JLAR 3 (2025)
10.15460/jlar.2025.3.2.1838

# The Brazilian Portuguese-Russian Corpus (BraPoRus) of older heritage speakers

Angelina Rubina[1], Olesya Kisselev[1], Aleksandra S. Skorobogatova[2], Anna Smirnova Henriques[3], & Irina A. Sekerina[4]*

[1]University of South Carolina, USA, [2]Universidade de São Paulo, Brazil, [3]Pontifícia Universidade Católica de São Paulo, Brazil, [4]The City University of New York, USA

**Abstract.** The *Brazilian Portuguese-Russian Corpus* (BraPoRus) is the first corpus of semi-spontaneous speech by cognitively healthy older heritage speakers of Russian (ages 60–100) who are primarily dominant in Brazilian Portuguese while retaining functional proficiency in heritage language (HL) Russian. BraPoRus includes over 200 hours of naturalistic speech from 43 participants and includes a subcorpus, BraPoRus-1.0, already pre-processed and available for research via BilingBank of TalkBank. BraPoRus-1.0 comprises 15-minute extracts from different interview sessions that contained descriptions of various life events from 16 participants (women $n = 10$, men $n = 6$; $M_{age} = 80.5$, age range 65–98). All recordings were transcribed into Cyrillic automatically by Sonix.ai and cross-checked manually; they were also processed with a BatchAlign-2 pipeline for automatic processing of media files. In addition to linguistic documentation, all BraPoRus participants underwent cognitive testing, enabling multimodal analyses. BraPoRus thus provides a critical resource for investigating HL maintenance, attrition, and linguistic aging. Its integration with psycholinguistic experiments enables cross-methodological studies. Its open-access design supports replication, comparative analysis, and tool development for HL pedagogy. As such, BraPoRus contributes a much-needed lifespan perspective to corpus-based bilingualism research.

**Keywords.** speech corpora, bilingualism, heritage language, Russian, aging

A peer-reviewed contribution to the *Journal of Language and Aging Research* (JLAR).
Submitted: 2025-06-13
Accepted: 2025-11-14
Published: 2025-12-29
DOI: 10.15460/jlar.2025.3.2.1838



*. Corresponding author, Irina.sekerina@csi.cuny.edu



Please cite as Rubina, Angelina, Kisselev, Olesya, Skorobogatova, Aleksandra S., Smirnova Henriques, Anna, Sekerina, Irina A. 2025. "The Brazilian Portuguese-Russian Corpus (BraPoRus) of older heritage speakers." *Journal of Language and Aging Research* 3(2): 239–253. 10.15460/JLAR.2025.3.2.1838.

## 1 Introduction

Nearly half of the world's residents today speak more than one language, with a bilingual population continuously increasing in proportion to monolingual speakers (Lewis, Simons, and Fennig 2021). This trend is intersecting with another important demographic trend, that of the aging population. The number of people over 60 years of age already outnumbers children under five and is set to double within 25 years from 1 to 2.1 billion (World Health Organization 2024). Despite this demographic convergence, research on how bilingual competencies evolve across lifespan remains scarce, with most studies focusing on cognitive comparisons between young monolinguals and bilinguals or on neuroprotective "bilingual advantage" (Rothman 2024).

While corpus-linguistic research has grown considerably in the past three decades for young adults and children, resources dedicated to documenting speech of older monolingual adults remain extremely scarce, with such notable exceptions as the LangAge Corpus (Gerstenberg 2011) and the multimodal CorpAGEst (Bolly and Boutet 2018) for French, and the Corpus of Spoken Yiddish in Europe (Bleaman and Nove 2025).

A handful of other resources focus on examining speech associated with aging though pathologies, such as AphasiaBank (MacWhinney et al. 2011), Carolina Conversation Collection (CCC, Davis 2010), DementiaBank (Lanzi et al. 2023), and VIntAGE (Duboisdindien and Bolly 2024), with the latter two targeting communication impairments in older people with dementia.

Corpus-based studies of bilingual young speakers are more abundant; however, the bulk of current research focuses on second language (L2) bilingualism (e. g., the International Corpus of Learner English, Granger 2004) and the written genre. For heritage language (HL) bilinguals, a small number of spoken corpora have started to appear only recently, e. g., the Heritage Language Documentation Corpus (Nagy 2009) and the RUEG corpus (Wiese et al. 2021) featuring young heritage speakers (HSs). Despite these significant advances of corpus linguistics studies in bilingualism, the lack of speech corpora on healthy aging in bilinguals is surprising (Dovetto and Marra 2024). One noticeable exception is the Corpus of American Nordic Speech (CANS, Johannessen 2015) collected between 1935 and 2017 in the U.S. CANS v.3.1. is based on approximately 30 minutes of spontaneous speech of 242 older HSs (80 years on average) who grew up in the families of Norwegian and Swedish descent in the U.S. The HSs were the 3rd and later generations, and their spontaneous speech was short and laborious.

To the best of our knowledge, except CANS (Johannessen 2015), there is no other spoken corpus of older bilingual adults who still speak the HL. Given that HL bilingualism is the most common type of bilingualism today, there is an urgent need to extend corpus-based research to older HL speakers "...to attain the fullest possible understanding of the developmental trajectories of heritage grammars across the lifespan" (D'Alessandro, Natvig, and Putnam 2021, 2).

