The MULTI-TRAD Project: Parallel Corpora and Multidimensional Analysis of Human, Machine and Post-Edited Translation in the Third Social Sector

M.M. Sánchez Ramos[1], D. Biber[2], C. Cano Fernández[1], I. Fuentes Pérez[1], D. González Pastor[3], L. Goulart da Silva[4], M. Yuri Himoro[5], D. Kenny[6], L. M. Mónaco[7], M. T. Ortego Antón[8], I. Peñuelas Gil[8], C. Plaza Lara[9], V. Redondo Astilleros[10], C. Rico Pérez[11], T. Salvador Blázquez[12], M. Shakir[13], F. J. Vigier Moreno[14], M. Aenlle Curras[15]

[1]Universidad de Alcalá, [2]Northern Arizona University, [3]Universitat de València, [4]Montclair State University, [5]UNED, [6]Dublin City University, [7]Universidade da Coruña, [8]Universidad de Valladolid, [9]Universidad de Málaga, [10]Universitat Oberta de Catalunya, [11]Universidad Complutense de Madrid, [12]UDIMA, [13]University of Münster, [14]Universidad Pablo de Olavide, [15]Independent Consultant

Abstract

Domain adaptation remains a major challenge for machine translation, particularly in institutional communication. This paper presents the MULTI-TRAD project, which develops English–Spanish parallel corpora for the Third Social Sector communication. The project integrates three complementary objectives: (i) the compilation of a domain-specific parallel corpus, (ii) the analysis of linguistic variation across human translation (HT), machine translation (MT), and post-edited (PE) texts using Multidimensional Analysis (Biber, 1988), and (iii) the development of a domain-adapted neural machine translation system. In particular, the project investigates how different translation processes give rise to distinct functional profiles, related to phenomena such as translationese and post-editese. This paper presents the project design and initial progress.

1 Introduction

Communication within the Third Social Sector plays a central role in global public discourse. Organizations in this sector regularly produce multilingual content aimed at informing the public, mobilizing support and documenting institutional activities (Tesseur, 2017). Despite their importance, these texts remain underrepresented in corpus-based studies of translation and multilingual communication. In the field of machine translation, large-scale English–Spanish parallel corpora such as Europarl or ParaCrawl have played a central role in training neural systems. However, these resources mainly reflect general or institutional domains and do not adequately capture the communicative practices of the Third Social Sector, limiting both domain adaptation and evaluation in this context.

The MULTI-TRAD project addresses this gap through the development of English–Spanish parallel corpora designed to investigate linguistic variation across translation modes and to support domain-adapted machine translation. In addition to its relevance for translation studies, the project contributes to ongoing efforts in machine translation to move beyond purely metric-based evaluation by incorporating linguistically informed approaches to analyzing MT output.

2 Project overview

MULTI-TRAD is a research project funded by the Spanish Ministry of Science, Innovation and Universities (Grant PID2024-157849OB-I00) running from 2025 to 2028. The project involves a consortium of academic institutions and collaborating Third Social Sector organizations (e.g., Fundación Abrazando Ilusiones, Caritas, and Plataforma del Voluntariado en España).

The project integrates three complementary objectives. The first objective is the compilation of an English–Spanish parallel corpus. Texts are collected in collaboration

with Third Social Sector organizations and include institutional, advocacy, and web-based communication texts (Tesseur, 2017). The corpus is designed following a protocol that aims to capture a wide range of communicative practices within the Third Social Sector. The corpus includes texts originally produced in both English and Spanish. The dataset is designed for Multidimensional Analysis (MDA), including aligned human translation (HT), machine translation (MT), and post-edited (PE) versions. The MT texts will be generated using the domain-adapted neural machine translation system developed within the project. The second objective is the linguistic characterization of translation processes using MDA (Biber, 1988). This approach enables the investigation of functional variation across HT, MT, and PE outputs, including phenomena such as translationese and post-editese. The third objective is the development of a domain-adapted neural machine translation system tailored to Third Social Sector discourse. The system is trained on in-domain parallel data compiled from institutional materials using the MTUOC framework and the Marian NMT toolkit (Junczys-Dowmunt et al., 2018).

3 Multidimensional Analysis

The analysis is based on MDA (Biber, 1988), a framework grounded in register theory. MDA identifies co-occurring linguistic features and groups them into functional dimensions of variation. The corpus is tagged using the Multi-Feature Tagger of English (MFTE, Le Foll and Shakir, 2023), an open-source tool for multivariate analysis of English corpora. For the Spanish sub-corpus, we are currently evaluating state-of-the-art NLP frameworks (such as spaCy) to build a custom pipeline capable of extracting the equivalent lexico-grammatical features.

4 Preliminary results

The project is currently in the corpus compilation and preprocessing stage. On the one hand, an English–Spanish parallel corpus is being compiled for multidimensional linguistic analysis, including texts representing different domains and their human translations (HT). The corpus design follows established practices in MDA, with an initial target of approximately 25 texts per domain across four domains (Tesseur, 2017) and three translation modes (HT, MT, and PE), resulting in a balanced dataset of around 300 texts. Each text contains around 1,000 words to ensure reliable analysis. The corresponding MT and PE versions will be generated once the domain-adapted neural machine translation system has been trained. On the other hand, in-domain parallel data are being collected and curated for the training of the neural machine translation system. Although both strands rely on data from the same domain, the datasets are designed for different purposes: the corpus used for MDA requires carefully controlled and aligned HT, MT, and PE versions, whereas the training data consist of a larger and more heterogeneous in-domain collection.

References

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Junczys-Dowmunt, Marcin et al. 2018. Marian: Cost-effective high-quality neural machine translation in C++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation (WNMT)*, 129–135.

Le Foll, Elen and Shakir, Muhammad. 2023. MFTE Python (Version 1.0) [Software]. Available at: https://github.com/mshakirDr/MFTE

Tesseur, Wine. 2017. The translation challenges of NGOs: Professional and non-professional translation at Amnesty International. *Translation Spaces*, 6(2), 209–229.