

## New Speakers of Belarusian: A Spoken Corpus Creation

A growing body of research (O'Rurke & Ramallo, 2015; Wicherkiewicz et al., 2018; Kudriavtseva, 2025) has been focusing on new speakers of minority languages, who are typically defined as “those who have learned a minority language not within their home or community settings but through bilingual education, immersion or migration” (Williams, 2023, p.1). From 2020, Belarusian, an endangered East Slavic language (Moseley, 2010), has been growing in popularity among many Russian-dominant Belarusians, with some of them completely switching to Belarusian in their everyday life (Głaz, 2025; Tratsiak, 2025). This wave of language revitalization is the fourth one reported in connection to Belarusian (Woolhiser, 2013), being potentially the largest so far (Masliankova, 2025).

The current study focuses on the creation of the first spoken corpus of Belarusian used by new speakers in the Belarusian diaspora from Poland and Lithuania. The approximate number of speakers is 42 people of different Belarusian proficiency levels: low-proficiency speakers (corresponding to CEFR A1-A2), medium-proficiency speakers (corresponding to CEFR B1-B2), high-proficiency speakers (corresponding to CEFR C1-C2), with each group having 14 speakers. Ideally, each group will be gender-balanced, with 7 males and 7 females. To understand the speakers' proficiency and language background, the author intends to give the speakers Rapid Automatized Picture Naming task for both Belarusian and Russian (similar to Borges, 2019) and Bilingual Language Profile task (Birdsong et al., 2012) combined with some qualitative questions asking about their language attitudes and beliefs (adapted from Kravchenko, 2025). The speakers' spontaneous speech will be collected with the help of a sociolinguistic interview (Labov, 1984), while semi-spontaneous speech will involve the speakers recorded doing a map task (Drager, 2018). The speech of the speakers will be transcribed with the help of a finetuned Whisper Automatic Speech Recognition system (navalnica, 2022) and aligned using Montreal Forced Aligner (McAuliffe et al., 2017).

The finalized methodology connected with the corpus creation, as well as the initial results of data collection, will be presented during the conference. The finished corpus will be freely accessible upon publication, available to the community and all researchers interested in the Belarusian language.

(346 words)

## References

- Birdsong, D., Gertken, L. M., & Amengual, M. (2012). *Bilingual language profile: An easy-to-use instrument to assess bilingualism*. COERLL, University of Texas at Austin. Retrieved from <https://blp.coerll.utexas.edu>
- Borges, R. (2019). Rapid automatized picture naming as a proficiency assessment for endangered language contexts: Results from Wilamowice. *Journal of Communication and Cultural Trends*, 1(1), 1–25. <https://doi.org/10.32350/jcct.11.01>
- Drager, K. K. (2018). *Experimental research methods in sociolinguistics*. Bloomsbury Academic Publishing Plc.
- Głaz, A. (2025). Struggling with words and worlds after February 24: Ukrainian and Belarusian students in Poland. In *Movy ta kul'tury pid chas viyny: Kolektyvna monohrafiya* = Languages and cultures in times of war: Collective monograph (pp. 191–206). Uzhhorod National University.
- Kravchenko, K. (2025). *An analysis of Ukrainians' language attitudes and ideology: A conflict-catalyzed language and identity shift* (Master's thesis, Brigham Young University). BYU ScholarsArchive. <https://scholarsarchive.byu.edu/etd/11084>
- Kudriavtseva, N. (2025). New speakers of Ukrainian: Ideologies of linguistic conversion. *Language in Society*, 1–20. <https://doi.org/10.1017/S0047404525101711>
- Labov, W. (1984). Field methods of the project on linguistic change and variation. In J. Baugh & J. Sherzer (Eds.), *Language in use: Readings in sociolinguistics* (pp. 28–54). Prentice-Hall.
- Masliankova, P. (2025). Mova i toesnasc': Karani matyvacyi vyvuchennja belaruskaj movy sjarod daroslyx belarusaŭ. *Acta Albaruthenica*, (25). <https://doi.org/10.31338/2720-698Xaa.25.8>

- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kald. *Interspeech 2017*, 498–502. <https://doi.org/10.21437/Interspeech.2017-1386>
- Moseley, C. (Ed.). (2010). *Atlas of the world's languages in danger* (3rd ed.). UNESCO Publishing.
- navalnica. (2022). *whisper-finetuning-be* [Computer software]. GitHub. <https://github.com/navalnica/whisper-finetuning-be>
- O'Rourke, B., & Ramallo, F. (2015). Neofalantes as an active minority: Understanding language practices and motivations for change amongst new speakers of Galician. *International Journal of the Sociology of Language*, 2015(231), 147–165. <https://doi.org/10.1515/ijsl-2014-0036>
- Tratsiak, V. (2025). The situation of the Belarusian language after the 2020 presidential elections. *Studia Białorutenistyczne*, 18, 241–252.
- Wicherkiewicz, T., Król, T., & Olko, J. (2018). Awakening the language and speakers' community of Wymysiöeryś. *European Review*, 26(1), 179–191. <https://doi.org/10.1017/S1062798717000424>
- Williams, C. H. (2023). *Language policy and the new speaker challenge: Hiding in plain sight* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/9781009047326>
- Woolhiser, C. (2013). New speakers of Belarusian: Metalinguistic discourse, social identity, and language use. In D. M. Bethea & C. Y. Bethin (Eds.), *American contributions to the 15th International Congress of Slavists* (pp. 63–115).