D.O.I. http://dx.doi.org/10.1590/2176-4573p55885

ARTIGOS

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

*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*[**********]

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

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

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*

*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 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*

A análise do *corpus* de discurso oral permite a realização de uma avaliação multidimensional da fala em pessoas com várias deficiências de linguagem e distúrbios neurológicos e psiquiátricos, assim como em falantes saudáveis. *Corpora* anotados apresentam uma fonte importante para a pesquisa fundamental em neuro e psicolinguística, análise automatizada da linguagem em populações clínicas, e também na área de patologias de fala. Neste artigo, apresentamos o *Discourse Diversity Database* (3D), uma coleção de gravações de áudio coletadas em falantes de russo com tumores cerebrais antes e depois da remoção do tumor, falantes com transtornos do espectro da esquizofrenia e indivíduos neurológica e mentalmente saudáveis. A estrutura do artigo é a seguinte: na seção 1, fornecemos uma visão geral sobre alguns dos *corpora* clínicos existentes; na seção 2, descrevemos as características específicas da fala de pessoas com tumores cerebrais e esquizofrenia, assim como de pessoas saudáveis, dependendo da sua idade e condição; na seção 3, descrevemos os motivos de coletar gêneros variados de discurso e os estímulos usados na elicitação de fala para o 3D; na seção 4, descrevemos os *subcorpora* do 3D, incluindo os dados sobre os participantes e sobre o procedimento da coleta de dados; e, na seção 5, apresentamos uma visão geral sobre o esquema de anotação do *corpus*.

## 1 Os *corpora* das amostras de fala em linguística clínica: uma visão geral

Na linguística clínica, a análise do *corpus* de discurso oral segue dois objetivos principais. O primeiro é a investigação de déficits de linguagem específicos em vários níveis linguísticos e como eles dependem das características específicas dos pacientes e dos diagnósticos. Este objetivo é, de maneira geral, uma parte da pesquisa fundamental, embora os resultados possam ser usados para aperfeiçoar os critérios de avaliação dos pacientes ou em terapia de fala. Tais *corpora* geralmente são anotados em vários níveis manualmente. O segundo objetivo consiste em treinamento de modelos para a análise automatizada da fala que possam ser usados para detectar sintomas precoces de diferentes distúrbios; tais *corpora* nem sempre têm anotação manual.

Um dos maiores e mais conhecidos bancos de dados de amostras de fala coletadas em diferentes populações é o *TalkBank* (https://talkbank.org) (MACWHINNEY, 2007) que contém cinco *corpora* clínicos: *Aphasiabank* (MACWHINNEY *et al*., 2011); *DementiaBank* (FORBES; FROMM; MACWHINNEY, 2012); *RHDBank*, um banco de dados de amostras de linguagem coletadas em sujeitos com lesões no hemisfério direito (MINGA *et al*. 2021); *TBIBank*, o banco de amostras de fala coletadas em sujeitos com traumas cerebrais; e *ASD Bank*, um banco de amostras de fala coletadas em sujeitos com autismo. Os *corpora* de *TalkBank* contêm um conjunto de tarefas para a coleta de amostras de acordo com um mesmo protocolo: as tarefas de indução de fala espontânea, descrição de imagens, contação de histórias e discurso procedural. As gravações são anotadas de acordo com o formato *Codes for the Human Analysis of Transcripts* (CHAT) (MACWHINNEY, 2010) e codificadas para análise com o programa *Computerized Language Analysis* (CLAN) (MACWHINNEY, 2017). A anotação fornece as informações sobre fluência, conteúdo, dispositivos lexicais, disfluências e erros lexicais e gramaticais. Embora os *corpora* de *TalkBank* contenham as amostras de discurso em mais de 41 idiomas diferentes, a maioria das amostras é em inglês.

Há uma variedade de *corpora* em diferentes idiomas com foco na anotação em diferentes níveis de linguagem e na coleta de diferentes tipos de discurso. Por exemplo, o *Cambridge Cookie-Theft Corpus* (WILLIAMS *et al*., 2010) contém descrições de imagens e amostras de fala espontânea produzida por pessoas com lesões cerebrais e indivíduos saudáveis; posteriormente, os dados do *corpus* foram anotados usando a

transcrição ortográfica no programa *Praat* (BOERSMA; WEENINK, 2005). O *Greek Corpus of Aphasic Discourse* (VARLOKOSTA, 2016) foi anotado manualmente com o programa *EUDICO Linguistic Annotator* (ELAN) (WITTENBURG *et al*., 2006), mas usando um esquema de anotação diferente, que incluía os fenômenos linguísticos e extralinguísticos, e também características microlinguísticas (palavras, classes gramaticais, erros gramaticais, semânticos e fonológicos, tipos de orações etc.) e características do discurso, como unidades da estrutura narrativa, eventos principais e dispositivos de avaliação. De maneira parecida, o *corpus Russian Clinical Pear Stories* (*Russian CliPS*) (KHUDYAKOVA *et al*., 2016) contém as narrativas produzidas por falantes de russo com lesões cerebrais e falantes neurologicamente saudáveis que recontam as histórias do filme *Pear film* (CHAFE, 1980), em que todas as narrativas são anotadas com ELAN nos níveis micro e macrolinguísticos (cf. BERGELSON; KHUDYAKOVA, 2017). O *corpus Night dream stories* (KIBRIK; PODLESSKAYA, 2009) foi criado com o foco nas características fonéticas e prosódicas da fala; as narrativas sobre os sonhos noturnos contadas por crianças falantes de russo com e sem distúrbios neuróticos são anotadas com o foco nos tipos e funções de pausas, acentos discursivos, fases ilocucionárias e internas com a distinção entre suas realizações canônicas e não canônicas.

