We need to combine the social side of language & the technical side of language modelling!

Variation is the Norm: Embracing Sociolinguistics in NLP
Anne-Marie Lutgen, Alistair Plum, Verena Blaschke, Barbara Plank, Christoph Purschke

Sociolinguistic NLP Card

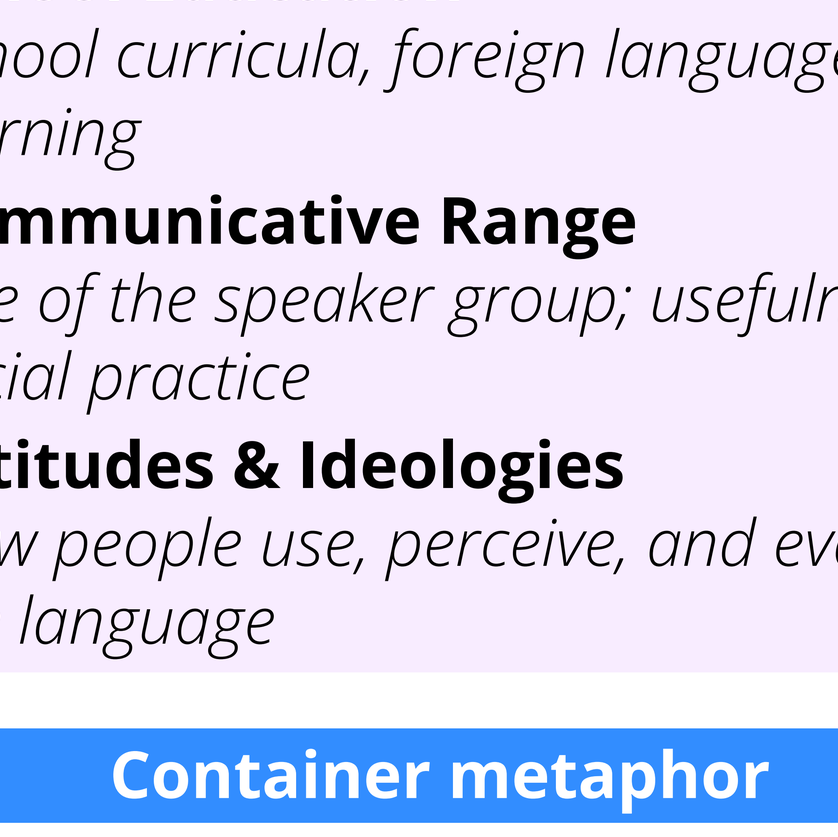
Sociolinguistic Dimensions

- **Sociolinguistic Setting**
  Socio-pragmatic context; multilingualism; language contact
- **Institutional Support**
  Political status; language policy & societal anchoring
- **Degree of Codification**
  Orthographic, grammatical & lexical standardisation
- **Domain Specificity**
  Usage situations; attributed functions
- **School Education**
  School curricula, foreign language learning
- **Communicative Range**
  Size of the speaker group; usefulness in social practice
- **Attitudes & Ideologies**
  How people use, perceive, and evaluate the language

NLP Dimensions

- **Knowing & Selecting your Data**
  Which varieties are (or should be) represented?
- **Preprocessing & Normalization**
  Normalizing text removes rich social signals
- **Tokenization**
  Sensitivity to language variation
- **Pre-training & Fine-tuning**
  Strategies depend on available data & type of downstream task
- **Evaluation of Non-Standard Varieties**
  NLG metrics are not robust to variation
- **Usage & Safety of Language Technologies**
  User preferences & safety

Container metaphor



Case study: Luxembourgish

- Sociolinguistic NLP card in paper!
- Case study: **Effect of orthographic variation on classification tasks**
- Fine-tuning models on a **mix of standard and non-standard data** creates the most robust models

Université du Luxembourg
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN