A Dialectal Corpus for Ukrainian: Collection, Classification, and Standardization

Yuliia Frund **Sina Ahmadi**
Department of Computational Linguistics
University of Zurich
{yuliia.frund,sina.ahmadi}@uzh.ch

Abstract

Ukrainian dialects remain largely excluded from the digital linguistic landscape despite their active everyday use. We present a regional dialect corpus covering 18 administrative regions of Ukraine, compiled from digitized fieldwork collections and an online dialect atlas. The corpus comprises over 284,000 tokens of dialect text, annotated by region and partially accompanied by manually standardized translations. Using these resources, we investigate language identification and dialect-to-standard standardization. Baseline language identification yields an F-score of 0.75, rising to 0.99 with dialect-inclusive training. Dialect classification reaches 0.58, with confusion patterns reflecting known regional boundaries. For standardization, the best-performing LLM achieves a COMET score of 0.80, though BLEU scores remain low (0.21–0.23) across all models. We release the corpus, labelled datasets, model outputs, and reference translations to support future work on inclusive language technologies for non-standard varieties. Our resource is openly available at https://github.com/yfrund/ukrainian_dialects.

Keywords: Ukrainian dialects, dialect corpus, language identification, dialect classification, low-resource NLP

1. Introduction

Dialects are actively used in everyday communication by millions of speakers worldwide, yet modern language technologies remain predominantly trained on standard language varieties. This creates a growing disconnect: as natural language processing (NLP) systems become more embedded in public services, healthcare, education, and digital communication, their inability to handle non-standard input excludes a significant portion of speakers. Recognizing this gap, recent years have seen increasing efforts to develop dialectal resources and tools across languages such as Arabic (Keleg et al., 2025), Greek (Chatzikyriakidis et al., 2023), Italian (Ramponi, 2024), and Germanic varieties (Borin, 2025), reflecting a broader shift toward inclusive language and speech technologies (Joshi et al., 2025).

Ukrainian is a language where this gap is particularly evident. While resources for Standard Ukrainian have expanded considerably, including large annotated corpora (Shvedova et al., 2017–2022), parliamentary proceedings (Kopp et al., 2023), and treebanks, its dialects remain largely absent from the digital landscape. Ukrainian encompasses three major dialect groups spanning the country's administrative regions (Del Gaudio, 2017), with substantial variation at the phonetic, morphological, syntactic, and lexical levels. For instance, Northern dialects exhibit distinctive diphthongization patterns and loss of consonant palatalization, South-western varieties preserve archaic morphological forms, and the South-eastern group, while closest to the standard variety, still displays notable phonetic and lexical divergence. This diversity is both a challenge for NLP systems and a valuable resource for building more robust tools, yet the lack of digitized, annotated dialect data has limited progress in this direction.

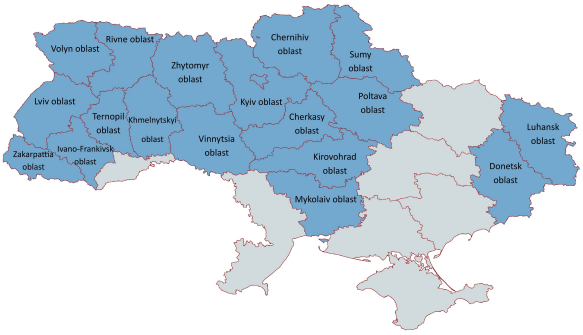

Figure 1: Regions in Ukraine for which our corpus contains dialectal data (highlighted in blue).

In this paper, we address this gap by collecting and curating dialectal data from digitized fieldwork collections and an online dialect atlas, covering 18 administrative regions of Ukraine with over 284,000 tokens. Using these resources, we investigate how current NLP tools handle dialectal input and whether they can be adapted to better accommodate non-standard variation. Our contributions are as follows:

- A dialectal corpus of Ukrainian annotated by region of origin;
- A dialect-to-standard reference set for machine translation, created through LLM-generated outputs with manual correction;
- Baseline experiments on language identification and dialect standardization.

