EUROPEAN LANGUAGE EQUALITY 2

FSTP Project Report

# MUMIST – Multilingual and Mixed Language Data for Inclusive Speech Technology

|  |  |
|---|---|
| Authors | A. Seza Doğruöz |
| Organisation | Universiteit Gent |
| Dissemination level | Public |
| Date | DD-MM-2023 |

# About this document

| | |
|---|---|
| Project | European Language Equality 2 (ELE2) |
| Grant agreement no. | LC-01884166 – 101075356 ELE2 |
| Coordinator | Prof. Dr. Andy Way (DCU) |
| Co-coordinator | Prof. Dr. Georg Rehm (DFKI) |
| Start date, duration | 01-07-2022, 12 months |
| FSTP Project | MUMIST – Multilingual and Mixed Language Data for Inclusive Speech Technology |
| Authors | A. Seza Doğruöz |
| Organisation | Universiteit Gent |
| Type | Report |
| Number of pages | 9 |
| Status and version | Draft |
| Dissemination level | Public |
| Date of delivery | DD-MM-2023 |
| EC project officer | Susan Fraser |
| Contact | European Language Equality 2 (ELE2)<br>ADAPT Centre, Dublin City University<br>Glasnevin, Dublin 9, Ireland<br>Prof. Dr. Andy Way – andy.way@adaptcentre.ie<br>European Language Equality 2 (ELE2)<br>DFKI GmbH<br>Alt-Moabit 91c, 10559 Berlin, Germany<br>Prof. Dr. Georg Rehm – georg.rehm@dfki.de<br>http://www.european-language-equality.eu<br>© 2023 ELE2 Consortium |

# Consortium

| | | | |
|---|---|---|---|
| 1 | Dublin City University (Coordinator) | DCU | IE |
| 2 | Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator) | DFKI | DE |
| 3 | Univerzita Karlova (Charles University) | CUNI | CZ |
| 4 | Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country) | UPV/EHU | ES |
| 5 | Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis | ILSP | GR |
| 6 | European Federation of National Institutes for Language | EFNIL | LU |
| 7 | Réseau européen pour l’égalité des langues (European Language Equality Network) | ELEN | FR |

# Contents

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| AI | Artificial Intelligence |
| AI4EU | AI4EU (EU project, 2019-2021) |
| CLAIRE | Confederation of Laboratories for AI Research in Europe |
| CLARIN | Common Language Resources and Technology Infrastructure |
| CRACKER | Cracking the Language Barrier (EU project, 2015–2017) |
| DARIAH | Digital Research Infrastructure for the Arts and Humanities |
| DLE | Digital Language Equality |
| EC | European Commission |
| ECSPM | European Civil Society Platform for Multilingualism |
| EFNIL | European Federation of National Institutes for Language |
| ELE | European Language Equality |
| ELE2 | European Language Equality *(this project)* |
| ELE Programme | European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)* |
| ELEN | European Language Equality Network |
| ELEXIS | European Lexicographic Infrastructure |
| ELG | European Language Grid (EU project, 2019-2022) |
| ELRA | European Language Resource Association |
| ELRC | European Language Resource Coordination |
| ELT | European Language Technology |
| EP | European Parliament |
| ERIC | European Research Infrastructure Consortium |
| ESCO | European Skills, Competences, Qualifications and Occupations classification |
| GDPR | General Data Protection Regulation |
| KPI | Key Performance Indicator |
| LT | Language Technology/Technologies |
| META | Multilingual Europe Technology Alliance |
| META-NET | EU Network of Excellence to foster META |
| ML | Machine Learning |
| MT | Machine Translation |
| NCC | National Competence Centre |
| NCP | National Contact Point |
| NLP | Natural Language Processing |
| STOA | Science and Technology Options Assessment |

## Abstract

There are millions of multilingual speakers who speak more than one language and mix them in daily communication around the world. However, current speech and language technologies are built with monolingual assumptions ignoring the variation (e.g., social, linguistic and cultural) among different types of speakers/users and there is a lack of multilingual and mixed language data to build speech technologies in Europe. This leads to language inequalities according to the strategic agenda of ELE especially for low resource languages. To fill this gap, this project focuses on collecting and transcribing multilingual and mixed language spoken data focusing on a bilingual/multilingual community in Belgium. In addition, we provide some recommendations about data collection and transcription for other low resource languages based on the challenges and solutions we have encountered in this project.