Nem todos os *corpora* clínicos fornecem uma extensa anotação linguística; alguns contêm apenas a transcrição básica e são usados principalmente para a análise automatizada. Por exemplo, em *Carolina Conversations Collection*, um banco de dados de conversas com pessoas com doença de Alzheimer, os enunciados são transcritos ortograficamente enquanto as informações complementares, como a taxa de elocução, são calculadas automaticamente (DAVIS; POPE, 2011). Com o desenvolvimento da análise automática da fala espontânea, uma atenção considerável passou a ser dada aos biomarcadores da fala. Nevler e colegas (2019) realizaram uma série de pesquisas em que procuraram investigar os biomarcadores específicos da prosódia a partir das características acústicas de fala em pacientes com afasia progressiva primária em comparação com um grupo controle de participantes saudáveis. Eles também usaram um protocolo de análise automática de fala para extrair e, posteriormente, analisar as características como frequência fundamental, durações de fala e de pausas silenciosas (NEVLER *et al*., 2019).

Alguns *corpora* clínicos integram amostras de fala de tamanhos diversos que variam de fonemas únicos até curtos trechos de discurso. Por exemplo, o *corpus* da fala de sujeitos com síndrome de Down *PRAUTOCAL* (ESCUDERO-MANCEBO *et al*., 2021) contém frases obtidas de falantes com síndrome de Down pronunciadas durante um videogame, avaliadas qualitativamente por vários especialistas. O *EasyCall* é um banco de amostras de fala disártrica que consiste de comandos com maior probabilidade de serem usados no manuseio de aplicativos controlados por voz (TURRISI *et al*., 2021). No *corpus Atlanta Motor Speech Disorders Corpus* (LAURES-GORE *et al*., 2016), os dados incluem as amostras obtidas de falantes de diferentes dialetos do inglês com distúrbios motores da fala: vogais únicas, palavras únicas, frases e trechos de discurso. De maneira parecida, o *corpus Carcinologic Speech Severity Index* (C2SI) (WOISARD *et al*., 2021) consiste de amostras de áudio de tamanhos variados: vogais sustentadas únicas, pseudopalavras, frases, passagens lidas em voz alta e amostras de fala espontânea de pacientes após o tratamento do câncer (cirurgia, radioterapia e quimioterapia). O *corpus* inclui avaliação qualitativa acústica e prosódica, e também análise automática.

## 2 Fala em diferentes grupos de indivíduos

### 2.1 Linguagem antes e depois da remoção do tumor

As lesões das estruturas cerebrais críticas para a produção da linguagem resultam em dificuldades e deficiências de fala. Por exemplo, tumores cerebrais localizados em áreas corticais e tratos de substância branca associados à linguagem podem levar a deficiências permanentes de fala. Assim, a afasia persistente é comum em pacientes com tumores cerebrais que desenvolvem condições cerebrais patológicas por um período prolongado de tempo. No entanto, a mudança gradual deixa tempo para a reorganização funcional da linguagem devido à neuroplasticidade, que é, até certo ponto, possível em qualquer idade (BRODTMANN *et al*., 2012; CAI et al., 2016). Muitas vezes, as pessoas com tumores cerebrais preservam o processamento da linguagem quase intacto e o aparecimento de déficit neurológico não é observado, embora o tecido cerebral patológico possa ser grande em volume e localizado em áreas eloquentes do cérebro (ANDERSON; DAMASIO; TRANEL, 1990; DUFFAU, 2005). Além disso, mesmo após a remoção do

tumor, os déficits de linguagem são transitórios e podem ser observados diretamente após a cirurgia com o desaparecimento subsequente depois de algumas semanas ou meses, de acordo com os resultados dos testes clínicos padrão (DUFFAU, 2005; WILSON *et al*., 2015).

Ainda assim, após o tratamento neurocirúrgico, muitos pacientes encaram problemas de comunicação diária que afetam significativamente a sua qualidade de vida (PAPAGNO *et al*., 2012). Atualmente, a natureza dessas dificuldades de comunicação em períodos pós-cirúrgicos precoces e tardios ainda é pouco estudada. Não é estabelecido de forma confiável em que nível ocorre o comprometimento da linguagem e quais são as dinâmicas das mudanças no status da comunicação após a cirurgia. As ferramentas clínicas padrão para avaliação da linguagem impedem uma caracterização detalhada do auge da capacidade da linguagem humana, a fala encadeada. A subseção neurocirúrgica do *corpus* 3D contém as amostras de discursos de pessoas antes e depois da remoção do tumor cerebral, assim como as informações sobre suas pontuações em um teste padrão de avaliação de linguagem e os dados de neuroimagem.