The Brazilian Portuguese-Russian Corpus of Older Speakers (BraPoRus, (Sekerina et al. 2023)) described in the present article provides an important resource for investigating the HL Russian language spoken in Brazil by second generation descendants of Russian immigrants. BraPoRus is a targeted collection of naturalistic speech of older bilingual speakers, aged 60–100, whose dominant language is Brazilian Portuguese (BP). The BraPoRus Corpus will add to the small but growing number of

JLAR 3 (2025)
10.15460/jlar.2025.3.2.1838

language corpora of older adults' monolingual speech (e. g., Bleaman and Nove 2025; Bolly and Boutet 2018; Gerstenberg 2011).

It also aims at establishing a methodological blueprint for future studies of interaction between aging and heritage bilingualism in order to represent the full lifespan continuum in studying HLs.

## 2 Brazilian Portuguese-Russian Corpus (BraPoRus)

The BraPoRus Corpus compilation began in 2021 as part of a larger project on the investigation of heritage HL bilingualism in aging. The project overall contains two major components:

1. the actual corpus of spontaneous naturalistic speech recorded from older bilingual BP-HL Russian speakers who currently live in Brazil
2. psycholinguistic experiments with the same participants.

Full transcription and linguistic analysis of BraPoRus data are under way, with a subset of data already pre-processed and available for research. This sub-set, BraPoRus-1.0, includes sixteen speech samples (15-minutes each) from 16 participants (Section 3) available in open access in TalkBank (Sekerina, Skorobogatova, and Smirnova Henriques 2025).

### 2.1 Participants

Bilingual BP-HL Russian speakers constitute a unique, geographically isolated HL community. More than 130 000 Russian-speaking migrants arrived in Brazil in the first half of the 20th century from the former Russian Empire, mostly via China or Europe. The number of Russian speakers that arrived in Brazil as children or were born there in the 1940–1950s and are still alive today is around 1 500 people (Skorobogatova et al. 2021). Only a part of them speak Russian, and previous studies on the history of the Russian immigration in Brazil collected interviews with the older immigrants in Portuguese (Ruseishvili 2016). For this project, we carefully searched for older speakers who had advanced functional proficiency in Russian, preserved in their families after leaving the Russian Empire. They were mostly highly educated, cognitively healthy people, with middle- and upper-class socioeconomic status.

At the beginning of the study in 2021, the target number of participants for BraPoRus was set at 50. From the time the technical description of BraPoRus was published (Sekerina et al. 2023) based on short extracts from 8 participants, the number of BraPoRus participants increased from 26 to 43 (5 participants, 3 men and 2 women, whose data were recorded since passed). All BraPoRus participants have provided demographic information and have been tested with three cognitive assessments. Based on the country of birth, they are classified into four categories: Brazil, China (mostly Harbin), Europe, and Russia. Table 1 provides the basic demographic characteristics for 43 (BraPoRus, all) and the 16 (BraPoRus-1.0) participants whose data are described in Section 3.

Table 1: BraPoRus and BraPoRus-1.0: Demographics

| **Category** | **BraPoRus (all, n=43)** | **BraPoRus-1.0 (n=16)** |
|---|---|---|
| **Gender** | 27 women, 16 men | 10 women, 6 men |
| **Age: Mean (range), years** | 79.5 (52–103) | 80.5 (65–98) |
| **Birth place** | | |
| Brazil | 17 | 4 |
| China | 9 | 8 |
| Europe | 12 | 3 |
| Russia | 5 | 1 |
| **Current residence** | | |
| São Paulo | 28 | 14 |
| Rio de Janeiro | 13 | 2 |
| Paranáą | 2 | 0 |

BraPoRus inclusion criteria were as follows:

1. Aged 60 years and older (range: 60–100; there were 2 exceptions).
2. Have lived in Brazil for most of their life, or were born in Brazil, and speak BP as a societal language.
3. Have proficiency in HL Russian sufficient to maintain a conversation for an hour.
4. No long-term residence in Russia.
5. No documented cognitive impairment.

## 2.2 Data collection and pre-processing

Over the past years (2021 through present), approximately 200 hours of spoken samples were recorded with 43 participants. All tasks were audio- or video recorded on a PC using Zoom™ or on a smartphone. The workflow of all stages for processing included cataloging the recordings and their processing stages in an Excel spreadsheet and maintaining the participants' characteristics in Lameta (V3.2-beta, Hatton and Hirt 2025), a metadata tool to help with organizing collections of files.

The media files and the data generated in subsequent processing are stored privately in the business Dropbox account, inaccessible to the general public (for the structure of the interview see Smirnova Henriques et al. 2022).

Some technical characteristics of BraPoRus have been described in Sekerina et al. (2023), in a paper that focused mainly on the quality of spoken data remotely collected during COVID-19 pandemic that comprises much of the corpus. Having analyzed a subset of data from eight participants, which were transcribed both manually and with the help of automated transcription software Sonix.ai, the researchers found automated transcription to be a viable alternative to manual (and labor-intensive) methods with the word error rate under 9.86%. The recordings were also found suitable for phonetic and acoustic analysis, including F0, F1, and F2 formants. The data used to distribute the corpus are in .mp3 format (MPEG Audio layer ½, stereo, sample rate 44100 Hz, bits per sample 32 kb/s).

Following in the steps of Sekerina et al. (2023), data pre-processing of the rest of the data in the BraPoRus (including the current subcorpus described in Section

3 below) proceeded in the same manner. In particular, data were anonymized by substituting participant names with alpha-numeric codes, e. g., AVG_m_70 (the 3 initials_gender_age, see Table 2). All the participants whose data are described in the present article chose the most permissive consent options allowing their recordings and data to be made publicly available to the researchers registered at TalkBank (see Ethical Considerations statement below).

## 3 BraPoRus-1.0: A subcorpus of 16 participants

Here we present BraPoRus-1.0, a small standalone subcorpus of BraPoRus in general, that is available in open access in the BilingBank of TalkBank (Sekerina, Skorobogatova, and Smirnova Henriques 2025). Sixteen participants were selected in a pseudorandomized manner, 8 of which were already in the technical description of quality of the BraPoRus speech samples (Sekerina et al. 2023). Table 2 represents their demographic information.

The age was the age at the time of recording.

Table 2: BraPo-Rus-1.0: Participants

| **Participant and session** | **Gender** | **Age** | **Born in** | **Topic** |
|---|---|---|---|---|
| EAB_S9 | F | 65 | Brazil | Childhood, traditions |
| IMK_S3 | M | 82 | Brazil | Childhood |
| AVM_S1 | M | 69 | Brazil | Family history, taking care of the old mother |
| SAP_S1 | F | 71 | Brazil | Family history |
| EKS_S4 | F | 73 | China | House |
| GAA_S1 | F | 85 | China | Life in China, leaving for Brazil |
| VVG_S1 | M | 80 | China | Family history |
| ENL_S1 | F | 72 | China | Ties to Russian |
| AVG_S2 | M | 70 | China | Russian boarding school in Brazil |
| NVM_S3 | M | 76 | China | Parents and guests, life in São Paulo, music band |
| ZVH_S1 | F | 78 | China | Family history, grandmother, religion |
| SAK_S2 | F | 73 | China | First house in Brazil, childhood |
| LNI_S5 | F | 70 | Europe | Moving to and first location in Brazil |
| TNK_S5 | F | 90 | Europe | Moving, trips, Russian culture in Brazil |
| TYL_S1 | F | 98 | Europe | Childhood and teen years in Europe |
| FFK_S1 | M | 83 | Europe | Childhood, leaving the village |

JLAR 3 (2025)
10.15460/jlar.2025.3.2.1838

The participants were selected from the three largest categories for the birth place (i. e., Brazil, China, and Europe), with more women (n=10) than men (n=6), and varying in age (Mage = 80.5, range 65–98) to reflect the composition and variability of the entire BraPoRus data. We chose 15-min. extracts from different sessions (i. e., S1–S9) that contained descriptions of various life events.

All recordings were transcribed automatically by Sonix.ai into Cyrillic and then cross-checked manually by trained research assistants who listened to the audio file while reading the transcript and corrected the occasional errors. Resulting recordings were originally in either audio .wav format or video .mp4 format.

Sonix transcription tool allows for either format as its input, and it converted the .wav and .mp4 media files to the.txt format for transcriptions as its output (Sonix, Inc.). Each transcription file was named using participant's code, with session number added (e.g., AVG_m_70_S2.txt, see Table 2). It was done to de-identify participants' personal information whose full details were recorded separately in Excel spreadsheets.

The transcriptions were simply utterances in Cyrillic, without any morphological or syntactic tagging. Some transcripts included single words in Brazilian Portuguese (proper names referring to places, people, etc.). These words were marked as non-Russian and were manually checked by a speaker of Brazilian Portuguese.

To avoid labor-intensive manual tagging, we used a BatchAlign-2 (BA2) pipeline for automatic processing of media files (Liu and MacWhinney 2024) that are meant for contribution to TalkBank (MacWhinney 2019). BA2 is a Python-based software that transcribes media files (using such AI-components as Whisper.ai, Radford et al. 2022) in a TalkBank .cha (MacWhinney 2000), force-aligns the resulting transcriptions with the media source files, morphologically and syntactically tags them with the tiers (%mor and %gram). Trained research assistants compared the Sonix transcriptions in .txt with the transcriptions of the same media files in .cha performed by BA2 and corrected all the discrepancies by listening to the media sources.

The BA2 then forced-aligned the .cha transcripts with the media files, morphologically (%mor) and syntactically (%gra) tagged them, and translated into English (%xtra). The indexing methods followed the procedures described in the CLAN Manual (MacWhinney 2000).

Figure 1 illustrates how the transcript is presented in the BilingBank of TalkBank, Figure 2 presents the sonogram, with the transcription of the related *.cha file.

Table (3) and Table (4) illustrate the first utterance of the resulting .cha file for Participant AVG_m_70_S2 from BraPoRus-1.0 (Table 2).

The .cha format is meant to be used with the CLAN (Computerized Language Analysis) program (MacWhinney 2000) that performs linguistic analysis of language samples transcribed with CHAT. We selected the .cha format because we chose to upload the subcorpus, both media and transcribed and tagged .cha files to BilingBank of TalkBank (Sekerina, Skorobogatova, and Smirnova Henriques 2025).