2. Related Work

Dialectal NLP. The challenges of processing dialectal text have received growing attention in the NLP community. Joshi et al. (2025) provide a comprehensive survey of methods applied to dialects, identifying data scarcity, inconsistent orthography, and evaluation gaps as persistent bottlenecks. On the evaluation side, benchmarks such as CoDET (Alam et al., 2024) and DialectBench (Faisal et al., 2024) have been introduced to enable systematic assessment of NLP systems on dialectal data. Notably, Alam et al. (2024) report consistent performance drops in neural machine translation systems when processing dialectal input, further illustrating the impact of dialectal underrepresentation in training data. Language-specific efforts have addressed this gap through dedicated resource creation and evaluation for English (Ziems et al., 2023), Arabic (Keleg et al., 2025), Greek (Chatzikyriakidis et al., 2023), Italian (Ramponi, 2024), Kurdish (Ahmadi et al., 2024), and Germanic varieties (Borin, 2025; Plüss et al., 2023). A recurring finding across these efforts is that explicitly dialectal training data are often required to close the performance gap, as simply increasing data volume does not consistently help (Kantharuban et al., 2023).

Dialect normalization and standardization. A key challenge in dialect processing is that NLP tools are typically designed around standard orthographic conventions, making them ill-suited for dialectal text whose written forms do not conform to these norms. Prior work has explored rule-based (Shvedova et al., 2022), semi-automatic (Samardžić et al., 2016), and neural approaches (Kuparinen et al., 2023) to map non-standard text to standard forms, with transformer-based systems generally achieving the most robust results. However, these methods typically address only orthographic form, while dialects also differ in vocabulary and syntax. This raises the question of whether LLMs can standardize both form and vocabulary, bypassing the extensive annotation traditionally required, a possibility we investigate for Ukrainian dialects.

Resources for Ukrainian. Digital resources for Standard Ukrainian have grown in recent years. The GRAC corpus (Shvedova et al., 2017–2022) provides a large annotated collection of standard texts with advanced filtering capabilities. Kopp et al. (2023) introduced an annotated corpus of Ukrainian parliamentary proceedings. Community-driven projects such as Lang-UK[1] and the Universal Dependencies treebank for Ukrainian[2] further contribute to the standard-language resource ecosystem. However, these resources focus exclusively on Standard Ukrainian or on a specific dialect, as for Hutsul (Kyslyi et al., 2025). Dialectal text remains underrepresented, and to the best of our knowledge, no publicly available parallel corpus for Ukrainian dialects existed prior to this work.

3. Ukrainian Dialects

Ukrainian dialects are traditionally grouped into three major categories: Northern, South-western, and South-eastern (Del Gaudio, 2017). The Northern group is spoken across the Volyn, Rivne, Zhytomyr, Kyiv, Chernihiv, and Sumy regions. The South-western group spans from the western border to the west of Kyiv, Cherkasy, Kirovohrad, and Mykolaiv regions. The South-eastern group covers the remaining regions, from Kyiv and Chernihiv in the north to Odesa and Crimea in the south. Our corpus covers 18 of Ukraine's administrative regions across all three groups (see Figure 1).

The **Northern group** is phonetically the most distinct. Notable features include the development of etymological [o] into diphthongs in closed syllables, the pronunciation of unstressed [o] closer to [a], and the loss of palatalization in certain consonants; for example, standard буряк [burʲak] ('beetroot') becomes бурак [burak]. The group also exhibits morphosyntactic particularities, such as non-standard future-tense constructions (му ходити alongside standard буду ходити 'I will be going') and special numeral agreement patterns.

The **South-western group** displays significant local variation and preserves archaic forms. The Zakarpattia dialect retains што ('what'), and the reflexive particle ся functions as a clitic rather than the verbal suffix found in the standard variety. Other features include variation in demonstrative pronouns (e.g., сей/цей for 'this'), long and short possessive pronoun forms (моему/мойому 'mine', dative), and consonant hardening in certain noun declensions.

The **South-eastern group**, whose core lies in the Cherkasy and Poltava regions, forms the basis of modern Standard Ukrainian and is the most homogeneous of the three. Nevertheless, it displays features such as the approximation of [o] towards [u], realization of /f/ as [xv] (e.g., хвабрика instead of фабрика 'factory'), and parallel infinitive forms (робити/робить 'to do') (Del Gaudio, 2017).

This variation poses considerable challenges for language technologies trained exclusively on standard text, and motivates the broad geographical coverage of our corpus.

[1] https://lang.org.ua/en

[2] https://universaldependencies.org/uk

Name	Sources	Metadata
Dialect Map	Dialect terms added by philology students from various linguistic resources.	Part of speech, source, synonyms, use in context, set phrases
Hibeba and Lesnova (2019)	Recorded by field linguists, school teachers, students, language enthusiasts, and researchers.	Oblast, place of residence, speaker's name and birth year, date of recording
Martynova (2012)	Recorded by the compiler.	Place of residence, speaker's name, birth year, education
Martynova et al. (2013)	Recorded by the compilers and dialectology students.	Place of residence, speaker's name, birth year, education
Lesnova (2013)	Recorded by the compiler.	Place of residence, speaker's name and birth year
Kovalenko and Kovalenko (2019)	Recorded by the compilers.	Place of residence, speaker's name, birth year and education

Table 1: Overview of the resources used, along with their sources and additional metadata.

4. Methodology

4.1. Data Collection

Due to the lack of large-scale dialect corpora for Ukrainian, especially those annotated by region, we compiled a dataset from several publicly available resources. While dialect text can also be found on social media or in local publications, the former is noisy due to slang and inconsistent spelling, and the latter tends to focus on isolated word lists. We therefore limited our sources to online platforms with contextualized examples and digitized academic collections of connected dialect text:

- *Dialect Map*,[3] an interactive atlas providing usage examples across seven regions (Lviv, Ivano-Frankivsk, Volyn, Zakarpattia, Mykolaiv, Rivne, and Vinnytsia);
- Hibeba and Lesnova (2019), dialectal speech transcriptions covering 17 oblasts in which speakers describe the preparation of borshch—a culturally significant dish chosen to elicit naturalistic, topic-controlled speech across regions;
- Martynova (2012) and Martynova et al. (2013), dialect texts from Poltava and Cherkasy regions respectively;
- Lesnova (2013) and Kovalenko and Kovalenko (2019), complementary texts from Luhansk and Khmelnytskyi oblasts.

Table 1 provides an overview of these resources and the metadata they supply.

The raw data required substantial preprocessing. The web-based data were parsed from HTML, while the remaining collections were extracted from digitized PDFs. During filtering, only dialect content was retained; speaker metadata and interviewer questions in the standard variety were excluded. Since the sources represent dialect speech using phonetic transcription, we removed diacritics and phonetic symbols, converting non-standard characters to their closest standard equivalents (e.g., non-syllabic [ў] and [ĭ] to у and і). The raw data are structured into phonetic phrases delimited by pauses: short pauses were mapped to commas and long pauses to sentence boundaries as follows:

Raw Text
но ў на́шўў с'імйі́ шчи йе йиеде́н рецепт боуршчу́ // його́ варит мо́а ма́ма лиеше́ на с'а́ткы //
Cleaned Data
1. но у нашуу сімйі шчи йе йиден рецепт боршчу 2. його варит моа мама лише на саткы

Table 2: Raw vs. cleaned data ("But in our family there is one more recipe of borshch. My mother cooks it only for celebrations."). Non-standard tokens are underlined.

Another challenge was character encoding inconsistency in PDF-extracted text, where visually identical Latin and Cyrillic characters (e.g., *i*, *e*, *a*) had different Unicode values, affecting both statistics and model training. All Latin characters were mapped to their Cyrillic equivalents. For the purposes of this work, a dialect is defined as corresponding to an administrative region (oblast), yielding 18 classes. While this does not capture the full variation within each region, it provides a practical framework that balances linguistic relevance with

[3] https://dialectmap.org/en

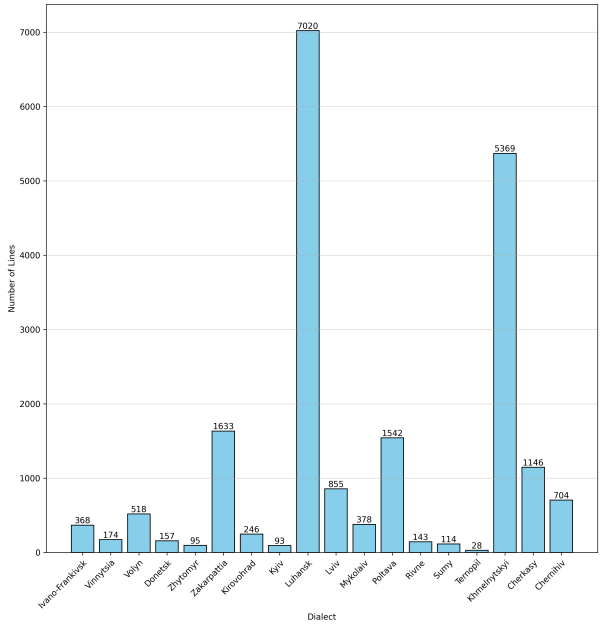

Figure 2: Class distribution of dialect data by oblast. Luhansk is the largest class and Ternopil the smallest.

