University of Zurich[UZH]

Department of Computational Linguistics

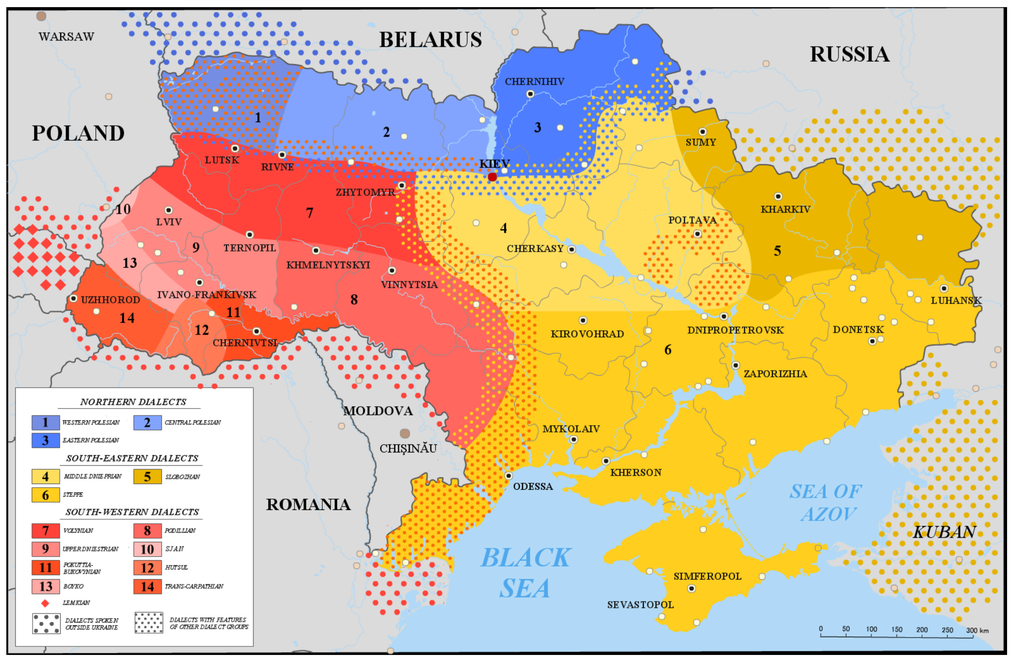
LREC 2026 Palma

A Dialectal Corpus for Ukrainian

Yuliia Frund **Sina Ahmadi**
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 variation
- **Can current NLP tools handle dialectal Ukrainian input, and can LLMs standardize it?**
- **This matters to inclusive language technology, dialect preservation, low-resource NLP, public services for non-standard speakers**



Methodology

- Sources: 6 collections (1 web atlas + 5 academic PDFs) → 18 oblasts; the largest source elicits speech via a shared prompt
- 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

Sources

6
collections
web atlas + PDFs

Preprocessing

Clean & segment
Strip diacritics
Pauses to punctuation

LLM + manual
4 models, round-robin
Single annotator review

Output

284k tokens
20,583 lines
18 dialects

1,800 pairs
dialect to standard
100 per dialect

Worked example: one Zakarpattia sentence traced through the pipeline

1. Raw source
но ў на́шўў с і́мйі́
шчи йе йиеде́н рецепт //

2. Cleaned
но у нашуу сімйі
шчи йе йиден рецепт.

3. Labeled (corpus)
`__label__zakarpattia`
но у нашуу сімйі шчи йе...

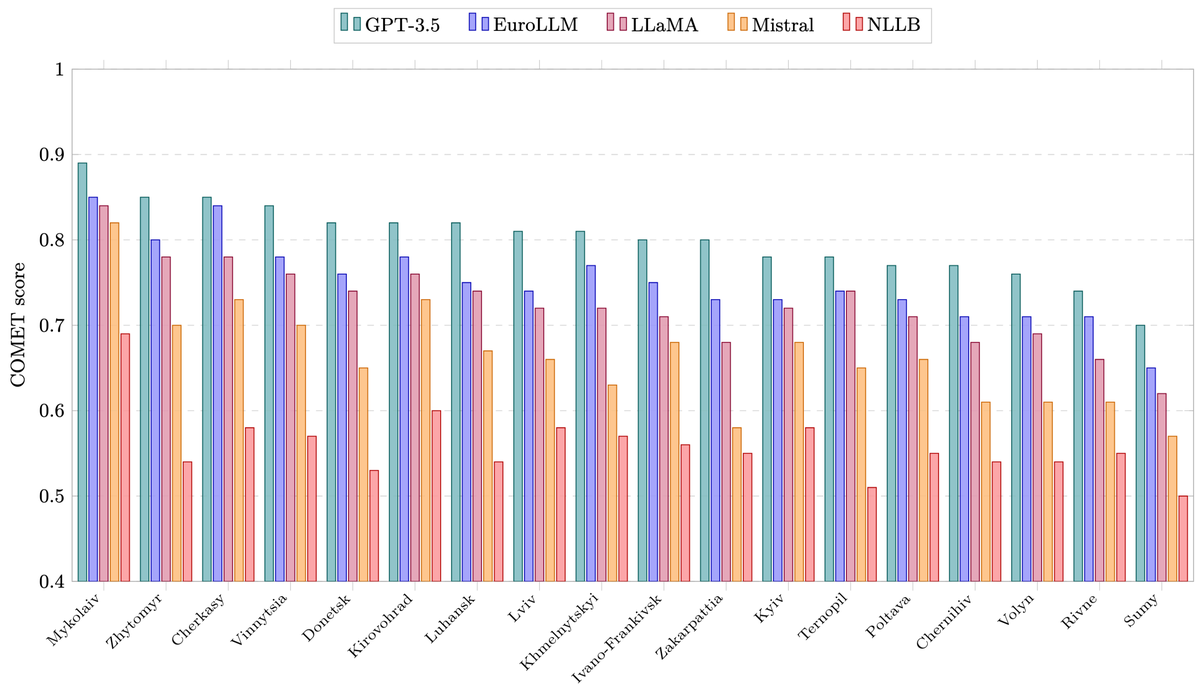
4a. LLM draft
але в нашій сім'ї
є ще один рецепт борщу

