CorCoDial – Machine translation techniques for corpus-based computational dialectology

Yves Scherrer **Olli Kuparinen** **Aleksandra Miletić**

Department of Digital Humanities, University of Helsinki, Finland

firstname.lastname@helsinki.fi

Abstract

This paper presents CorCoDial, a research project funded by the Academy of Finland aiming to leverage machine translation technology for corpus-based computational dialectology. In this paper, we briefly present intermediate results of our project-related research.

1 Introduction

Dialectology is concerned with the study of language variation across space. Over the last decades, dialectologists have collected large datasets, which typically consist of transcribed interviews with informants. Unfortunately, these interviews cannot easily be compared with each other as they differ considerably in length and content. If informant A does not use word x, this does not necessarily mean that the word does not exist in A's dialect. It may just be that A chose to talk about topics that did not require the use of word x. The CorCoDial (*Corpus-based computational dialectology*) project aims to introduce comparability in dialect corpora with the help of machine translation techniques. CorCoDial is funded by the Academy of Finland during the period 2021–2025.

The core of the project focuses on the dialect-to-standard normalization process, which is a sequence-to-sequence task that maps the phonetic transcriptions to the standardized spellings. We are not only interested in the results of the normalization process, but also in the emerging representations of dialects and speakers that the (statistical or neural) normalization models learn. These representations allow us to provide new visualisations of dialect landscapes and to confirm or challenge traditional dialect classifications.

Traditional dialect corpora are costly to produce: informants need to be found and interviewed, and the recorded interviews need to be transcribed and annotated. To circumvent this data bottleneck, researchers have increasingly turned to user-generated content (UGC), i.e., to texts published by laypeople on social media. We also investigate to what extent normalization methods trained on "clean" data transcribed by dialectologists generalize to noisier UGC datasets.

The main goals of the CorCoDial project are:

1. to improve the automatic normalization of dialect texts by using state-of-the-art machine translation methods,
2. to extract, visualize, compare and interpret the dialectal patterns emerging from the normalization models, and
3. to contrast the patterns found in traditional dialectological corpora with those found in user-generated content.

In the following sections, we present some results of our ongoing research.

2 Benchmarking dialect-to-standard normalization systems

In contrast to historical text normalization (Bollmann, 2019; Bawden et al., 2022) and UGC standardization, there have not been any multilingual evaluations of dialect-to-standard normalization systems. In order to establish dialect normalization as a distinct task, we compiled a multilingual benchmark dataset from existing sources, covering Finnish, Norwegian, Swiss German and Slovene.

We evaluate different sequence-to-sequence models that have been previously employed for normalization tasks:[1] statistical machine translation with character-level segmentation; neural machine translation with RNN and Transformer architectures, character-level and BPE segmentation,

[1] Note that normalization tasks, in contrast to other translation tasks, are monotonic. Although specific monotonic NMT architectures have been proposed, we follow earlier evaluations and focus on vanilla architectures. We leave the evaluation of normalization-specific architectures to future work.

and full-sentence and word-trigram windows; and the pre-trained multilingual ByT5 model using byte-level segmentation.

Our results show that the Transformer is the most successful model architecture on all four datasets. This is somewhat surprising since recent related work (Bollmann, 2019; Partanen et al., 2019; Bawden et al., 2022) found SMT and RNN-NMT to be competitive. Using word trigram windows instead of full sentences, as in Partanen et al. (2019), is also effective in our setup, although the gap towards full-sentence models is considerably lower than in their work. Finally, the pre-trained ByT5 model only outperforms vanilla Transformers on the Norwegian dataset.

3 Analyzing speaker representations in multi-dialectal NMT

Language labels are often used in multilingual neural language modeling and machine translation to inform the model of the language(s) of each sample. As a result of the training process, the models learn embeddings of these language labels, which in turn reflect the relationships between the languages (Östling and Tiedemann, 2017). Following Abe et al. (2018), we apply this idea to the Finnish and Norwegian parts of the normalization dataset introduced in the previous section. We use distinct labels for each speaker in the corpus and analyze their representations obtained by the Transformer-based normalization models.

We find that (1) the speaker label embeddings of two speakers coming from the same village are very similar, and that (2) the embeddings of all speaker labels taken together reflect the traditional dialect classifications precisely. Detailed results of this analysis are given in Kuparinen and Scherrer (2023).

4 Collecting Finnish dialect tweets

In order to extend our dialectological research to more modern and realistic types of data, we collected and annotated a dataset of dialectal Finnish tweets. We take advantage of Murreviikko ('dialect week'), a Twitter campaign initiated at the University of Eastern Finland, which promotes the use of dialects on Finnish social media. The campaign lasts for a week in October and has run for three years (2020–2022). We collected tweets containing the keyword *murreviikko* or *#murreviikko* via the Twitter API from all three years.

This collection resulted in a total of 465 tweets, 344 of which were written in a dialect of Finnish. The tweets were manually annotated by a dialectologist with the dialect region and normalized to Standard Finnish on sentence level.

In contrast to the "clean" Finnish dialect dataset used in our benchmark (Section 2), the Murreviikko data is much noisier.[2] In terms of normalization performance, the SMT model has been found to perform best, followed by the pre-trained ByT5 model. These two approaches turned out to be much more robust to noise than the vanilla Transformers.

The corpus collection process, the normalization results and the modalities of access are described in detail in Kuparinen (2023).[3]

References

Abe, Kaori, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. 2018. Multi-dialect neural machine translation and dialectometry. In *Proceedings of PACLIC*, pages 1–10, Hong Kong, China.

Bawden, Rachel, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. 2022. Automatic normalisation of early Modern French. In *Proceedings of LREC*, pages 3354–3366, Marseille, France.

Bollmann, Marcel. 2019. A large-scale comparison of historical text normalization systems. In *Proceedings of NAACL-HLT*, pages 3885–3898, Minneapolis, Minnesota, USA.

Kuparinen, Olli and Yves Scherrer. 2023. Dialect representation learning with neural dialect-to-standard normalization. In *Proceedings of VarDial*, pages 200–212, Dubrovnik, Croatia.

Kuparinen, Olli. 2023. Murreviikko - a dialectologically annotated and normalized dataset of Finnish tweets. In *Proceedings of VarDial*, pages 31–39, Dubrovnik, Croatia.

Östling, Robert and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of EACL*, pages 644–649, Valencia, Spain.

Partanen, Niko, Mika Hämäläinen, and Khalid Alnajjar. 2019. Dialect text normalization to normative standard Finnish. In *Proceedings of W-NUT*, pages 141–146, Hong Kong, China.

[2] The Murreviikko tweet authors are laypersons who do not follow any transcription conventions used by trained dialectologists. Some of the tweets also mix dialectal and standard features. Finally, the tweets contain a lot of social-media specific artifacts (emojis, hashtags, etc.) that are completely absent from the clean dataset.

[3] The public part of the corpus is available at `https://github.com/Helsinki-NLP/murreviikko`.