We introduce PRODIS, a speech corpus which has been constructed to close a conspicuous gap existing in current language resources for Polish. PRODIS aims to be the first large, publicly available Polish speech corpus of excellent acoustic quality. The design is 50 speakers, 30 hours of read and conversational speech, specific tasks include reading (Wikipedia entires), experimental reading (a prosodic production task) and a spontaneous dialogue. The most recent and largest database for Polish, SpokesBiz (Pęzik et al. 2023), remedies the neglects on the infrastructure scene by providing a database of 590 speakers and 650 hours with a manually verified transcription and ASR. SpokesBiz in its current version does not yet include phonemic alignment and the Presentation part contains recordings done via MS Teams, representing academic discourse.

Prodis will be a useful resource for phonetic studies and speech technology applications due to a number of unique features:

1) application of Automatic Speech recognition (ASR, Whisper AI) to proces the recordings
2) Manual verification of ASR generated orthographic transcripts
3) phonemic alignment using WebMAUS (Schiel et al. 2024); we have used WebMAUS instead of Montreal Forced alignment due to the pipeline which does not require to cut the files
4) Manual verification of forced alignment
5) Nearly automated (90 per cent) speech processing

Moreover, PRODIS offers a phoneme level Polish language model based on Nano GPT architecture by Karpathy (2023) and trained on Wikipedia texts. We trained the language model on machine-readable phoneme transcriptions (X-SAMPA) of the OSCAR-Mini text database. The OSCAR-Mini data was transcribed using an X-SAMPA inventory identical to the one used by the WebMaus aligner - assuring compatibility with speech alignment labels. The language model extracts and analyzes acoustic estimates of contextual predictability which is our research objective. This way, PRODIS provides a tool for measuring surprisal (Jaeger et al. 2017, Aylett and Turk, 2006, Malisz et al. 2018, Turnbull et al. 2015) and, in future, may calculate other estimates of predictability in speech.

PRODIS is not just another speech database as it incorporates a state-of-the-art, freely available tools (such as a phoneme-based langauge model, a phonemizer, a character tokenizer) enabling database expansion or adaptation to additional languages.

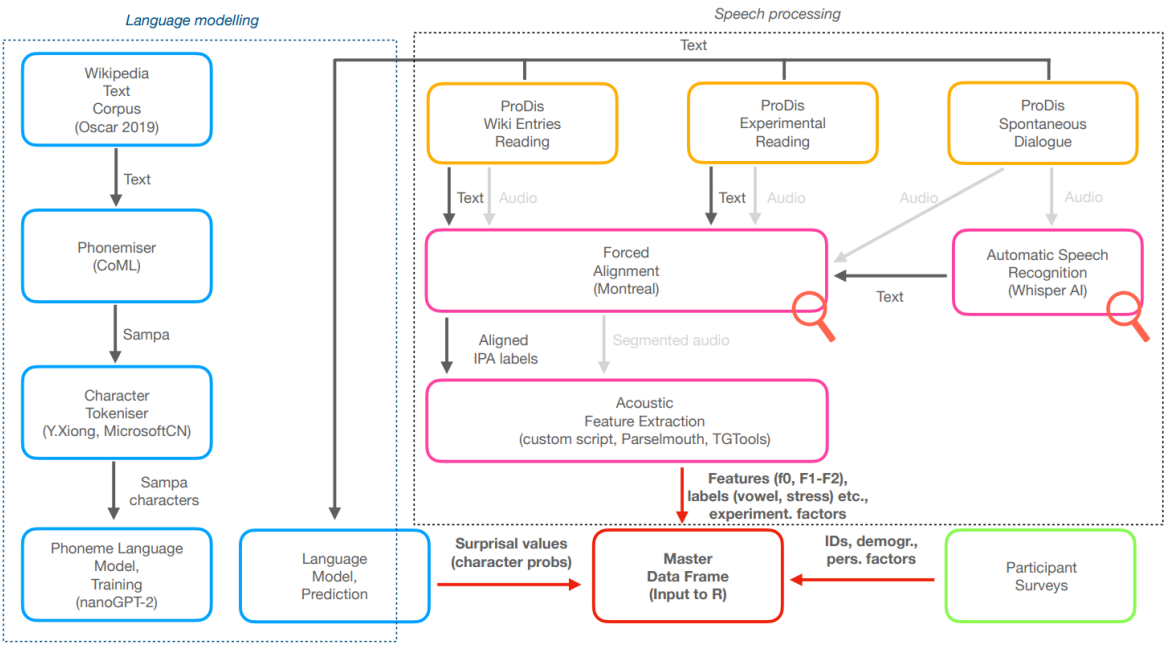

Figure 1: PRODIS: database design, speech processing and language modeling pipelines. The magnifying glass symbolises manual evaluation processes.

Selected references:

Aylett, Z. and A. Turk. (2006). *Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei*. Journal of Acoustic Society of America, 119:30–48.

https://prodis-opus19.github.io/

Jaeger, T. F., and Buz, E. (2017). Signal reduction and linguistic encoding. *The handbook of psycholinguistics*, 38-81.

Karpathy, A. (2023) *Nano-GPT.* https:// github.com/karpathy/nanoGPT

Malisz, Z., Brandt, E., Möbius, B., Oh, Y. M., and B. Andreeva. (2018). Dimensions of segmental variability: Interaction of prosody and surprisal in six languages. *Frontiers in Communication*, *3*, 25.

McAuliffe, M. and M. Sonderegger. (2022). *Polish mfa dictionary v2.0.0a. Technical reeport*, https://mfa-models.readthedocs. io/pronunciationdictionary/Polish/ PolishMFAdictionaryv2_0_0a.html.

OpenAI. Whisper. https://github.com/ openai/whisper. Accessed: 2022-12-06.

Ortiz Suarez, P. J., Romary, L., and B. Sagot. (2020). *A monolingual approach to contextualized word embeddings for mid-resource language*s. In Proceedings of the 58th Annual Meeting of the Association for Compuatational Linguistics, pages 1703–1714, Online. Association for Computational Linguistics.

Pęzik, P., Karasińska, S., Cichosz, A., Jałowiecki, Ł., Kaczyński, K., Krawentek, M. and S. Marszałkowski. (2023). *SpokesBiz--an Open Corpus of Conversational Polish*. arXiv preprint arXiv:2312.12364.

Schiel, F., Draxler, C., & Harrington, J. (2024). *Munich AUtomatic Segmentation (MAUS)*. https://www.bas.uni-muenchen.de/Bas/BasMAUS.html

Socolof, M., Stengel-Eskin, E., Mihuc, S. Wagner, M., McAuliffe, M. and M. Sonderegger. (2017). *Montreal forced aligner* [computer program]. version 1.0.0.

Turnbull, R., Burdin, R. S., Clopper, C. G.,and J. Tonhauser. (2015). Contextual predictability and the prosodic realisation of focus: A cross-linguistic comparison. *Language, Cognition and Neuroscience*, *30*(9), 1061-1076.