4b. Corrected (reference)
але в нашій сім'ї є ще
один рецепт борщу.

Key Findings

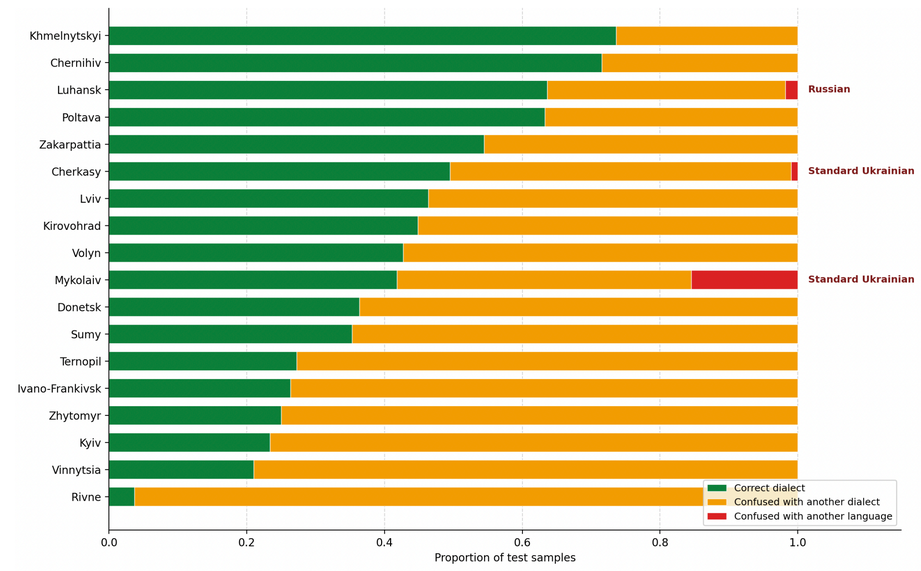
Finding 1: Standardization quality varies sharply by region

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



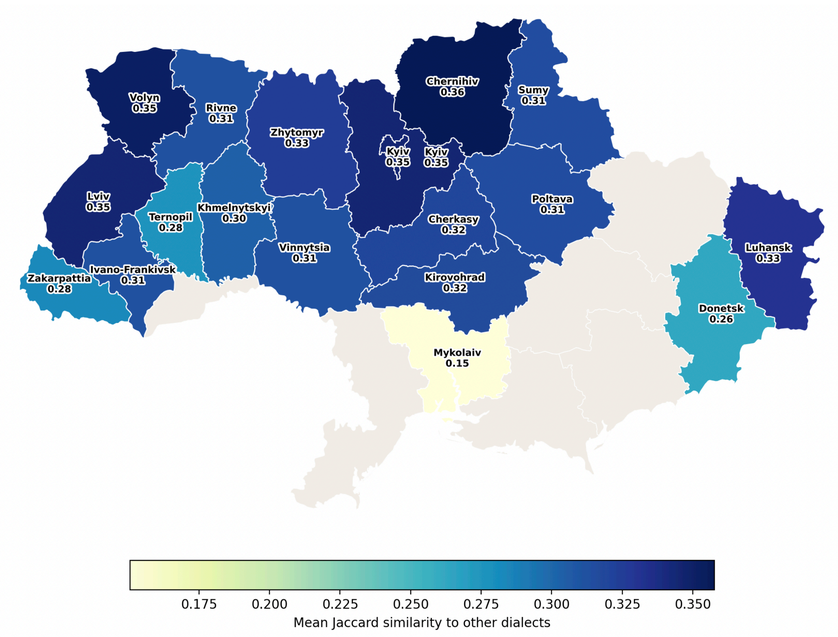
Finding 2: Dialect classification

Misclassifications are intra-Ukrainian and lexically motivated



Finding 3: Linguistic distinctiveness

Mean Jaccard similarity per dialect, a measure of lexical



Conclusion

- First comprehensive dialectal corpus for Ukrainian
- Standard LID tools systematically misclassify dialect input as related Slavic languages
- Inter-dialect confusion is lexically driven
- LLMs are viable zero-shot dialect-to-standard standardizers
- Dialect-aware training closes the LID gap: F-score 0.75 → 0.99
- Inclusive language technology for Ukrainian is achievable with modest annotation effort