Metric	Value
Number of dialects	18
Number of lines	20,583
Number of tokens	284,045
Number of unique tokens	87,510
Avg. tokens per line	13
Avg. unique tokens per line	4
Avg. token length	5

Table 3: General corpus statistics.

scalability. Table 3 presents overall corpus statistics and Figure 2 shows the class distribution.

4.2. Reference Translations

To evaluate how well models can translate dialectal text to other languages and even Standard Ukrainian, we need parallel data pairing dialect input with its standard equivalent. Since no such resource exists for Ukrainian dialects, we create one semi-automatically by generating translations with multiple LLMs and then manually correcting them.

For each of the 18 dialect classes, 100 sentences were sampled (with a minimum length of 5 tokens to ensure sufficient context); for smaller classes, as many sentences as possible were taken and the remainder supplemented from larger classes, yielding 1,800 sentence pairs in total. Four models were used to generate initial translations: three quantized models (EuroLLM-9B-Instruct (Martins et al., 2025), Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Mistral-7B-Instruct-v0.1 (Jiang et al., 2023)), along with GPT-3.5-turbo. All models received the prompt: *"Translate the text into Standard Ukrainian. Do not provide any commentary."* with a temperature of 0 to ensure reproducibility.

To avoid biasing the reference set toward any single model's outputs, translations were sampled in a round-robin manner: for each source sentence, the reference was taken from a different model in turn. The sampled translations were then manually reviewed and corrected with close attention to the source material. Since the data represent transcribed speech, sentence structure was preserved to avoid favouring any particular model's rephrasing style, and dialect-specific terms were retained when their status as non-standard was uncertain. Additionally, a set of translations was generated using NLLB (Team, 2024), but these were excluded from the reference set and evaluated separately.

4.3. Experimental Setup

We conduct experiments on two tasks: language identification and dialect-to-standard text standardization.

Language identification. We use fastText (Joulin et al., 2017, 2016) for language identification, as it supports 176 languages and can be trained efficiently on a standard CPU. Our baseline is the pretrained `lid.176.bin` model. We evaluate it by mapping all dialect text to a single Ukrainian label to assess how well dialect input is recognized as Ukrainian.

We then train two additional models. The *macro model* is trained on dialect data labelled as Ukrainian alongside samples from languages that caused frequent misclassifications in the baseline (Bulgarian, Macedonian, Serbian, Tatar, Russian, and Belarusian), sourced from the Leipzig Corpora Collection (Goldhahn et al., 2012). The *micro model* is trained on the same data but with dialect-specific labels (one per oblast), enabling inter-dialect classification. Since fastText does not support fine-tuning, pretrained word vectors were used to initialize representations for the macro model; the micro model was trained from scratch. Given the uneven class distribution, downsampling and upsampling were applied to balance classes at 500 samples each, with an 80/20 train/test split. Upsampling was performed after splitting to prevent data leakage. All models are evaluated using macro F-score. Table 4 reports the final hyperparameters.

Parameter	Macro Model	Micro Model
Vector dimensions	16	300
Learning rate	1.0	0.8
Epochs	60	80
Word n-grams	1	2
Min char n-gram	3	1
Max char n-gram	6	5
Pretrained vectors	Yes	No

Table 4: Final hyperparameter settings for fastText model training.

Dialect-to-standard standardization. We evaluate five systems on the task of mapping dialect text to Standard Ukrainian: three locally-run quantized LLMs (EuroLLM-9B-Instruct, Meta-Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.1), GPT-3.5-turbo, and the NLLB machine translation model (Team, 2024). The LLMs were quantized using GGUF formats (Q3_K_S, Q4_K_M, and Q4_K_S respectively) to enable inference on consumer-grade hardware (Lang et al., 2024). Translation quality is assessed using two complementary metrics: BLEU (Papineni et al., 2002), which measures surface-level n-gram overlap, and COMET (Rei et al., 2020), which evaluates semantic adequacy. Each model receives a dialectal sentence as input and is prompted to produce its Standard Ukrainian equivalent. The generated output is then compared against the manually corrected reference translations using BLEU and COMET. We additionally test a more explicit prompt—*"Translate the text into Standard Ukrainian. Use standard spelling and standard vocabulary. Do not provide any commentary."*—to investigate prompt sensitivity. Scores are reported per dialect and in aggregate.