### 2.2 Linguagem nos transtornos do espectro da esquizofrenia

A esquizofrenia é uma condição mental grave caracterizada, por um lado, pela percepção distorcida da realidade e comportamento desorganizado e, por outro lado, pelo declínio cognitivo e emocional significativo (OWEN; SAWA; MORTENSEN, 2016). Um dos sintomas básicos da esquizofrenia desde a introdução do termo tem sido o transtorno do pensamento formal (*formal thought disorder*, FTD) (PERALTA; CUESTA, 2011). O FTD refere-se a aberrações no processo de pensamento, que geralmente se apresentam como distúrbios da fala e da linguagem. A classificação mais abrangente existente até agora divide o FTD em dois grupos: positivo, por exemplo, descarrilamento, tangencialidade, associações soltas; e negativo, por exemplo, alogia e bloqueio de pensamento (CAVELTI *et al*., 2018).

A fala incoerente ou desordenada é uma das principais características do FTD e um importante critério diagnóstico. Acredita-se que seja o reflexo das interrupções nos processos de pensamento normais (as que surgem no FTD, ver HART; LEWINE, 2017). Existem dois tipos principais de teorias que explicam as origens da incoerência discursiva

observada na esquizofrenia: teorias de disfunção executiva (também conhecidas como teorias de cognição prejudicada) e teorias de associação frouxa (ver DITMAN; KUPERBERG, 2010 para uma revisão). As primeiras teorias afirmam que a falta de controle sobre o processo do pensamento é típica do transtorno do pensamento negativo. As mais recentes, em contrapartida, explicam as incoerências em termos de tangencialidade e associações soltas, que são características do transtorno do pensamento positivo.

Dependendo do tipo de FTD e da gravidade do distúrbio, a fala pode ser afetada em vários níveis linguísticos (ver KUPERBERG, 2010 para revisão). As pessoas com esquizofrenia podem produzir, no seu discurso, menos palavras previsíveis (SALZINGER *et al*., 1970; Hart; PAYNE, 1973; SALZINGER; PORTNOY; FELDMAN, 1979) e mais neologismos, padrões de pausa não comuns (SPITZER et *al*., 1994) e uma quantidade mais elevada de erros gramaticais e lexicais do que falantes saudáveis (MARINI *et al*., 2008). O discurso de pessoas com esquizofrenia também é caracterizado por menor diversidade lexical. Estudar os desvios de linguagem e fala em pacientes com esquizofrenia com FTD é de alta importância prática, porque tais estudos podem fornecer ferramentas de diagnóstico mais objetivas e fáceis de usar (como análise automatizada da fala) e sustentar as classificações baseadas em evidências.

Uma das questões importantes em pesquisa de transtornos mentais é a definição da norma. De acordo com a 10ª revisão da Classificação Internacional de Doenças (CID-10) (Organização Mundial da Saúde (OMS), 1993) e da quinta edição do Manual Diagnóstico e Estatístico de Transtornos Mentais (DSM-5) (AMERICAN PSYCHIATRIC ASSOCIATION, 2013), as classificações clínicas contemporâneas de transtornos mentais e comportamentais mais utilizadas, uma condição só pode ser qualificada como transtorno mental se estiver associada ao sofrimento percebido ou à incapacidade funcional (ÜSTÜN; KENNEDY, 2009). Em vista dessa definição, os pesquisadores frequentemente usam os grupos de controle de sujeitos autorreferidos saudáveis como norma psiquiátrica. Mas há uma crescente consciência da utilidade limitada das classificações clínicas para fins de pesquisa, e novas classificações estão surgindo. Por exemplo, a *Hierarchical Taxonomy of Psychopathology* (HiTOP) (KOTOV; KRUEGER; WATSON, 2018) usa clusters de sintomas covariantes para a formação de diagnóstico, diferente das classificações clínicas tradicionais (CID-10 e

DSM-5), que usam uma abordagem categórica. A abordagem categórica afirma a presença ou ausência de uma condição patológica, enquanto a abordagem dimensional, como em HiTOP, observa os sintomas separados em um *continuum* com diferentes níveis de gravidade. A abordagem dimensional é muito mais requisitada na pesquisa psiquiátrica hoje em dia, mas ela não permite a separação de pessoas 'normais' e 'doentes' tão claramente quanto a abordagem categórica. Portanto, devemos ter em mente que os mesmos sintomas podem estar presentes em diferentes condições psiquiátricas e até mesmo na população não clínica em diferentes graus de gravidade. Por causa disso, em pesquisa, a avaliação cuidadosa de participantes saudáveis autorreferidos é tão importante quanto a avaliação de sujeitos com transtornos mentais diagnosticados. No *corpus* 3D, criamos uma subseção que descreve a norma psiquiátrica; ela contém as amostras de falantes que não apresentaram nenhum sintoma de transtornos mentais conforme avaliação por psiquiatra.

### 2.3 Variação linguística em adultos saudáveis

Na linguística clínica, a análise de discurso oral inclui a avaliação de características como comprimento da frase, diversidade lexical, taxa de elocução, conteúdo informativo, complexidade gramatical, parafasias, informatividade e coerência (PRINS; BASTIAANSE, 2004; BRYANT; FERGUSON; SPENCER, 2016). Para cada população clínica, é importante ter um controle para comparação. Dessa maneira, é crucial ter dados de controle equilibrados, que considerem toda a variabilidade possível. Abaixo, apresentamos alguns dos fatores que podem afetar as características da fala em falantes saudáveis.