JLAR 3 (2025)
10.15460/jlar.2025.3.2.1838

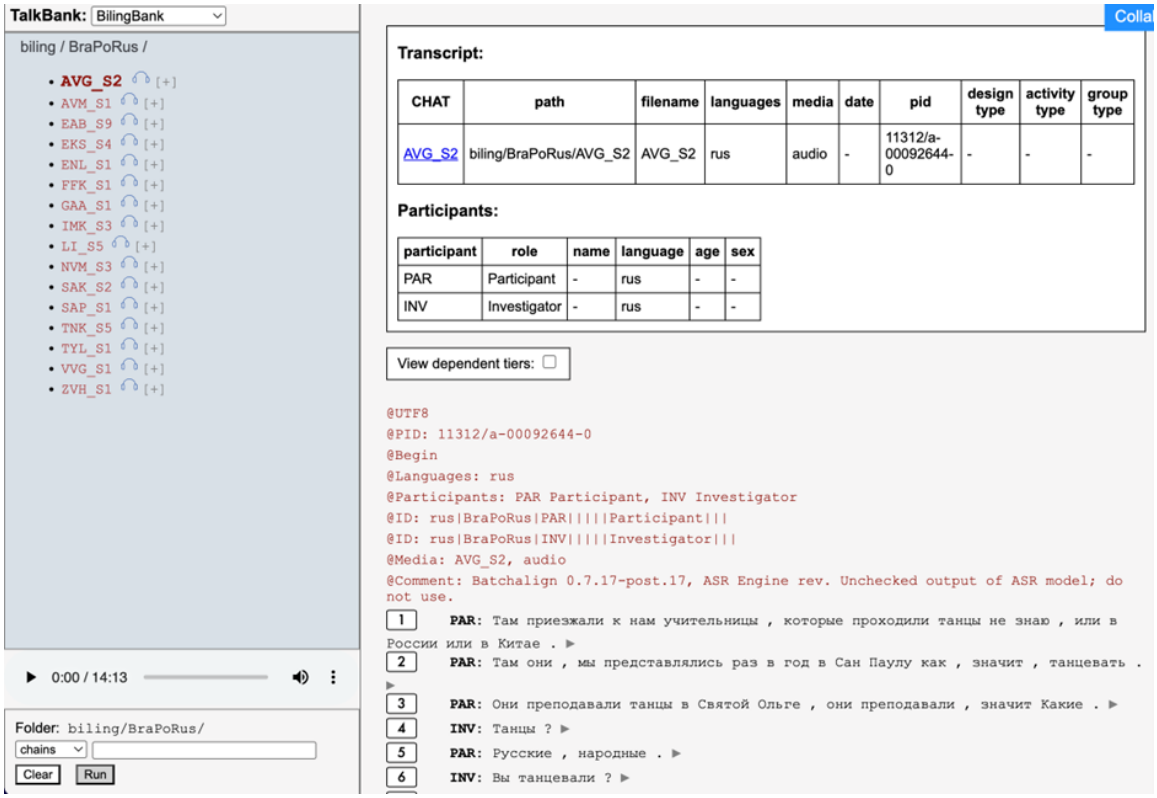


Figure 1: BraPoRus-1.0 transcript presentation in TalkBank

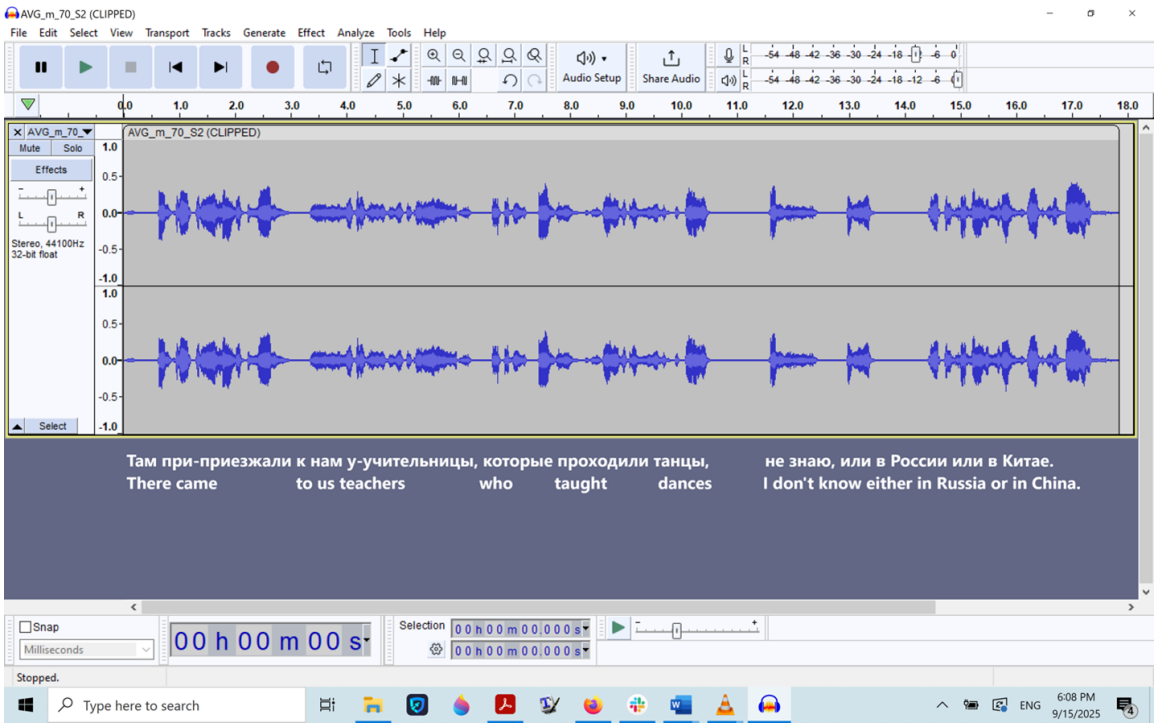


Figure 2: A sonogram of an utterance by participant AVG_m_70_S2 from BraPoRus-1.0

Table 3: .cha file header information

| Field | Value |
|---|---|
| @Begin | |
| @Languages | rus |
| @Participants | PAR Participant, INV Investigator |
| @ID: rus\|BraPoRus\|PAR | \|\|\|\|\|Participant\|\|\| |
| @ID: rus\|BraPoRus\|INV | \|\|\|\|\|Investigator\|\|\| |
| @Media | AVG_S2, audio |
| @Comment | Batchalign 0.7.17-post.17, ASR Engine rev. Unchecked output of ASR model. |

Table 4: Illustration of first utterance in .cha file of AVG_m_70_S2

| Tier | Content |
|---|---|
| *PAR: | Там приезжали к нам учительницы , которые проходили танцы не знаю , или в России или в Китае . |
| %mor: | adv\|там verb\|приезжать-Fin-Imp-Ind-Past-P adp\|к pron\|мы-Prs-Dat-P1 noun\|учительница-Fem-Plur-Nom cm\|cm pron\|который-IntRel-Nom-S1 verb\|проходить-Fin-Imp-Ind-Past-P noun\|танец-Masc-Plur-Acc part\|не verb\|знать-Fin-Imp-Ind-Pres-S1 cm\|cm cconj\|или adp\|в propn\|Россия-Fem-Loc cconj\|или adp\|в propn\|Китай-Masc-Loc . |
| %gra: | 1\|2\|ADVMOD 2\|18\|ROOT 3\|4\|CASE 4\|2\|OBL 5\|2\|NSUBJ 6\|8\|PUNCT 7\|8\|NSUBJ 8\|5\|ACL-RELCL 9\|11\|OBJ 10\|11\|ADVMOD 11\|8\|CONJ 12\|15\|PUNCT 13\|15\|CC 14\|15\|CASE 15\|11\|CONJ 16\|18\|CC 17\|18\|CASE 18\|15\|CONJ 19\|2\|PUNCT |
| %xtra: | There were teachers who came to us , who danced , I don’t know , either in Russia or China . |

## 4 Sample analysis: Lexical complexity

Our sample analysis of BraPoRus-1.0 data focuses on linguistic complexity. Linguistic complexity is a subject of much debate in bilingualism (see Laleko and Kisselev (2021) and Polinsky and Putnam (2024), specifically with regard to HLs). In the BraPoRus project, we refer to linguistic complexity in its formal sense as the range of forms found in the language of the speakers and the degree of sophistication of these forms (Housen et al. 2019). Of particular interest to us is the concept of lexical complexity, which we measure by inspecting the variation and sophistication of lexical items used by our participants (see Table 5). The analysis of the 16 texts reveals notable variation in lexical sophistication and diversity across the texts. Average word length ranged from 4.18 to 5.21 letters, with most texts clustering around 4.5 to 5.0, suggesting moderate lexical sophistication, yet highly consistent with what was found for standard Russian oral production (cf. Bogdanova-Beglarian, Martynenko, and Sherstinova 2015).

Syllabic and morphological complexity were also relatively consistent, with averages near 1.3 syllables and 1.4 to 1.5 morphemes per word, although LNI stood out with the highest morphological complexity (1.61), possibly reflecting the frequent use of derived or compound words. Lexical diversity, measured by the number of unique tokens and lemmas and MTLD-based type-token ratios, varied more widely across the participants. ZVH produced the most lexically rich text (1086 tokens, 850 lemmas), while TNK and LNI showed the lowest diversity. MTLD scores further highlighted

FFK, ZVH, AVG, and NVM as having particularly diverse lexicons, whereas TNK and LNI relied on more repetitive vocabulary. Taken together, these patterns suggest that participants like ZVH and AVG may demonstrate higher lexical proficiency or more advanced stylistic choices, while TNK and LNI reflect simpler or more redundant lexical patterns. Most other texts fall within a mid-range band, balancing moderate word complexity with a fairly rich vocabulary.

While specific TTR values for either Russian monolingual speech or Russian heritage speaker oral production are not readily available, future analysis of this parameter will have to include a comparative lens: while comparing the performance of our participants to other groups of Russian speakers (such as young HL speakers, first-generation immigrant bilinguals, and native speakers, both young and older adults), we will be able to co-relate these lexical parameters with other factors, cognitive and sociolinguistics, that contribute to the overall linguistic performance of our participants.

