourse Coherence in Schizophrenia: Approximation of Manual Measures. LREC 2020 Language Resources and Evaluation Conference 11-16 May 2020. pp.98-101, 2020.

SALING, L. L.; LAROO, N.; SALING, M. M. When More Is Less: Failure to Compress Discourse with Re-Telling in Normal Ageing. *Acta Psychologica*, v. 139, n. 1, pp.220-224, 2012.

SALZINGER, K. *et al*. The Immediacy Hypothesis and Response-Produced Stimuli in Schizophrenic Speech. *Journal of Abnormal Psychology*, v. 76, n. 2, pp.258-264, 1970.

SALZINGER, K.; PORTNOY, S.; FELDMAN, R. S. The Predictability of Speech in Schizophrenic Patients. *British Journal of Psychiatry*, v. 135, pp.284-287, 1979.

SPITZER, M. *et al*. Contextual Insensitivity in Thought-Disordered Schizophrenic Patients: Evidence from Pauses in Spontaneous Speech. *Language and Speech*, v. 37, n. 2, pp.171-185, 1994.

STARK, B. C. A Comparison of Three Discourse Elicitation Methods in Aphasia and Age-Matched Adults: Implications for Language Assessment and Outcome. *American Journal of Speech-Language Pathology*, v. 28, n. 3, pp.1067-1083, 9 Aug. 2019. Available on: http://pubs.asha.org/doi/10.1044/2019_AJSLP-18-0265. Access on: 28 May 2019.

TURRISI, R. *et al*. EasyCall *Corpus*: A Dysarthric Speech Dataset. 2021.

ULATOWSKA, H. K.; NORTH, A. J. D.; MACALUSO-HAYNES, S. Production of Narrative and Procedural Discourse in Aphasia. *Brain and Language*, v. 13, pp.345-371, 1981.

ÜSTÜN, B.; KENNEDY, C. What Is "Functional Impairment?" Disentangling Disability from Clinical Significance. *World Psychiatry*, v. 8, n. 2, p.82, 2009. Available on: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2691163/. Access on: 20 Sep. 2021.

VARLOKOSTA, S. A Greek *Corpus* of Aphasic Discourse: Collection, Transcription, and Annotation Specifications. *LREC 2016 Workshop Resources and Processing of Linguistic and Extra- Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016)*. no. May, pp.14-21, 2016.

VERHAEGHEN, P. Aging and Vocabulary Scores: A Meta-Analysis. *Psychology and Aging*, v. 18, n. 2, pp.332-339, 2003.

WEBSTER, J.; MORRIS, J. Communicative Informativeness in Aphasia: Investigating the Relationship Between Linguistic and Perceptual Measures. *American Journal of Speech-Language Pathology*, v. 28, n. 3, pp.1115-1126, 9 Aug. 2019. Available on: https://doi.org/10.1044/2019_AJSLP-18-0256.

WILLIAMS, C. *et al*. The Cambridge Cookie-Theft *Corpus*: A *Corpus* of Directed and Spontaneous Speech of Brain-Damaged Patients and Healthy Individuals. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*. pp.2824-2830, 2010.

WILSON, S. M. *et al*. Transient Aphasias after Left Hemisphere Resective Surgery. *Journal of Neurosurgery*, v. 123, n. 3, pp.581-593, 2015.

WITTENBURG, P. *et al*. ELAN: A Professional Framework for Multimodality Research. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*. pp.1556-1559, 2006.

WOISARD, V. *et al*. C2SI *Corpus*: A Database of Speech Disorder Productions to Assess Intelligibility and Quality of Life in Head and Neck Cancers. *Language Resources and Evaluation*, v. 55, n. 1, pp.173-190, 2021.

WORLD HEALTH ORGANIZATION(WHO). *The ICD-10 Classification of Mental and Behavioural Disorders*. World Health Organization, 1993.

Funding

Collection of the Neurosurgery and the Age-balanced self-reported norm sub*corpora* was supported by the Russian Science Foundation, project No. 20-18-00-399. Collection of the Schizophrenia and the Psychiatric norm sub*corpora* was supported by the Center for Language and Brain NRU Higher School of Economics, RF Government Grant, Ag. No. 14.641.31.0004.

Statement of Author's Contribution

Mariya Khudyakova conceived and planned the collection of the *corpus*, the stimuli, and the annotation scheme. Mariya Khudyakova, Natalia Gronskaya and Olga Dragoy planned and supervised the collection of the Self-reported norm and

Neurosurgery subsection. Konstantin Yashin, Igor Medyanik, Andrey Zuev, Alina Minnigulova, Natalia Antonova, Maria Nelubina, Anastasia Surova and Anna Vorobyova collected the data for the Neurosurgery sub-section, Alina Minnigulova, Natalia Antonova and Anastasia Surova analyzed the neuroimaging data for the neurosurgery sub-section, Natalia Antonova, Maria Nelubina, Anastasia Surova and Anna Vorobyova collected the data for the Self-reported norm subsection and annotated the data for the Self-reported norm and Neurosurgery sub-sections.

Tatiana Shishkovskaya and Galina Ryazanskaya collected and annotated the data for the Psychiatric norm and Schizophrenia sub-sections. Mariya Khudyakova collected and annotated data for the Norm in active and tired state sub-section.

Mariya Khudyakova, Natalia Antonova and Tatiana Shishkovskaya wrote the manuscript. All authors provided critical feedback and helped shape the database and manuscript.

The collected data include speech samples from people with various impairments. Such recordings are not to be made publicly available. We put the available information on the OSF platform and provided the link in the article.

Received October 07, 2021

Accepted August 22, 2022

Reviews

Due to the commitment assumed by *Bakhtiniana. Revista de Estudos do Discurso* [*Bakhtiniana. Journal of Discourse Studies*] to Open Science, this journal only publishes reviews that have been authorized by all involved.

Research Data and Other Materials Availability

Data cannot be made publicly available. The collected data include speech samples from people with various impairments. Such recordings are not to be made publicly available. We put the available information on the OSF platform and provided the link in the article.