A idade é um dos fatores de influência que pode afetar as características da fala em adultos saudáveis. Vários estudos afirmam que adultos saudáveis demonstram mudanças na linguagem falada com declínio no desempenho em diferentes domínios da linguagem (NADEAU, 2019). Ao contrário dos jovens, os idosos demonstram um aumento no número total de palavras (BORTFELD *et al*., 2001), menor conteúdo informativo (SALING; LAROO; SALING, 2012), mais estados de ponta da língua (por exemplo, BURKE; SHAFTO, 2004; GOLLAN; BROWN, 2006; ABRAMS; FARRELL, 2011), dificuldades em produzir e compreender frases sintaticamente complexas ou

ambíguas (por exemplo, KEMTES; KEMPER, 1997; KEMPER; HERMAN; LIAN, 2003; KEMPER; CROW; KEMETS, 2004) e vocabulários mais extensos (VERHAEGHEN, 2003). A tendência de idosos de produzir mais palavras coincide com um aumento no número de disfluências, tais como preenchimentos lexicais e não lexicais, repetições de palavras, pausas silenciosas longas e palavras vazias (KEMPER et al., 1990; HELLER; DOBBS, 1993; BORTFELD *et al*., 2001). Acredita-se que o aumento de disfluências relacionados à idade aconteça em adultos mais velhos por conta da dificuldade na recuperação de palavras (LOVELACE; TWOHIG, 1990; BORTFELD *et al*., 2001), as disfluências poderiam servir ao propósito de dar para eles mais tempo para encontrar a palavra pretendida.

## 3 Banco de dados *Discourse Diversity Database*

### 3.1 Os tipos de discurso e suas características de fala

Os tipos de discurso, ou gêneros, diferem em suas características de fala. Na linguística clínica, costumam ser analisadas as amostras de fala bastante curtas; o tipo de discurso depende da tarefa de elicitação. Os métodos mais utilizados na elicitação de discurso envolvem as tarefas que propõem descrever uma imagem ou uma sequência de imagens (WILLIAMS et al., 2010; BRYANT; FERGUSON; SPENCER, 2016), produzir um discurso narrativo que implica em contar uma história pessoal ou uma história baseada em uma trama bem conhecida (BEHRNS et al., 2009; OLNESS; ULATOWSKA, 2011), discurso procedural (ULATOWSKA; NORTH; MACALUSO-HAYNES, 1981; STARK, 2019) ou participar de uma conversa (WEBSTER; MORRIS, 2019).

Diferentes tarefas de elicitação de fala são associadas aos diferentes processos cognitivos (OLNESS, 2006; FERGADIOTIS; WRIGHT, 2011; GORNO-TEMPINI et al., 2011; STARK, 2019). Por exemplo, há uma diferença na linguagem do discurso narrativo e das tarefas de descrição de imagens: a primeira costuma ser mais complexa ao contar uma história pessoal ou recontar uma história conhecida ou evento acontecido com as próprias palavras (FERGADIOTIS; WRIGHT, 2011; MACWHINNEY *et al*., 2011). As tarefas de produção de narrativas sem estímulos visuais estimulam uma variedade mais elevada na fala com maior diversidade lexical (FERGADIOTIS;

WRIGHT, 2011; STARK, 2019) do que tarefas de descrição de imagem, nas quais muitas palavras descritivas são observadas (OLNESS et al., 2002). As tarefas de discurso procedural, por outro lado, pressupõem um esquema rigoroso que leva ao uso mais frequente de palavras de ação (PRITCHARD *et al*., 2015).

Presumimos que, para a avaliação adequada da capacidade linguística individual em todos os aspectos da fala real, diferentes tarefas de elicitação devem ser usadas. Escolhemos três tarefas de elicitação de fala, com e sem estímulos visuais, de dois gêneros diferentes - as tarefas de produção de narrativas e de discurso procedural.

### 3.2 Tarefas de elicitação de fala usadas no *corpus* 3D

Para o *corpus* 3D, obtivemos as amostras de discurso para análise através de três tipos de tarefas de elicitação de fala: narrativas induzidas por imagens, histórias pessoais e instruções baseadas em imagens (discurso procedural). Cada tarefa continha três opções de estímulos. Em tarefas de produção de narrativas obtidas através da descrição de imagens organizadas em sequências, usamos um dos três quadrinhos de Herluf Bidstrup ("Superman", "Discovery of the World", "Wonderful Day"). Para a elicitação das histórias pessoais, usamos uma das três perguntas sobre as ocasiões notáveis na vida do participante: (1) Por favor, conte-me sobre o melhor e o mais marcante presente que você já ganhou; (2) Por favor, conte-me sobre a melhor ou a mais marcante viagem que você já fez; (3) Por favor, conte-me sobre a melhor ou a mais marcante festa que você já foi. Como estímulos para as instruções baseadas em imagens, usamos os manuais de montagem de móveis da loja IKEA: uma cadeira, uma mesa e um banco. Em cada subseção, os diferentes procedimentos foram utilizados para garantir o equilíbrio e randomização (ver Tabela 1 para detalhes). A ordem de tarefas foi fixa.

## 4 *Subcorpora*

### 4.1 Visão geral

O *corpus* 3D integra duas coleções: de adultos com diagnósticos neurológicos e psiquiátricos e de adultos neurotípicos. O banco de dados consiste em amostras de fala de

dois grupos clínicos: (1) pacientes com tumores cerebrais (N = 45); (2) pessoas com transtornos do espectro da esquizofrenia (N = 26); e três grupos controle: (3) adultos saudáveis autodeclarados, intervalo de idade 18-80 anos (N = 84); (4) adultos jovens sem transtornos mentais diagnosticados por psiquiatra (N = 22); (5) adultos jovens autorreferidos neurologicamente saudáveis foram gravados em dois momentos de tempo: no estado ativo e no estado de cansaço (N = 10). A coleta de dados para o grupo (3) do *corpus* está concluída e, para os grupos (1), (2), (4) e (5), está em andamento. O resumo para todos os grupos é apresentado na Tabela 1[1].