## 5 Reuse potential

The subcorpus BraPoRus-1.0 presented in this article and its parent BraPoRus corpus (whose processing still continues) offer exceptional re-use potential across a range of scholarly domains. Documenting spontaneous speech from cognitively healthy older heritage speakers, it captures a disappearing variety of HL Russian spoken by BP-dominant bilinguals. Following D'Alessandro, Natvig, and Putnam (2021), this variety may be classified as moribund, spoken by the final generation of highly proficient users, who did not transmit it to their children. The corpus' naturalistic speech samples illuminate more than phonetic or grammatical patterns – they reveal how HL Russian in Brazil evolved in isolation from the baseline for over 60 years, offering insights into narrative competence, cultural continuity, and community memory in a context of linguistic attrition.

Moreover, BraPoRus is part of a broader research framework that triangulates corpus-based methods with experimental psycholinguistic and cognitive assessment protocols, all conducted with the same participants. This integrative, multidisciplinary approach increases the corpus's value for cross-methodological studies that explore HL representation, processing, and maintenance across lexical, grammatical, and discourse domains. The fact that the BraPoRus-1.0 subcorpus is already in open access at TalkBank enables not only replication and comparison, but also development of tools for speech processing and HL pedagogy.

The data and metadata in BraPoRus-1.0 meet community needs of HL speakers of languages other than English. Although English has long held the status of a global language, other languages continue to maintain and, in some cases, increase their significance due to political and global security considerations. Russian, spoken by 255 million people worldwide (Lu 2024), is among the 10 most spoken languages. Approximately 30 million Russian speakers reside outside of Russia, spanning post-Soviet states, Eastern and Western Europe, and the Americas, a number that has grown since Russia's large-scale military invasion of Ukraine (Dubinina and Kisselev 2025). Recognizing its strategic importance, the U.S. government designates Russian as a language critical for national security and economic prosperity; however, it remains underrepresented among American speakers (Rivers and Brecht 2018).

While Americans have a rather poor record learning second languages, the country boasts a largely untapped source of linguistic diversity in the form of various HLs (Rivers and Brecht 2018). The sheer number of HLs (40+) and their speakers (28+ million; Nagano (2015)) in the United States makes investigating and supporting research on HLs from early childhood to older age an economically viable option to advance our nation's competitiveness in the modern globalized society through linguistic diversity. Current research on HL bilingualism is piecemeal and focused almost exclusively on children and college-aged adults. Our project is the first comprehensive study that puts HL bilingualism in later life in the center and employs a combination of methodological approaches. Therefore, the BraPoRus project will go beyond simply horizontally extending the field of HL bilingualism. Instead, it has the potential to advance the field vertically, by opening it up to the interaction with aging and its impact on language.

## 6 Author contributions

Angelina Rubina: Data curation, Writing – original draft, Writing – review and editing. Olesya Kisselev: Formal analysis, Writing – original draft, Writing – review and editing. Aleksandra Skorobogatova: Data curation, Investigation, Methodology, Investigation, Writing – review & editing. Anna Smironova Henriques: Conceptualization, Methodology, Writing – review & editing. Irina Sekerina: Conceptualization, Data curation, Investigation, Methodology, Resources, Supervision, Funding acquisition, Writing – original draft, Writing – review and editing.

## 7 Acknowledgements

We would like to thank all BraPoRus participants for their enthusiastic support of our project. Our special thanks go to Brian MacWhinney for his generous help with BA2 processing and Anton Vakhranev for checking the automatic results.

## 8 Data availability

BraPoRus-1.0 includes sixteen speech samples (15 minutes each) from 16 participants available in open access in the BilingBank of TalkBank (Sekerina, Skorobogatova, and Smirnova Henriques 2025).

## Funding

The last author (IAS) was partially supported by the PSC-CUNY grant (Trad-B 66406-00 54) and a Fulbright Core Program award *Bilingualism in Brazil: The Case of Russian-Brazilian Portuguese*. During the data collection, the third author (ASS) was supported by FAPESP (2022/01119-0), and the fourth author (ASH), by PNPD/CAPES fellowship (88882.315378/2019-01).

## Ethics statement

Data collection was approved by the Ethics Committee of Pontifícia Universidade Católica de São Paulo (CAAE 09079219.9.0000.5482). The consent form, written in Brazilian Portuguese, the dominant language of the bilingual participants, included three options for authorization for the recordings and their analysis (publicly available without anonymization, publicly available with anonymization, or not publicly available. All participants emphasized that their goal in taking part in the BraPoRus project was a desire to preserve their memories for posterity. Accordingly, all the participants whose data are described in the present article chose the most permissive consent options, allowing their recordings and data to be made publicly available to registered researchers with names disclosed.

## Conflict of interest

The authors confirm that there is no known conflict of interest associated with this publication and that no significant financial support could have influenced its outcome.

## References

Bleaman, Isaac L., and Chaya R. Nove. 2025. “The corpus of spoken Yiddish in Europe: Goals, methods, and applications.” *Language Documentation and Conservation* 19:142–157. https://hdl.handle.net/10125/74812.

Bogdanova-Beglarian, Natalia, Gregory Martynenko, and Tatiana Sherstinova. 2015. “The "One Day of Speech" corpus: Phonetic and syntactic studies of everyday spoken Russian.” In *International Conference on Speech and Computer*, edited by Andrey Ronzhin, Rodmonga Potapova, and Nikos Fakotakis, 429–437. Cham: Springer. https://doi.org/10.1007/978-3-319-23132-7_53.