## 1 Introduction

Multilingualism is widely spread among millions of speakers/users around the world (e.g., Europe, South East Asia, Africa, South America) and it refers to speaking more than one language/dialect on a daily basis and/or mixing them (Doğruöz et al., 2021; Sridhar, 2002; Treffers-Daller, 2009; Herkenrath, 2012). However, most speech technologies are currently built with monolingual assumptions and they are not able to handle multilingual and mixed language communication (Doğruöz and Sitaram, 2022). Speech data representing multilingual speakers and/or mixed languages/dialects in Europe are largely missing. In accordance with the language equality principles of the strategic agenda of ELE, this study focuses on collecting conversational and mixed language data from an immigrant community in Belgium which is a multilingual country with three official languages (i.e., Dutch, French, German). Dutch (also referred as ”Flemish”) is spoken in the Flanders region and Brussels. Although English is not an official language, it is widely spoken as well. In addition, there are also languages (e.g., Turkish, Arabic) spoken by the immigrant communities. Although there is some earlier research on Turkish-Dutch bilingualism in the Netherlands (Doğruöz and Backus, 2009), these data sets do not represent the multilingual communication for similar communities in the Belgian context since there are also dialectal differences (e.g., vocabulary, intonation, grammar) between Dutch spoken in Belgium and the Netherlands. Therefore, there is a need for new speech data collection. Turkish community is estimated to be around 300,000 in Belgium and it is one of the largest immigrant communities with more than 60 years of immigration history. Similar to other multilingual immigrant contexts around the world, younger generations of the community grow up bilingually (Turkish and Dutch simultaneously) and make use of English in their daily lives as well. So far, speech data regarding the multilingual language use among the members of this immigrant community is lacking. To fill this gap, this project focuses on collecting and transcribing multilingual (i.e. Turkish, Dutch and English) conversational data in this context.

## 2 Data

This project focused on multilingual and spoken data collection in Belgium focusing on Turkish-Dutch bilinguals who also made use of English words and/or phrases in some parts of the conversations. The data was collected through an audio recorder and it involves natural conversations between bilingual speakers in pairs. In total, there are 10 recordings (approx. 1 hour each) between 20 participants (between the ages of 18-35). All the participants were compensated for their participation in the project based on the regulations of the host institute. Considering the limited time frame for this project, it was not possible to conduct a language proficiency test to measure the language skills of the participants in detail. Moreover, both Polinsky (2018) Doğruöz (2022) report challenges and drawbacks about conducting proficiency tests for heritage language speakers in immigrant contexts. Instead of the proficiency test, all the participants in this project reported the languages they speak briefly in a survey. In general, they were all born into immigrant families in Belgium and grew up bilingually (Turkish-Dutch) starting from an early age onwards. Table (1) presents the number of words per recorded conversations between the participants for this data set.

| Conversations | Word Counts |
|---|---|
| 1 | 7346 |
| 2 | 6535 |
| 3 | 7637 |
| 4 | 4285 |
| 5 | 8646 |
| 6 | 5975 |
| 7 | 7671 |
| 8 | 7131 |
| 9 | 8772 |
| 10 | 9180 |
| Total | 73178 |

Table 1: Speech Data Set and Word Counts

# 3 Methodology

Before starting the project, it was necessary to apply for an internal ethical approval at the host institute and we followed their recommendations throughout the project in terms of data collection and GDPR regulations. To reach out to the target population, multilingual student assistants who were active in the community were hired. In the meanwhile, a short survey was developed to collect brief background information (e.g., languages, age) about the speakers who will participate in the data collection. Upon hiring the student assistants, they were given trainings about the data collection procedures and how to use the audio recording devices. The participants were approached through the networks of the student assistants and snowball sampling method. During the data collection, the speakers were instructed to converse freely without any topic limitations to achieve natural and mixed language conversations as they occur in real life. Since there are no automatic tools available to transcribe the mixed language communication in this data set, we opted for a manual transcription method. After searching for the best transcription tool, we decided on ELAN which fits the purposes of this project better. The student assistants received a training about how to use this tool for transcription purposes. Based on the earlier experiences of the researcher, there are differences in the spoken and written language conventions. In addition, mixed language use creates extra difficulties for novice transcribers (e.g. the student assistants in this project). In order to minimize these difficulties, we held regular meetings with the transcribers for training purposes and also sharing the updates with respect to the commonly encountered challenges and solutions. To protect the privacy of the speakers, a first iteration for the pseudonymization of the sensitive information (e.g., NERs) in the transcribed data has also been performed manually. If necessary, additional measures will be taken to protect the privacy of the participants as well.