Tabela 1. *Subcorpora* do *corpus* 3D

| | *Subcorpora* | | | | |
|---|---|---|---|---|---|
| | Clínicos | | Controle | | |
| | Neurocirúrgico | Espectro da esquizofrenia | Indivíduos autodeclarados saudáveis, grupo equilibrado por idade | Norma psiquiátrica | Indivíduos controle no estado ativo e no estado de cansaço |
| N de participantes | 74 | 26 | 84 | 22 | 10 |
| N de sessões | 3 | 1 | 1 | 1 | 2 |
| Distribuição de estímulos | Equilibrados em listas experimentais, ordem aleatória | Duas tarefas de descrição de imagens (*Sportsman* e *Adventure*) | Equilibrados em listas experimentais, ordem aleatória | Duas tarefas de descrição de imagens (*Sportsman* e *Adventure*) | Equilibrados em listas experimentais, ordem pseudoaleatória |
| Idade, anos | M = 49,7, SD = 14,6 | M = 28,8, SD = 4,3 | 18-29 anos (M = 21,2, SD = 2,6); 30-49 anos (M = 38,1, SD = 6,6); 50-64 anos (M = 57, SD = 3,8); 65+ anos (M = 72, SD = 7,0) | M = 23,9, SD = 4,3 | M = 28,80, SD = 2,86 |
| Diagnósticos | Tumores cerebrais | Esquizofrenia, transtorno esquizoafetivo | Não há | Não há | Não há |
| Metadados | MRI (ressonância magnética), RAT (*Russian Aphasia Test*) CETI (*Communicative Effectiveness Index*) | CID-10 (Classificação Internacional de Doenças), PANSS (*Positive and Negative Symptoms Scale*) | - | SCL-90-R (*Symptom Checklist-90-Revised*), PANSS (*Positive and Negative Symptoms Scale*) | Teste de autoavaliação diferencial de um dos estados funcionais |

[1] As informações detalhadas sobre cada participante, distribuição de estímulos entre as listas e dados adicionais em https://osf.io/2wvdz/?view_only=a38147409b5042bbabf1fc7560b32805. No presente momento, o *corpus* 3D não está disponível publicamente.

## 4.2 *Subcorpus* neurocirúrgico

### 4.2.1 Participantes

As amostras de discurso de pacientes submetidos a cirurgia para a ressecção de tumor foram coletadas na Universidade de Pesquisa em Medicina Privolzhsky (PIMU), em Nizhny Novgorod, e no Centro Médico e Cirúrgico Nacional de N. I. Pirogov, em Moscou. Os critérios de seleção para o grupo clínico exigiam que os indivíduos tivessem um tumor no hemisfério esquerdo em áreas críticas para a produção da linguagem, conforme avaliado com ressonância magnética.

Até o momento, aplicamos testes em 74 falantes nativos de russo (31 mulheres; a média de idade = 49,7 anos, SD = 14,6 anos; intervalo de idade - 19-72 anos; a média de anos de escolaridade - 14 anos, SD = 2,7 anos). Três pacientes desistiram de realizar as tarefas 1-2 dias antes da cirurgia devido ao déficit grave de fala ou à recusa de participação. Em seis pacientes, dentro de 3-7 dias após a cirurgia, as tarefas e o teste *Russian Aphasia Test* (RAT) (IVANOVA *et al*., 2019) não foram aplicados devido aos problemas de organização de testes, razões médicas ou devido à recusa do paciente. Cinco pacientes não realizaram todas as três tarefas no período de 3-7 dias após a cirurgia devido à recusa do paciente, déficit grave de fala ou razões médicas. 28 pacientes não completaram o teste (tarefas de fala e/ou RAT) três meses após a cirurgia por razões médicas ou pela impossibilidade de realizar o teste. Sete pacientes não terminaram o experimento por motivo de falecimento. Outros oito pacientes serão testados três meses após a cirurgia até o final do ano. Um paciente foi excluído da análise posterior devido a erro cometido durante a coleta de dados. Planejamos coletar as amostras de fala e RAT em mais de 60 pacientes em três períodos de tempo. Todos os participantes assinaram o termo de consentimento antes do experimento.

### 4.2.2 Protocolos de coleta de dados

De acordo com o mecanismo específico de reorganização cerebral, as amostras de fala encadeada foram coletadas em três períodos de tempo: 1-2 dias antes da cirurgia, 3-7 dias após a cirurgia e 3 meses após a cirurgia. Todos os participantes completaram três

tarefas em cada período de tempo e todos os participantes completaram diferentes versões de uma tarefa em cada um dos períodos de tempo. A distribuição das versões foi equilibrada entre as tarefas de coleta de dados (para as informações detalhadas, consulte OSF[2]). Os estímulos foram apresentados no *tablet* com a tela de 10,4 polegadas. Todos os pacientes foram testados em um local silencioso.

Em cada período da coleta de dados, os pacientes completaram o teste *Russian Aphasia Test* (RAT) (IVANOVA *et al*., 2019), um teste completo padronizado para a avaliação da produção e compreensão linguística em diferentes níveis de linguagem e o teste *Token Test* (de RENZI; VIGNOLO, 1962), uma ferramenta padronizada para a avaliação rápida da compreensão da linguagem em afasia. Antes da cirurgia e três meses após a cirurgia, os familiares dos pacientes preencheram o questionário do índice de eficácia comunicativa (*Communicative Effectiveness Index*, CETI) (LOMAS *et al*., 1989).

## 4. 3 Subcorpus de transtornos do espectro da esquizofrenia

### 4.3.1 Participantes

26 pacientes do Centro de Pesquisa em Saúde Mental em Moscou, Rússia, participaram do experimento (21 mulheres; a média de idade = 28,8 anos, SD = 4,3 anos; intervalo de idade de 18-42 anos; a média de anos de escolaridade = 13,6 anos, SD = 2,3 anos). Todos os pacientes foram admitidos na clínica com psicose aguda do espectro da esquizofrenia (diagnósticos CID-10: esquizofrenia F20 ou transtorno esquizoafetivo F25) e não tinham histórico de distúrbios neurológicos ou abuso de substâncias. Todos os participantes assinaram o termo de consentimento. A pesquisa foi aprovada pelo Comitê de Ética do Centro de Pesquisa em Saúde Mental.

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

### 4.3.2 Protocolos de coleta de dados

As amostras de discurso foram coletadas em períodos quando os pacientes estavam em remissão parcial, avaliados com as notas 3 (minimamente melhorado) ou 2 (muito melhorado) na escala *Clinical Global Impression Scale* (HARO *et al*., 2003). Cada participante completou duas versões da tarefa de narrativas induzidas por imagens e uma versão de cada uma das outras duas tarefas. As versões foram distribuídas em listas experimentais (para as informações detalhadas, consulte OSF[3]). Os estímulos foram apresentados na folha de papel em ordem fixa.

Além do diagnóstico clínico padrão, todos os pacientes foram avaliados com a escala padrão de sintomas psiquiátricos positivos e negativos (*Positive and Negative Symptoms Scale*, PANSS) (KAY; FISZBEIN; OPLER, 1987), que estima sintomas de esquizofrenia positivos, negativos e gerais, bem como a gravidade geral.

## 4.4 *Subcorpus* controle equilibrado por idade formado por sujeitos autodeclarados saudáveis

### 4.4.1 Participantes

As amostras de discurso do grupo controle foram coletadas em 84 falantes de russo como L1 sem distúrbios neurológicos ou psiquiátricos. Os participantes se distribuíram em quatro faixas etárias: 18-29 anos (N = 21; 16 mulheres; a média de idade = 21,2 anos, SD = 2,6 anos; a média de anos de escolaridade = 14 anos, SD = 2,0 anos), 30-49 anos (N = 23; 15 mulheres; a média de idade = 38,1 anos, SD = 6,6; a média de anos de escolaridade = 16,5 anos, SD = 2,9 anos), 50-64 anos (N = 20; 16 mulheres; a média de idade = 57 anos, SD = 3,8 anos; a média de anos de escolaridade = 16,4 anos, SD = 2,1 anos) e 65+ anos (N = 20; 15 mulheres; a média de idade = 72 anos, SD = 7,0 anos; a média de anos de escolaridade = 16 anos, SD = 3,1 anos). Todos os participantes assinaram o termo de consentimento antes do experimento. A pesquisa foi aprovada pelo Comitê de Pesquisas Interuniversitárias e Avaliação Ética em Pesquisas Empíricas

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

(*Committee on Interuniversity Surveys and Ethical Assessment of Empirical Research*) da Escola Superior de Economia (*High Economics School,* HSE), Rússia.

### 4.4.2 Protocolos de coleta de dados

As amostras de discurso padrão para as faixas etárias 18-29, 30-49 e 50-64 anos foram coletadas *online* na plataforma *Finding Five* (FINDINGFIVE, 2019). As amostras de discurso padrão na faixa etária 65+ anos foram coletadas com os estímulos apresentados na folha de papel ou na tela de um laptop ou tablet de 10,4 polegadas em um local silencioso. Cada participante completou uma versão de cada tarefa. A distribuição das versões de tarefas foi equilibrada (para as informações detalhadas, consulte OSF[4]).

## 4.5 *Subcorpus* controle formado por sujeitos avaliados como saudáveis por psiquiatra

### 4.5.1 Participantes

48 sujeitos sem histórico de distúrbios psiquiátricos ou neurológicos e sem histórico de abuso de álcool ou de substâncias participaram do estudo. Após preencher os questionários e passar por um exame psiquiátrico, 22 pessoas participaram do estudo (19 mulheres, a média de idade = 23,9 anos, SD = 4,3 anos; intervalo de idade - 20-36 anos; a média de anos de escolaridade = 15,7 anos, SD = 1,7 anos). Todos os participantes assinaram o termo de consentimento. A pesquisa foi aprovada pelo Comitê de Ética do Centro de Pesquisa em Saúde Mental, Rússia.