Bolly, Catherine T., and Dominique Boutet. 2018. “The multimodal CorpAGEst corpus: Keeping an eye on pragmatic competence in later life.” *Corpora* 13 (3): 279–317. https://doi.org/10.3366/cor.2018.0151.

D’Alessandro, Roberta, David Natvig, and Michael T. Putnam. 2021. “Addressing challenges in formal research on moribund heritage languages: A path forward.” *Frontiers in Psychology* 12. https://doi.org/10.3389/fpsyg.2021.700126.

Davis, Boyd. 2010. “AlzTalk: The Carolinas Conversation Collection.” *American Association of Geriatric Psychiatry* (Savannah).

Dovetto, Francesca M., and Francesca Marra. 2024. “The FRA Corpus from the DISease and AGEing Project.” *Journal of Language and Aging Research (JLAR)* 2:64–78. https://doi.org/10.15460/jlar.2024.2.2.1522.

Dubinina, Irina, and Olesya Kisselev. 2025. "Shifting Russian-speaking diasporas: New directions in the study of Russian as a heritage language." In *Syntax in uncharted territories: Essays in honor of Maria Polinsky*, edited by Lauren Clemens, Vera Gribanova, and Gregory Scontras. Irvine: University of California. https://escholarship.org/uc/item/0xj6b5t3.

Duboisdindien, Guillaume, and Catherine Bolly. 2024. "Videos to study Interactions in AGEing (VIntAGE): A longitudinal, ecological and multimodal French corpus." *Journal for Language and Aging Research* 2:50–63. https://doi.org/10.15460/jlar.2024.2.2.1524.

Gerstenberg, Annette. 2011. *Generation und Sprachprofile im höheren Lebensalter: Untersuchungen zum Französischen auf der Basis eines Korpus biographischer Interviews.* Frankfurt am Main: Vittorio Klostermann. https://doi.org/10.5281/zenodo.7315875.

Granger, Sylviane. 2004. "Computer learner corpus research: Current status and future prospects." *Applied Corpus Linguistics* 52:123–145. https://doi.org/10.1163/9789004333772_008.

Hatton, John, and Christopher Hirt. 2025. *Lameta: The Metadata Editor for Transparent Archiving of Language Document Materials, V.2.3-beta 2025-09-06.* Summer Institute of Linguistics: GitHub / Summer Institute of Linguistics. https://github.com/onset/lameta/releases/tag/v2.3.16-beta.

Housen, Alex, Bastien De Clerqc, Folkert Kuiken, and Ineke Vedder. 2019. "Multiple approaches to complexity in second language research." *Second Language Research* 35 (1): 3–22. https://doi.org/10.1177/026765831880976.

Johannessen, Janne B. 2015. "The Corpus of American Norwegian speech (CANS)." In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015*, edited by Beáta Megyesi, 297–300. Sweden: Linköping University Electronic Press. https://tekstlab.uio.no/norskiamerika/english/corpus.html.

Laleko, Oksana, and Olesya Kisselev. 2021. "Heritage language complexity matters: The Editors' Introduction to the Special Issue." *Heritage Language Journal* 18 (2): 1–8. https://doi.org/10.1163/15507076-12340008.

Lanzi, Alyssa M., Anna K. Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L. Cohen. 2023. "DementiaBank: Theoretical rationale, protocol, and illustrative analyses." *American Journal of Speech-Language Pathology* 32 (2): 426–438. https://doi.org/10.1044/2022_AJSLP-22-00281.

Lewis, Paul, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World.* Twenty-fourth. Dallas, (TX): SIL International.

Liu, Houjun, and Brian MacWhinney. 2024. "Morphosyntactic analysis for CHILDES." *Language Development Research* 4 (1): 233–258. https://doi.org/10.34842/j97r-n823.

Lu, Marcus. 2024. *Ranked: The top languages spoken in the world.* Visual Capitalist. https://www.visualcapitalist.com/top-languages-spoken-in-the-world/.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk.* 3rd ed. Mahwah (NJ): Lawrence Erlbaum.

———. 2019. "Understanding spoken language through TalkBank." *Behaviour Research Methods* 51 (4). https://doi.org/10.3758/s13428-018-1174-9. https://talkbank.org/.

MacWhinney, Brian, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. "AphasiaBank: Methods for studying discourse." *Aphasiology* 25 (11): 1286–1307. https://doi.org/10.1080/02687038.2011.589893.

Nagano, Tomonori. 2015. "Demographics of the adult heritage language speakers in the United States: Differences by region and language and their implications." *The Modern Language Journal* 99 (4): 771–792. https://doi.org/10.1111/modl.12272.

Nagy, Naomi. 2009. *The Heritage Language Documentation Corpus (HerLD).* Toronto: University of Toronto. https://ngn.artsci.utoronto.ca/HLVC/1_4_corpus.php.

Polinsky, Maria, and Michael T. Putnam. 2024. *Formal approaches to complexity in heritage language grammars.* Berlin: Language Science Press.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. "Robust Speech Recognition via Large-Scale Weak Supervision [eess.AS]." *arXiv*, https://doi.org/10.48550/arXiv.2212.04356.

Rivers, William P., and Richard D. Brecht. 2018. "America's languages: The future of language advocacy." *Foreign Language Annals* 51 (1): 24–34. https://doi.org/10.1111/flan.12320.

