ZuBidasoa: Participatory Research for the Development of Linguistic Technologies Adapted to the Needs of Migrants in the Basque Country

Xabier Soto[1,2], Ander Egurtzegi[2], Maite Oronoz[1], Urtzi Etxeberria[2]
[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU,
[2]CNRS - IKER UMR5478
xabier.soto@ehu.eus

Abstract

Recent years have witnessed the development of advanced language technologies, including the use of audio and images as part of multimodal systems. However, these models are not adapted to the specific needs of migrants and Non-Governmental Organizations (NGOs) communicating in multilingual scenarios. In this project, we focus on the situation of migrants arriving in the Basque Country, nearby the western border between Spain and France. For identifying migrants' needs, we have met with several organisations helping them in different stages, including: sea rescue; primary care in refugee camps and *in situ*; assistance with asylum demands; other administrative issues; and human rights defence in retention centres. In these interviews, Darija has been identified as the most spoken language among the under-served ones. Considering this, we have started the development of a Machine Translation (MT) system between Basque and Darija (Moroccan Arabic), based on open-source corpora. In this paper, we present the description of the project and the main results of the participatory research developed in the initial stage.

1 Introduction

ZuBidasoa project aims to use MT as a bridge for improving the communication between migrants and NGOs. The project, developed between HiTZ - UPV/EHU and CNRS - IKER UMR5478, will last 3-4 years (from 2024 up to 2028) and is funded by the Basque government (project reference: POS_2023_1_0035).

The first stage of this project focuses on the participatory research carried out with 12 NGOs assisting migrants in the Basque Country, based in the cross-border cities of Donostia, Irun, Hendaia and Baiona.

For the first phase of this project, we have defined the following research questions[1]:

1. Among the NGOs working with migrants in the Basque Country, what is the knowledge and use of language technologies?

2. Are current Natural Language Processing (NLP) tools enough to meet the language needs of migrants and related NGOs?

3. How can we use MT to improve the communication between migrants and NGO members, as well as the internal work of NGOs?

2 Related Work

Recently, Maher et al. (2024) have broadly covered translation and migration research.

Extant work geographically closer to ours is done in Spain by Rico et al. (2020), describing a project developed with Caritas[2] and CEAR[3] to translate their documents from Spanish to English, French, Russian, Arabic and Chinese using *ad hoc* Neural Machine Translation (NMT) systems.

More specifically, Macken et al. (2024) presents a platform to be used in asylum reception centres in Belgium "to translate English, French or Dutch text messages into a set of at least 14 languages, including low-resourced languages such as Pashto, Somali and Tigrinya".

Compared to the previous work, the contributions of this project are the following:

1. We work in a cross-border location, where Basque, Spanish and French are spoken by many people, especially NGO members.

2. We consider the diglossic situation in the Basque Country, where Basque is minoritised with respect to Spanish and French.

[1] Adapted from Tesseur et al. (2022)

[2] https://www.caritas.es/

[3] https://www.cear.es/

3. We plan to develop MT systems for translating between two under-served languages, in our case Basque and Darija (Moroccan Arabic).

3 Main Results

Regarding the above research questions, from the NGO members interviewed we conclude that:

1. their knowledge and use of language technologies can be defined as basic. Most of the groups make use of Google Translate[4], one of the interviewed mentioned difficulties to use it, while another one used an MT tool and a dictionary better suited for Basque[5].

2. the current NLP tools are not enough to satisfy the needs of migrants and organisations working with them. Some NGOs prefer interpretation for dealing with medical or juridical issues, while others mention that automatic tools may suffice provided that these work better for specific domains and languages.

3. in all the cities under study, there is a linguistic/cultural gap between NGOs and Darija speaking migrants. Thus, a way to improve communication between migrants and NGO members would be the development of a Basque/Darija MT system, considering the possibility of translating audio and images.

The election of Darija as a language is confirmed by a recent study[6] by Gaindegia[7], stating that Morocco is the most common country of origin for migrants arriving in the Basque Country (after Spain and France). Even if Modern Standard Arabic is the main written language in Morocco, Darija is the most spoken language (HCP, 2024). When written, Darija uses both Arabic and Latin scripts.

During this initial research, we have identified a dataset (Outchakoucht and Es-Samaali, 2024)[8] with around 50,000 Darija/English sentences. In addition, both Basque and Darija are included in FLORES+[9], making it easier to evaluate future systems in a standardised way.

4 Conclusion and Future Work

Based on these and newly created corpora, we plan to develop MT models between Basque and Darija, using encoder-decoder NMT systems and instruction-tuned language models derived from Latxa (Etxaniz et al., 2024). In the future, we plan to adapt the systems to the legal domain and extend them to other languages. We will also explore the possibility of translating audio and images using visual and multimodal language models.

Acknowledgments

Xabier Soto is a researcher supported by the post-doctoral improvement program offered by the Basque Government (POS_2023_1_0035).

References

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.

Haut Commissariat au Plan du Maroc HCP. 2024. Recensement général de la population et de l'habitat 2024.

Lieve Macken, Ella Hest, Arda Tezcan, Michaël Lumingu, Katrijn Maryns, and July Wilde. 2024. MaTIAS: Machine translation to inform asylum seekers. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 6–7, Sheffield, UK. European Association for Machine Translation (EAMT).

Brigid Maher, Loredana Polezzi, and Rita Wilson, editors. 2024. *The Routledge Handbook of Translation and Migration (1st ed.)*. Routledge.

Aissam Outchakoucht and Hamza Es-Samaali. 2024. The evolution of darija open dataset: Introducing version 2. *Preprint*, arXiv:2405.13016.

Celia Rico, María Del Mar Sánchez Ramos, and Antoni Oliver. 2020. INMIGRA3: building a case for NGOs and NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 469–470, Lisboa, Portugal. European Association for Machine Translation.

Wine Tesseur, Sharon O'Brien, and Enida Friel. 2022. Language diversity and inclusion in humanitarian organisations: Mapping an ngo's language capacity and identifying linguistic challenges and solutions. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 21.

[4] https://translate.google.com/

[5] Elia: https://elia.eus/ and Elhuyar hiztegia: https://hiztegiak.elhuyar.eus/, respectively.

[6] https://shorturl.at/bIkTc

[7] https://www.gaindegia.eus/

[8] Darija Open Dataset: https://github.com/darija-open-dataset/dataset

[9] https://huggingface.co/datasets/openlanguagedata/flores_plus