### 4.5.2 Protocolos de coleta de dados

Todos os participantes completaram a versão *online* do *Symptom Checklist-90-Revised* (SCL-90-R) (DEROGATIS; SAVITZ, 1999), uma lista dos sintomas

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

psiquiátricos mais proeminentes, comumente usados para os fins de triagem. SCL-90-R avalia nove dimensões de sintomas, tais como somatização, sintomas obsessivos-compulsivos, sensibilidade interpessoal, depressão, ansiedade, hostilidade, ansiedade fóbica, ideação paranoica e psicoticismo. Para selecionar os participantes com menor chance de doença psiquiátrica não diagnosticada, definimos o limiar do índice geral de sintomas (*General Symptom Index*, GSI) = 0,55, com base no limiar previamente estabelecido para os estudantes russos de 17 a 20 anos sem nenhum distúrbio psiquiátrico diagnosticado (KIOSEVA, 2016). Após isso, 27 participantes foram convidados para uma entrevista psiquiátrica cujo objetivo era eliminar os participantes com os traços esquizotípicos; 22 sujeitos foram qualificados como grupo controle da norma psiquiátrica para a pesquisa atual.

Todas as amostras de fala foram coletadas *online* via *Skype*. Além disso, todos os participantes foram avaliados com a escala psiquiátrica PANSS para a posterior comparação com os pacientes. Cada participante completou duas versões da tarefa de indução de narrativa baseada em imagens e uma versão de cada uma das outras duas tarefas. As versões foram distribuídas em listas experimentais (para as informações detalhadas, consulte OSF[5]). Os estímulos foram apresentados na folha de papel em ordem fixa.

### 4.6 *Subcorpus* controle formado por sujeitos saudáveis no estado ativo e no estado de cansaço

#### 4.6.1 Participantes

Dez sujeitos sem histórico de transtornos psiquiátricos ou neurológicos, abuso de álcool ou substâncias participaram deste estudo (8 mulheres, a média de idade = 28,8 anos, SD = 2,86 anos; intervalo de idade = 23-33 anos; a média de anos de escolaridade = 16,9 anos; SD = 1,91 anos). Todos os participantes assinaram o termo de consentimento.

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

### 4.6.2 Protocolos de coleta de dados

As amostras de fala foram coletadas *online*. Cada sujeito participou de duas sessões de coleta de dados: uma no estado ativo e outra no estado de cansaço. Em cada sessão, os participantes completaram o teste de autoavaliação diferencial do estado funcional (BOSKIN et al., 1973). As versões das tarefas foram distribuídas em listas experimentais (para as informações detalhadas, consulte OSF[6]).

## 5. Esquema de anotação

A anotação das amostras foi realizada com ELAN (WITTENBURG *et al*., 2006) em várias camadas.

### 5.1 Transcrição e segmentação

A camada de transcrição está alinhada com os arquivos de mídia e contém a transcrição ortográfica da fala gravada. A maioria das palavras nesta camada aparece em sua ortografia regular, no entanto, nos casos de erros fonéticos ou pronúncia específica, a transcrição reflete essas divergências da norma linguística. No nível de transcrição, todas as pausas com mais de 70 ms são anotadas, tanto silenciosas quanto preenchidas (por exemplo: *ah*, *um*).

A unidade principal para a segmentação da fala é a unidade de fala elementar (*elementary discourse unit*, EDU), a extensão dela é aproximadamente igual à extensão de uma oração. A unidade EDU contém um predicado, ou um predicado omitido que pode ser semanticamente recuperado; no caso de repetição do predicado resultante de dificuldades de encontrar palavras, todos os lexemas repetidos são incluídos no mesmo EDU (para os exemplos, veja BERGELSON; KHUDYAKOVA, 2020). Os enunciados incluem a oração principal com todas as suas orações subordinadas; a proporção de orações e enunciados pode ser interpretada como uma medida de complexidade sintática (MARINI, 2012).

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

## 5.2 Anotação microlinguística

A transcrição lexical é uma camada técnica e contém as mesmas informações que a camada de transcrição segmentada em palavras e não-palavras.

As camadas de lemas e de classes gramaticais (*part-of-speech*, POS) contêm formas básicas e rótulos de classes gramaticais. O esquema de marcação das classes gramaticais e a lematização são baseados no manual do *corpus* nacional de russo, *Russian National Corpus* (http://www.ruscorpora.ru/en/corpora-morph.html).

Erros gramaticais, semânticos e fonéticos são anotados na camada de erros.

A camada de não-palavras contém anotações de pausas silenciosas, pausas preenchidas, falsos começos, repetições, palavras semanticamente vazias e expressões automáticas. As pausas silenciosas são segmentos de fala silenciosos com a duração maior de 70 ms, as pausas preenchidas são segmentos com a duração maior de 70 ms preenchidos com um som que não seja uma palavra (por exemplo, *ah* ou *um*) que também não apresenta o início de uma nova palavra. Os falsos começos são segmentos de palavras não finalizados (geralmente uma sílaba-longa), veja (1), onde o falso começo é marcado por =. As repetições incluem palavras e frases repetidas sem alteração (cf. MACWHINNEY, 2010, p.77). Na camada de não-palavras, as expressões são marcadas como repetições a partir da segunda vez a serem usadas, ver (1), onde as repetições são marcadas em itálico.

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

## 5.3 Anotação de componentes no nível macro

Cada elemento EDU é anotado com cinco componentes do nível macro. Os elementos EDU principais descrevem os eventos da história e as ações das instruções. Os elementos EDU do segundo plano contêm as informações gerais sobre os personagens da história, o cenário onde a história se passa ou, no caso das tarefas de instrução, a descrição dos elementos e espaço. Os elementos EDU de comentários contêm as opiniões e ideias do falante em relação aos eventos, personagens e detalhes da história. Ao contrário dos comentários, os metacomentários contêm a informação sobre as atitudes e os pensamentos do falante em relação ao processo de contar a história ou dar a instrução,

tais como os problemas de busca de palavras, por exemplo. Os elementos EDU reguladores organizam o fluxo da fala e não têm qualquer conteúdo informativo.