Rothman, Jason. 2024. "Harnessing the bilingual descent down the mountain of life: Charting novel paths for Cognitive and Brain Reserves research." *Bilingualism: Language and Cognition* 3:1–9. https://doi.org/10.1017/S1366728924000026.

Ruseishvili, Svetlana. 2016. *Ser russo em São Paulo. Os imigrantes russos e a reformulação de identidade após a Revolução Bolchevique de 1917.* São Paulo. https://teses.usp.br/teses/disponiveis/8/8132/tde-13022017-124015/pt-br.php.

Sekerina, Irina A., Aleksandra S. Skorobogatova, and Anna Smirnova Henriques. 2025. *Brazilian Portuguese-Russian Corpus (BraPoRus).* TalkBank. https://doi.org/10.21415/CJV6-JY66. https://biling.talkbank.org/access/BraPoRus.html.

Sekerina, Irina A., Anna Smirnova Henriques, Aleksandra S. Skorobogatova, Nataliya Tyulina, Tatiana V. Kachkovskaia, Svetlana Ruseishvili, and Sandra Madureira. 2023. "Brazilian Portuguese-Russian (BraPoRus) Corpus: Automatic transcription and acoustic quality of elderly speech during Covid-19 pandemic." *Linguistics Vanguard* 9 (s4): 375–388. https://doi.org/10.1515/lingvan-2021-0149.

Skorobogatova, Aleksandra S., Anna Smirnova Henriques, Svetlana Ruseishvili, Irina A. Sekerina, and Sandra Madureira. 2021. "Verbal working memory assessment in Russian-Brazilian Portuguese bilinguals." *Cadernos de Linguística* 2 (4): 1–24, e572. https://doi.org/10.25189/2675-4916.2021.V2.N4.ID572.

Smirnova Henriques, Anna, Aleksandra S. Skorobogatova, Tatiana V. Kachkovskaia, Pavel A. Skrelin, Svetlana Ruseishvili, Sandra Madureira, and Irina A. Sekerina. 2022. "BraPoRus, spoken corpus of heritage Russian in Brazil: Protocol of data collection." *Cadernos de Linguística* 3 (1): 1–20, e629. https://doi.org/10.25189/2675-4916.2022.V3.N1.ID629.

Wiese, Heike, Artemis Alexiadou, Shanley Allen, Oliver Bunk, Natalia Gagarina, Kateryna Iefremenko, Esther Jahns, Martin Klotz, Thomas Krause, and Annika Labrenz. 2021. *RUEG Corpus 0.4.0.* Zenodo. https://zenodo.org/records/5808870.

World Health Organization. 2024. "Aging and health." *Fact sheets* (Geneva), https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

Table 5: BraPo-Rus-1.0: Lexical complexity

| Part. | Average length of words in: | | | Number of unique: | | TTR by: | |
|---|---|---|---|---|---|---|---|
| | Letters | Syllables | Morphemes | Tokens | Lemmas | Token | Lemma |
| AVG | 5.06 (4.26) | 1.36 (1.29) | 1.46 (0.71) | 683 | 572 | 73.48 | 45.16 |
| AVM | 4.47 (3.68) | 1.35 (1.15) | 1.43 (0.73) | 644 | 503 | 67.83 | 49.19 |
| EAB | 4.42 (3.65) | 1.34 (1.08) | 1.48 (0.65) | 535 | 431 | 54.82 | 33.48 |
| EKS | 4.93 (4.14) | 1.29 (1.13) | 1.47 (0.71) | 544 | 435 | 69.4 | 42.06 |
| ENL | 4.18 (3.66) | 1.21 (1.07) | 1.42 (0.71) | 627 | 507 | 55.62 | 34.73 |
| FFK | 4.59 (3.82) | 1.30 (1.13) | 1.48 (0.65) | 589 | 466 | 80.89 | 47.05 |
| GAA | 4.32 (3.74) | 1.24 (1.11) | 1.38 (0.72) | 653 | 509 | 70.16 | 40.98 |
| IMK | 4.77 (4.21) | 1.20 (1.12) | 1.50 (0.79) | 665 | 523 | 54.05 | 30.54 |
| LNI | 4.28 (3.47) | 1.26 (1.09) | 1.61 (0.83) | 552 | 433 | 45.08 | 35.23 |
| NVM | 4.68 (4.03) | 1.25 (1.19) | 1.48 (0.75) | 544 | 454 | 71.11 | 52.37 |
| SAK | 4.76 (3.92) | 1.34 (1.10) | 1.49 (0.73) | 635 | 512 | 64.33 | 38.39 |
| SAP | 4.64 (4.12) | 1.26 (1.15) | 1.46 (0.80) | 538 | 426 | 47.76 | 31.19 |
| TNK | 4.82 (4.16) | 1.28 (1.22) | 1.44 (0.71) | 501 | 411 | 48.21 | 30.75 |
| TYL | 5.11 (4.36) | 1.24 (1.19) | 1.39 (0.67) | 705 | 579 | 69.46 | 41.34 |
| VVG | 5.21 (4.40) | 1.25 (1.25) | 1.39 (0.66) | 517 | 430 | 62.24 | 35.29 |
| ZVH | 4.70 (3.89) | 1.31 (1.16) | 1.48 (0.76) | 1086 | 850 | 74.44 | 51.59 |