A Dialectal Corpus for Ukrainian

Collection, Classification, and Standardization

Yuliia Frund Sina Ahmadi

DialRes-LREC 2026
Workshop on Dialects in NLP — A Resource Perspective
16 May 2026, Palma de Mallorca

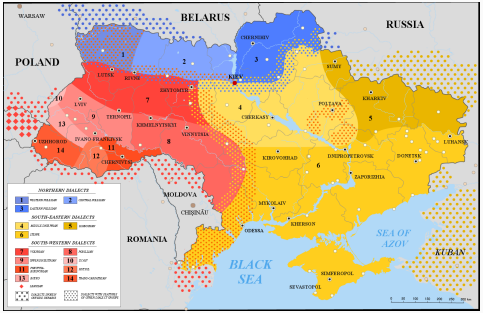
Department of Computational Linguistics
University of Zurich

Motivation

- Ukrainian dialects are largely absent from NLP
- Very few dialect resources exist, unlike for Standard Ukrainian
- Ukrainian has three major dialect groups with substantial phonetic, morphological, syntactic, and lexical variation
- **Research questions:**
  - Can current NLP tools handle dialectal Ukrainian input?
  - Can LLMs standardize dialect to Standard Ukrainian?
- **Why it matters:** inclusive language technology, dialect preservation, low-resource NLP, public services for non-standard speakers

Ukrainian dialects in three groups

- **Northern** — Volyn, Rivne, Zhytomyr, Kyiv, Chernihiv, Sumy
- **South-western** — Lviv, Ivano-Frankivsk, Zakarpattia, Ternopil, Khmelnytskyi, Vinnytsia
- **South-eastern** — Cherkasy, Poltava, Kirovohrad, Mykolaiv, Donetsk, Luhansk *(forms the basis of Standard Ukrainian)*

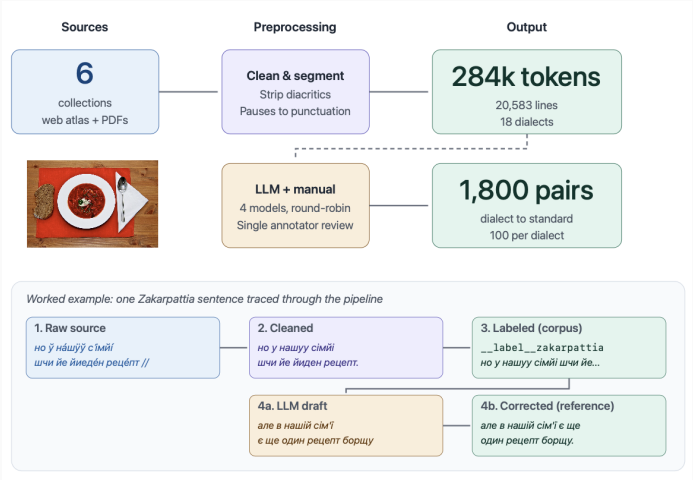


Dialect groups across Ukraine

Methodology

- **Sources:** 6 collections (1 web atlas + 5 academic PDFs) $\rightarrow$ 18 oblasts
  Largest source: speakers across 17 oblasts describing how they cook borshch
- **Preprocessing:** strip diacritics, fix encoding, segment by pauses, label by oblast
- **Reference set:** 1,800 pairs — 4 LLMs round-robin + manual correction
- **Task 1 — LID:** fastText baseline / macro / micro models, F-score
- **Task 2 — Standardization:** 4 LLMs + NLLB, BLEU + COMET vs reference