## Considerações finais

A estrutura e anotação do corpus 3D permitem o estudo de várias características da fala em vários níveis linguísticos, diversos gêneros discursivos e diferentes grupos de sujeitos.

As anotações manuais permitem extrair as medidas comumente usadas na avaliação da fala: as medidas de fluência (relação de pausa, taxas de elocução e articulação), comprimento médio de elementos EDU e enunciados medidos em palavras e milissegundos, diversidade lexical e o número de erros, falsos começos e repetições. Além disso, a anotação no nível macro permite extrair os dados apenas dos elementos EDU relacionados ao enredo da história e excluir da análise os comentários e reguladores, uma vez que os elementos EDU complementares podem afetar algumas das medidas, como a diversidade lexical (KINTZ; FERGADIOTIS; WRIGHT, 2016).

O *subcorpus* de neurocirurgia inclui as amostras de fala coletadas em três períodos diferentes, assim como os dados de neuroimagem, o que permite criar os perfis completos de linguagem dos pacientes nos estados pré e pós-operatório e investigar a maneira como o crescimento do tumor cerebral e as lesões que ele provoca nos tratos de substância branca afetam a linguagem. Além disso, devido à disponibilidade de dados obtidos por meio de testes padronizados de avaliação da linguagem no *subcorpus* de neurocirurgia, podemos analisar como os déficits em vários níveis de linguagem (fonético, lexical e sintático) se manifestam no discurso oral de vários gêneros.

A análise do discurso do *corpus* 3D não se limita à extração de dados da disponível anotação manual. Por exemplo, estamos atualmente executando uma análise automatizada da coerência global e local em sujeitos com esquizofrenia. Essa análise é baseada nas transcrições dos *subcorpora* da esquizofrenia e da norma psiquiátrica com a aplicação de modelos *word2vec* (ver a descrição e avaliação do método em RYAZANSKAYA; KHUDYAKOVA, 2020).

Criar um amplo *corpus* de amostras de fala obtidas em populações clínicas e grupos de controle saudáveis é um processo demorado e trabalhoso. No entanto, esses

recursos fornecem muitas possibilidades para a pesquisa linguística minuciosa, assim como para a análise automatizada, comparação de variados grupos de falantes e para o estudo do substrato neural da fala.

## REFERÊNCIAS


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.

**Financiamento**

A coleta de dados para o *subcorpus* neurocirúrgico e o *subcorpus* da norma autodeclarada equilibrado por idade foi apoiada pela fundação *Russian Science Foundation*, projeto No. 20-18-00-399. A coleta de dados para os *subcorpora* da esquizofrenia e da norma psiquiátrica foi apoiada pelo Centro de Linguagem e Cérebro da Escola Superior de Economia (*Center for Language and Brain NRU Higher School of Economics*), Bolsa do Governo da Federação Russa, Ag. No. 14.641.31.0004.

**Declaração da contribuição de autores**

Mariya Khudyakova concebeu e planejou a coleta de dados para o *corpus*, elaborou os estímulos e o esquema de anotação. Mariya Khudyakova, Natalia Gronskaya e Olga Dragoy planejaram e supervisionaram a coleta de dados para o *subcorpus* de norma autodeclarada e para o *subcorpus* neurocirúrgico. Konstantin Yashin, Igor Medyanik, Andrey Zuev, Alina Minnigulova, Natalia Antonova, Maria Nelubina, Anastasia Surova e Anna Vorobyova coletaram os dados para o *subcorpus* neurocirúrgico. Alina Minnigulova, Natalia Antonova e Anastasia Surova analisaram os dados de neuroimagem para o *subcorpus* neurocirúrgico. Natalia Antonova, Maria Nelubina, Anastasia Surova e Anna Vorobyova coletaram os dados para o *subcorpus* da norma autorreferida e anotaram os dados para *o subcorpus* da norma autorreferida e para o *subcorpus* neurocirúrgico.

Tatiana Shishkovskaya e Galina Ryazanskaya coletaram e anotaram os dados para os *subcorpora* da norma psiquiátrica e da esquizofrenia. Mariya Khudyakova coletou e anotou os dados para o *subcorpus* da norma nos estados ativo e de cansaço.

Mariya Khudyakova, Natalia Antonova e Tatiana Shishkovskaya escreveram o manuscrito. Todos os autores contribuíram com o parecer crítico e ajudaram na construção do banco de dados e na elaboração do manuscrito.

Traduzido por *Aleksandra S. Skorobogatova*

*Recebido em 07/10/2021*

*Aprovado em 22/08/2022*

**Pareceres**
Tendo em vista o compromisso assumido pela *Bakhtinina*. Revista de Estudos do Discurso com a Ciência Aberta, a revista publica somente os pareceres autorizados por todas as partes envolvidas.

**Disponibilidade de dados de pesquisa e outros materiais**
Os dados não podem ser disponibilizados publicamente. Os dados coletados incluem amostras de fala de pessoas com diversas deficiências. Tais gravações não devem ser disponibilizadas ao público. Colocamos as informações disponíveis na plataforma OSF e disponibilizamos o *link* no artigo.