5. Experiments

5.1. Language Identification

The baseline fastText model achieves an overall macro F-score of 0.75 when classifying dialect text as Ukrainian. Performance varies by region, ranging from 0.56 (Sumy) to 0.99 (Mykolaiv), with per-dialect scores reported in Table 5. Misclassifications stem primarily from confusion with other East and South Slavic languages—particularly Russian, Belarusian, Bulgarian, Serbian, Macedonian, and Tatar—as shown in Figure 3.

The macro model, trained with dialect data labelled as Ukrainian alongside confusable languages, raises the F-score to 0.99, correctly identifying virtually all dialect input as Ukrainian (Figure 4). On the other hand, the micro model, trained with dialect-specific labels, achieves an F-score of 0.58. Figure 5 reveals several clusters of inter-dialect confusion. A western cluster groups Zakarpattia, Lviv, and Volyn, while a central-eastern cluster links Cherkasy, Poltava, Kirovohrad, and Chernihiv. Some dialect pairs show symmetric confusion (e.g., Poltava–Cherkasy: 29 and 24 misclassifications respectively; Luhansk–Khmelnytskyi: 15 and 10), while others are asymmetric (e.g., Mykolaiv misclassified as Lviv 21 times, but Lviv as Mykolaiv only 5 times).

To investigate these confusion patterns, we computed the top-10 most frequent bigrams through five-grams for each dialect and measured pairwise Jaccard similarity (Jaccard, 1901). As Figure 6 shows, confusion patterns largely align with n-gram similarity: Cherkasy and Poltava share a similarity of 0.86, Lviv and Volyn 0.57, and Zakarpattia and Lviv 0.48. The unexpected Luhansk–

Dialect	F	Dialect	F
Cherkasy	0.86	Mykolaiv	0.99
Chernihiv	0.63	Poltava	0.89
Donetsk	0.67	Rivne	0.64
Ivano-Frankivsk	0.79	Sumy	0.56
Khmelnytskyi	0.86	Ternopil	0.71
Kirovohrad	0.75	Vinnytsia	0.79
Kyiv	0.70	Volyn	0.65
Lviv	0.73	Zakarpattia	0.65
Luhansk	0.78	Zhytomyr	0.81

Table 5: Baseline F-scores: performance varies widely across dialects, from 0.56 (Sumy) to 0.99 (Mykolaiv).

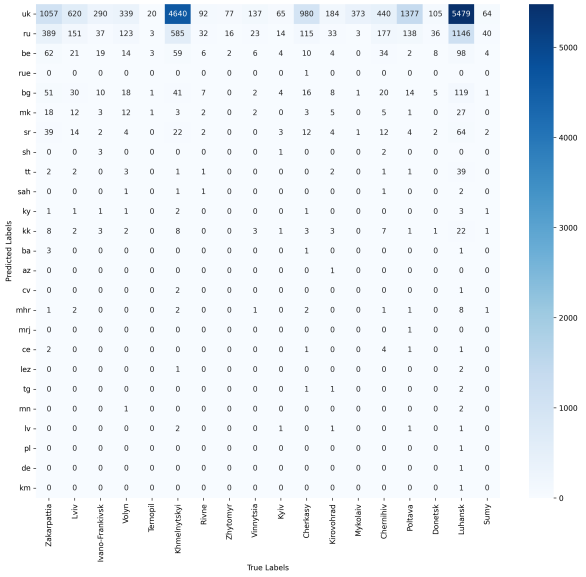

Figure 3: Baseline confusion matrix: Ukrainian dialect text is frequently misclassified as related Slavic languages. Dialects ordered West to East.

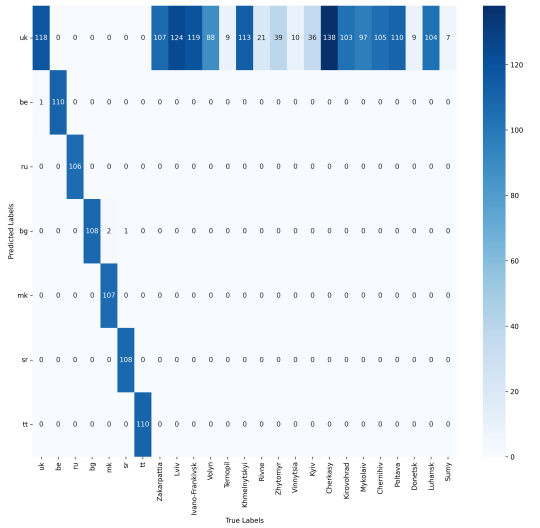

Figure 4: Macro model confusion matrix: including dialect data in training nearly eliminates cross-language misclassification (F=0.99).