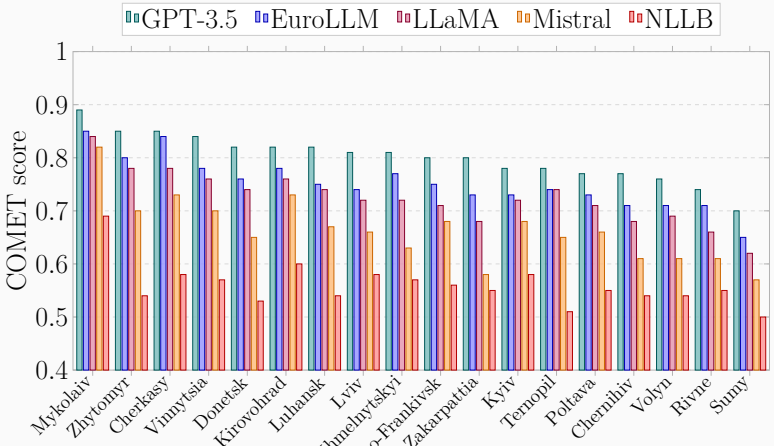
Corpus construction pipeline



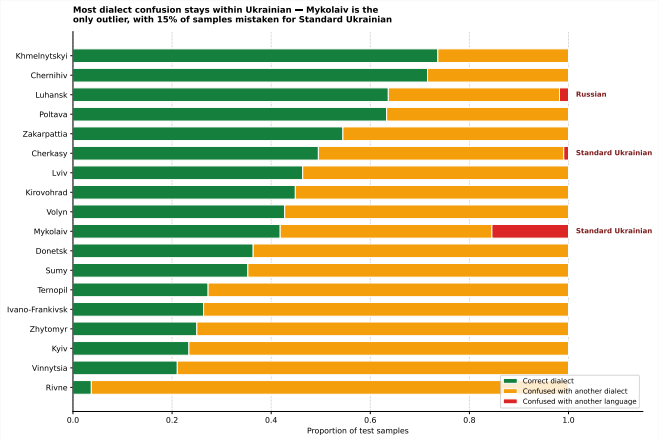
Worked example: one Zakarpattia sentence traced through every stage

Finding 1: Standardization quality varies sharply by region

Mykolaiv (0.89) standardizes cleanly, Sumy (0.70) resists every model



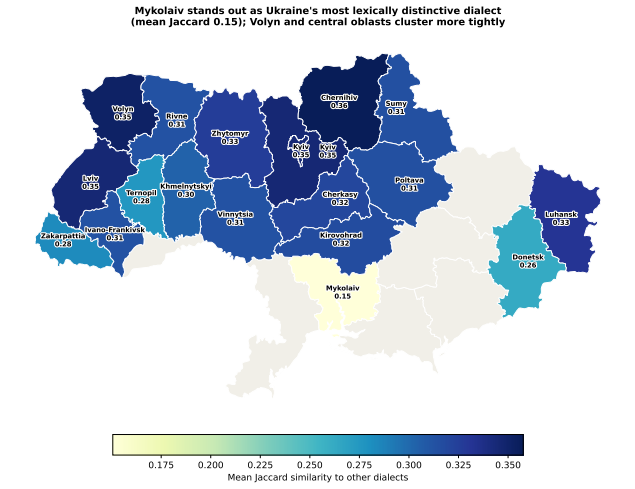
Finding 2: Misclassifications are intra-Ukrainian and lexically motivated



- Micro model: $F = 0.58$
- 95% of errors fall on *other Ukrainian oblasts*, not other Slavic languages
- **Mykolaiv** is the only outlier: 15% of samples mistaken for Standard Ukrainian

Finding 3: Lexical centrality across Ukraine

Mean Jaccard similarity per dialect — a measure of lexical centrality



A note on language identification

- **Baseline fastText:** F = 0.75
  Dialect text frequently misclassified as Russian, Belarusian, Bulgarian, Serbian, Macedonian, or Tatar
- **Macro model** (dialect data added to training): F = 0.99
- Adding even modest dialect data is enough to close the LID gap

Conclusion

- **First comprehensive dialectal corpus for Ukrainian**
  284k tokens, 18 oblasts, 1,800 parallel pairs
- Standard LID tools systematically misclassify dialect input as related Slavic languages
- Dialect-aware training closes the LID gap: F-score 0.75 $\rightarrow$ 0.99
- Inter-dialect confusion is lexically driven, not phonetically driven
- LLMs are viable zero-shot dialect-to-standard standardizers (GPT-3.5 COMET 0.80)
- Inclusive language technology for Ukrainian is achievable with modest annotation effort

Resources

All resources released openly:

github.com/yfrund/ukrainian_dialects

- Regional dialect corpus (284k tokens, 18 oblasts)
- Labelled datasets for fastText
- Model outputs and parallel reference translations

Thank you!

yuliia.frund@uzh.ch sina.ahmadi@uzh.ch