EUROPEAN LANGUAGE EQUALITY

META-NET
META-FORUM 2023

Multilingual & Mixed Language Data for Inclusive Speech Technologies

A. Seza Doğruöz (Ghent University)
as.dogruoz@ugent.be

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu

Background

- There are millions of multilingual speakers who speak more than one language and mix them for daily communication around the world and in Europe (Doğruöz et al, 2021).
- However, current language technologies are mostly built with monolingual assumptions and ignore the variation & diversity among different types of speakers/users (Doğruöz & Sitaram, 2022).
- This project aims at building a multilingual and mixed language speech data set in Belgium.

Goals

- Belgium is a multilingual country with three official languages (Dutch, French, German).
- In addition to widely spoken English, there are also other languages spoken by the immigrant communities.
- This project focused on collecting conversational and mixed language data (Turkish-Dutch and some English) in Belgium.

Outcomes

- Free conversations (approx. 10 hours) between (20) speakers from the target group were audio-recorded.
- There are no automatic tools to transcribe the mixed language use in these language pairs.
- The recorded conversations were transcribed manually according to the transcription guidelines developed for the project.
- The project contributes to the multilingual and mixed data collection for speech technologies in Europe and aligns with the language equality principles of the strategic agenda (ELE).

References

- Doğruöz, A.S., Sitaram, S., Bullock, B.E., Toribio, A.J. (2021). *A Survey of Code-switching: Linguistic and Social Perspectives for Language Technologies*, Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP'2021), Bangkok, Thailand.
- Doğruöz, A.S. & Sitaram, S. (2022). *Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights*. Proceedings of SIGUL at LREC'22. European Language Resources Association.

Thank you!

The European Language Equality project has received funding from the European Union under grant agreements № LC-01641480 – 101018166 (ELE) and № LC-01884166 – 101075356 (ELE2).

A. Seza Doğruöz (Ghent University)
as.dogruoz@ugent.be

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu