Phonetics Colloquium at UdS

on Wednesday, **4 February 2026**, 14:15

location: building C7.2 in room 5.09

Verena Blaschke

Ludwig-Maximilians-Universität München

Dialect NLP: How (and why) to process written and spoken dialect data

Abstract

NLP has improved by leaps and bounds when it comes to processing data from standardized languages with much available data, but lags behind when closely related non-standard varieties are concerned. In this talk, I will describe ways in which processing dialect data differs from processing standard-language data, and discuss some of the current challenges in dialect NLP research. For instance, the performance of text models can be hampered when ad-hoc pronunciation spellings result in infelicitous subword tokenization. Speech models, which deal with continuous inputs, appear to be more robust towards dialectal variation in content classification tasks. However, other tasks like automatic speech recognition are complicated by the fact that there is not just one way to correctly transcribe a given dialectal utterance. Furthermore, I argue that we should not only consider *how* to tackle dialectal variation in NLP, but also *why*. To this end, I will highlight perspectives of some dialect speaker communities on which language technologies should (or should not) be able to process or produce dialectal in- or output.

Bio

Verena Blaschke is currently finishing her PhD at LMU Munich where she researches NLP for non-standard dialects and other low-resource language varieties, investigating how robust language models are towards language variation (and how to make them more robust). Her research is supervised by Barbara Plank and co-supervised by Hinrich Schütze. She also completed a research internship at Apple where she worked on multilingual NLP, and she previously developed software for machine-assisted historical linguistics at the University of Tübingen.