Learning from dialect classifiers:
Detecting dialect features across different data sources

Dana Roemling, University of Birmingham / University of Helsinki

Neural models are increasingly used for dialect classification, yet the specific linguistic features that inform their predictions are often opaque. In this talk I present different dialect classification experiments across diverse, labelled text sources - including social media and historical texts - focusing on dialect feature identification.

In these experiments, our research group employed both a neural model for dialect classification and a logistic regression baseline across several language varieties: Finnish, German and Greek dialects, national varieties of Bosnian–Croatian–Montenegrin–Serbian, and historical and modern Low German. To better understand the model's decision-making process, we use a leave-one-out (LOO) approach, which systematically removes or masks individual words or subwords to reveal their impact on prediction probabilities. This method allows us to isolate and analyse the linguistic elements, such as morphological, phonological and lexical markers, that most strongly contribute to dialect distinction.

Our results demonstrate that the neural classifier performs well on modern language varieties, although historical data proves more challenging. Moreover, the LOO approach successfully highlights key dialectal features, aligning with known linguistic distinctions and providing valuable insights into model interpretability.

This work was done as part of the CorCoDial research project at the University of Helsinki together with Yves Scherrer, Aleksandra Miletić, Janine Siewert, Erofili Psaltaki and Olli Kuparinen.