## 4 Challenges & (possible) Solutions

Although collecting and analyzing language data (especially spoken data) is needed and valued in low resource and multilingual settings, there are also multiple challenges which need to be tackled. First, and perhaps most important challenge is to reach out to the community members to participate in the project. This becomes especially challenging when the researchers do not have a social network within the designated community. Hiring student assistants and/or local members of the community are viable options to overcome this challenge. Second, any type of data collection project (but especially spoken data) has to be screened through multiple committees (e.g, Ethical committees) in most academic host institutions (in Europe). In addition, there are multiple procedures to be followed to secure the privacy of the participants and the collected data. It takes time to learn about these procedures and follow their timelines (e.g. some committees meet only once a month) which becomes challenging for a short term ELE project. Although there are similarities between the regulations for data collection and preservation across different host institutions in Europe, there are also differences in terms of the procedures and timelines. In other words, it is not easy to come up with a standard guideline to speed up these procedures for each and every ELE project. Instead, it is highly recommended to get informed with the internal procedures of the host institution about the data collection and preservation in a timely manner to prevent delays if/when the ELE project is granted. Ideally, it is useful to collect language data from diverse sets of participants (with different backgrounds) to understand the language use across different communities/populations. However, collecting data from different populations (e.g. under 18 or above 65) may require different official procedures and regulations. Considering the limited time frame of the project, it is more feasible to restrict the project to the target groups who are easier to reach out both in terms of networking and following the related official procedures for data collection. Since there are no available tools for the automatic transcription of the data set in this project, manual transcription was necessary. However, finding skilled transcribers in low resource language contexts is not always easy. As mentioned by Polinsky (2018); Doğruöz (2022), heritage speakers may grow up bilingually but it does not mean that all aspects of their language skills (e.g. speaking vs. writing) are at the same level. While conducting a similar project in other low resource language settings, it is recommended to provide regular trainings for the data collectors and transcribers about the task in hand to prevent further complications. Morphemes of agglutinative languages (e.g., Turkish) may undergo reduction (e.g., tense or person markers on the verb) in conversational contexts and these reductions reflect the linguistic diversity and sociolinguistic variation in terms of the participants, context and medium of communication. However, transcribing the data as it is spoken in daily communication creates challenges for follow-up computational tasks (e.g., POS) since most tools which are available in these languages are developed and trained using written and standard language data (Nguyen et al., 2016). To overcome this challenge, we transcribed the collected data using standard language but also provided layers of explanations and comments to indicate spoken language conventions when necessary. Table (2) summarizes these challenges and potential solutions in this project with the goal of providing insights for the next ELE cohort.

| Challenges | Solutions |
|---|---|
| Lack of Community Network | Hiring student assistants from the same community |
| Multiple and Lengthy procedures for Data collection and preservation | Start with the official procedures very early to prevent delays |
| Different procedures for different types of data collection | Focus on one type of data within the limited period of time |
| Challenges for automatic transcription of mixed language data | Manual transcription |
| Challenges for transcription in low resource settings | Trainings and regular meetings with the transcribers |
| Spoken vs. Written language conventions for Transcription | If possible, provide both (with extra layers) |

Table 2: Challenges and Solutions for Multilingual Data Collection

## 5 Summary and Conclusions

Current speech technologies are built considering the needs and preferences of the groups of speakers communicating in the standard language and/or dialect. This assumption ignores the inherent sociolinguistic variation within the same language (e.g., different dialects, variation across speakers with different social and linguistic backgrounds) and the mixed language communication. As a result, the needs and preferences of underrepresented communities and their members (especially for languages which are low resourced and/or mixed) are not met. This project has aimed to fill this gap by collecting a conversational data set from the adult (+18) and bilingual speakers who are bilingual in Turkish and Dutch in Belgium. Although the duration of the project was limited and there were many challenges along the way, the project has achieved its goals successfully in terms of data collection, transcription and promotion (both the project and ELE) in international scientific venues. We hope the challenges and (potential) solutions provided in this study could also serve other researchers well and research teams who are planning to apply for the ELE competition in the future. In addition, new data sets for the same community could be also collected from different speaker groups to measure the similarities and differences between language use in the current data set.

## References

A. Seza Doğruöz and Sunayana Sitaram. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.sigul-1.12.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.131. URL https://aclanthology.org/2021.acl-long.131.

A. Seza Doğruöz. Issues about analysing multilingual communication in immigrant contexts. In Albert Ali Salah, Emre Eren Korkmaz, and Tuba Bircan, editors, *Data science for migration and mobility*, volume 251 of *Proceedings of the British Academy*, pages 334–349. British Academy, Oxford University Press, 2022. ISBN 9780197267103.

A. Seza Doğruöz and Ad Backus. Innovative constructions in dutch turkish: An assessment of on-going contact-induced change. *Bilingualism: Language and Cognition*, 11(2):185–220, 2009.

Annette Herkenrath. Receptive multilingualism in an immigrant constellation: Examples from turkish–german children's language. *International Journal of Bilingualism*, 16:287–314, 2012.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593, 2016.

Maria Polinsky. *Heritage languages and their speakers*, volume 159. Cambridge University Press, 2018.

Kamal Sridhar. Societal multilingualism, world englishes: their implications for tesol. *Indian Journal of Applied Linguistics*, 28(2):83–100, 2002.

Jeanine Treffers-Daller. Code-switching and transfer. In Barbara Bullock and Almedia Jaqueline Toribio, editors, *The Cambridge Handbook of Linguistic Code-switching*, pages 58–74. Cambridge University Press, Cambridge, 2009.