Dialect NLP
Thinking outside the box when processing non-standard and low-resource languages

Verena Blaschke
MaiNLP lab, LMU Munich
verena.blaschke@cis.lmu.de

May 15, 2025

This handout summarizes some of the considerations and tips for processing data from dialects and other low-resource languages (LRLs) mentioned in the lecture. In case you are especially interested in any of the subtopics below, I also provided pointers to related works.

1 Data availability & quality

Example paper: A Survey of Corpora for Germanic Low-Resource Languages and Dialects (Blaschke et al., NoDaLiDa 2023). github.com/mainlp/germanic-lrl-corpora

Starting points for finding datasets

- ACL Anthology (and other relevant conference proceedings), arXiv, Google Scholar, Semantic Scholar
- Google Dataset Search datasetsearch.research.google.com
- Zenodo zenodo.org
- European Language Grid live.european-language-grid.eu
- CLARIN Virtual Language Observatory vlo.clarin.eu
- OpenSLR openslr.org
- Text+ text-plus.org
- OLAC www.language-archives.org
- ORTOLANG www.ortolang.fr/market/corpora
- Hamburg Centre for Language Corpora (HZSK)
- OPUS opus.nlpl.eu

Recommendations for using dialect/LRL corpora

- Check the quality! (Any obvious issues? Language ID correct? Likely produced by actual speakers of the language?)
- Is the written representation suitable for your purposes? (E.g., phonemic transcription vs. ad-hoc spelling?)
- Are there overlaps between the pre-training, training/fine-tuning, development, and test datasets?
- Check data outside traditional NLP venues (e.g., works by linguists working on the language you're interested in)

Recommendations for creating dialect/LRL corpora

- Document the transcription guidelines / orthographies
- Share metadata like corpus size, data sources, annotation procedure; specify a license or access conditions (!)
- Used archives geared towards long-term storage (CLARIN, LRE Map, Zenodo, ...)

Other takeaways/considerations

- Which languages / language varieties are currently well-represented in NLP research and language technologies, and which aren't?
- How trustworthy and generalizable are your data?

Further pointers

- Language Varieties of Italy: Technology Challenges and Opportunities (Ramponi, TACL 2024)
- Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets (Kreutzer et al., TACL 2022)
- The State and Fate of Linguistic Diversity and Inclusion in the NLP World (Joshi et al., ACL 2020)
- Evaluating the Diversity, Equity, and Inclusion of NLP Technology: A Case Study for Indian Languages (Khanuja et al., EACL Findings 2023)

2 Modelling non-standard or low-resource language data

Example paper: Evaluating Pixel Language Models on Non-Standardized Languages (Muñoz-Ortiz et al., COLING 2025).

Be creative! Methods for dialectal and low-resource data can go in many directions beyond a straightforward "pre-train, then fine-tune" or "prompt a large language model" paradigm.

Examples for other kinds of methods and considerations

- A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios (Hedderich et al., NAACL 2021)
- Different perspectives on tokenization issues
  - Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise (Aepli/Sennrich, ACL Findings 2022) & Does Manipulating Tokenization Aid Cross-Lingual Transfer? A Study on POS Tagging for Non-Standardized Languages (Blaschke et al., VarDial 2023) – modifying the standard-language fine-tuning data
  - Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models (Ahia et al., EMNLP 2023)
- Data augmentation focused on syntactic variation: Multi-VALUE: A Framework for Cross-Dialectal English NLP (Ziems et al., ACL 2023)

Some general pointers on dialect NLP

- Natural language processing for similar languages, varieties, and dialects: A survey (Zampieri et al., Natural Language Engineering 2020)
- Quantifying the Dialect Gap and its Correlates Across Languages (Kantharuban et al., EMNLP Findings 2023)
- DIALECTBENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages (Faisal et al., ACL 2024)

3 What NLP tools should we build and why?

Example paper: What Do Dialect Speakers Want? A Survey of Attitudes Towards Language Technology for German Dialects (Blaschke et al., ACL 2024).

Takeaways

- Speakers of German dialects: Interest in language technologies processing dialectal input & speech-based technologies
- Speaker(group)s aren't monoliths!
- Sociolinguistic backgrounds are an important factor (but individual opinions exist too)
- Actively consider the wants & needs of the relevant speaker communities!

Further pointers

- NLP systems for low resource languages – hype vs. reality (Panel discussion, PML4DC @ ICLR 2023)
- Not always about you: Prioritizing community needs when developing endangered language technology (Liu et al., ACL 2022)
- What a Creole Wants, What a Creole Needs (Lent et al., LREC 2022)
- Local Languages, Third Spaces, and other High-Resource Scenarios (Bird, ACL 2022)
- Ethical Considerations for Machine Translation of Indigenous Languages: Giving a Voice to the Speakers (Mager et al., ACL 2023)
- Language Technologies as If People Mattered: Centering Communities in Language Technology Development (Markl et al., LREC-COLING 2024)
- My LLM might Mimic AAE [African American English] – But When Should It? (Sandoval et al., NAACL 2025)