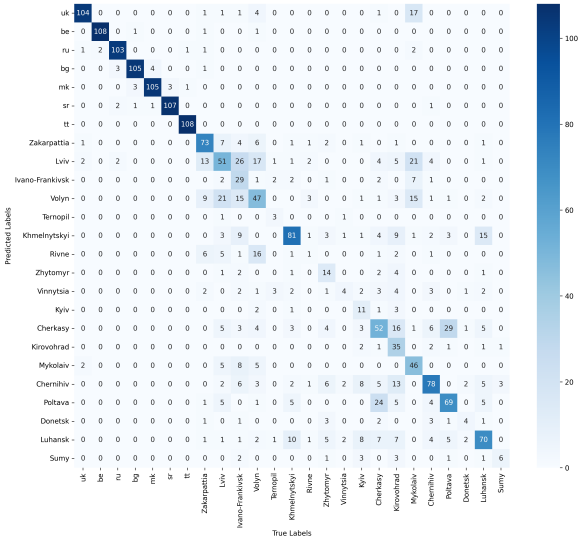

Figure 5: Micro model confusion matrix: inter-dialect confusion clusters emerge along geographic lines (F=0.58).

Khmelnytskyi confusion is also explained by their having the highest pairwise similarity among all non-neighbouring dialect pairs. However, high similarity does not always produce frequent misclassifications: Chernihiv and Sumy share a score of 0.54 yet yield only three misclassifications, suggesting the model can learn to distinguish similar dialects when sufficient training data are available.

5.2. Dialect-to-Standard Standardization

Table 6 presents BLEU scores across all models and dialects. Overall performance is low, with

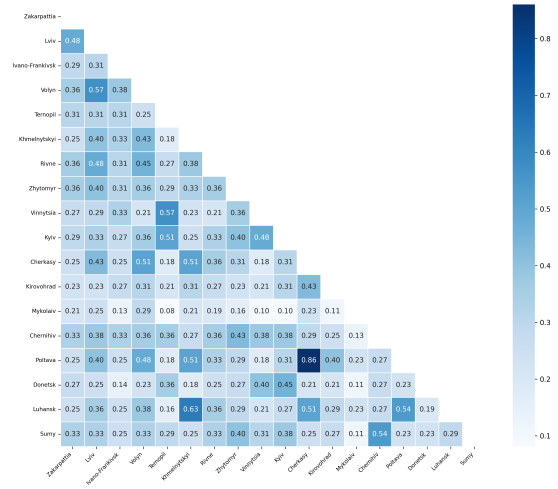

Figure 6: Jaccard similarity between dialect pairs based on n-gram overlap: geographically adjacent varieties tend to share more in common (e.g., Cherkasy–Poltava: 0.86, Lviv–Volyn: 0.57).

scores ranging from 0.21 (LLaMA, NLLB) to 0.23 (EuroLLM). All models consistently score lowest on Donetsk (0.07) and highest on Kyiv (0.36–0.45). The four LLMs perform comparably, while NLLB trails slightly. When it comes to COMET (Table 7), GPT-3.5 leads with 0.80, followed by EuroLLM (0.76), LLaMA (0.73), Mistral (0.66), and NLLB (0.56). All models perform worst on Sumy and best on Mykolaiv. The divergence between BLEU and COMET rankings indicates that while surface-level overlap with the reference remains limited, models, particularly GPT-3.5 and EuroLLM, preserve semantic content well.

Manual inspection reveals qualitative differences across models. GPT-3.5 produces the most consistent output in terms of capitalization and punctuation. EuroLLM occasionally enters infinite generation loops and sometimes conflates the system prompt with the output. Mistral shows issues with diacritics and inconsistent script, sometimes producing Latin-script transliterations instead of Cyrillic. LLaMA occasionally refuses to translate, misinterpreting dialectal input as incomplete or misleading. Note that NLLB translations were generated after the reference set was finalized and are therefore not included in it, placing NLLB at a disadvantage in direct comparison.

A more explicit prompt (*"Use standard spelling and standard vocabulary."*) produced only minor score variations, with the overall model ranking unchanged. The most notable shifts were observed for Vinnytsia (BLEU: 0.32→0.25) and Rivne (BLEU: 0.36→0.23).

Dialect	EuroLLM	GPT-3.5	LLaMa	Mistral	NLLB
Ivano-Frankivsk	0.19	0.17	0.17	0.18	0.16
Vinnytsia	0.33	0.32	0.32	0.34	0.30
Volyn	0.13	0.13	0.13	0.13	0.06
Donetsk	**0.07**	**0.07**	**0.07**	**0.07**	**0.06**
Zhytomyr	0.22	0.18	0.21	0.23	0.18
Zakarpattia	0.21	0.21	0.15	0.21	0.19
Kyiv	**0.45**	**0.41**	**0.39**	**0.39**	**0.36**
Kirovohrad	0.29	0.27	0.25	0.28	0.25
Luhansk	0.10	0.10	0.10	0.10	0.09
Lviv	0.38	0.36	0.31	0.36	0.31
Mykolaiv	0.22	0.22	0.21	0.22	0.22
Poltava	0.27	0.26	0.25	0.27	0.23
Rivne	0.34	0.36	0.32	0.34	0.32
Sumy	0.25	0.26	0.26	0.26	0.12
Ternopil	0.14	0.13	0.13	0.14	0.13
Khmelnytskyi	0.08	0.12	0.09	0.11	0.11
Cherkasy	0.15	0.15	0.14	0.14	0.14
Chernihiv	0.32	0.32	0.30	0.34	0.28
Total	**0.23**	**0.22**	**0.21**	**0.22**	**0.21**

Table 6: Performance of Ukrainian dialect-to-standard translation based on BLEU.

Dialect	EuroLLM	GPT-3.5	LLaMa	Mistral	NLLB
Ivano-Frankivsk	0.75	0.80	0.71	0.68	0.56
Vinnytsia	0.78	0.84	0.76	0.70	0.57
Volyn	0.71	0.76	0.69	0.61	0.54
Donetsk	0.76	0.82	0.74	0.65	0.53
Zhytomyr	0.80	0.85	0.78	0.70	0.54
Zakarpattia	0.73	0.80	0.68	0.58	0.55
Kyiv	0.73	0.78	0.72	0.68	0.58
Kirovohrad	0.78	0.82	0.76	0.73	0.60
Luhansk	0.75	0.82	0.74	0.67	0.54
Lviv	0.74	0.81	0.72	0.66	0.58
Mykolaiv	**0.85**	**0.89**	**0.84**	**0.82**	**0.69**
Poltava	0.73	0.77	0.71	0.66	0.55
Rivne	0.71	0.74	0.66	0.61	0.55
Sumy	**0.65**	**0.70**	**0.62**	**0.57**	**0.50**
Ternopil	0.74	0.78	0.74	0.65	0.51
Khmelnytskyi	0.77	0.81	0.72	0.63	0.57
Cherkasy	0.84	0.85	0.78	0.73	0.58
Chernihiv	0.71	0.77	0.68	0.61	0.54
Total	**0.76**	**0.80**	**0.73**	**0.66**	**0.56**

Table 7: Performance of Ukrainian dialect-to-standard translation based on COMET.

6. Conclusion

We presented a dialectal corpus for Ukrainian covering 18 administrative regions, along with a dialect-to-standard parallel reference set built by combining LLM outputs with manual correction. We used these resources to run experiments on language identification and dialect-to-standard standardization. The results show that standard NLP tools struggle with dialectal input, a baseline fastText model scores only 0.75 on dialect text, but that adding even a modest amount of dialect data to training brings this up to 0.99. When we push the task further to distinguishing individual dialects, performance drops to 0.58, with confusion patterns that track geographic proximity and surface-level similarity between varieties. On the standardization side, GPT-3.5 achieves a COMET score of 0.80, which suggests that the semantic content is largely preserved, though low BLEU scores (0.21–0.23) across all models point to limited surface-level overlap with the reference. From a practical standpoint, running dialectal text through an LLM as a pre-processing step can make it more compatible with downstream tools.

Future Work Looking ahead, the most immediate need is more data: expanding the corpus to additional regions and domains would help address class imbalance and improve coverage. On the modelling side, transformer-based classifiers and few-shot or fine-tuned LLMs are natural next steps for both identification and standardization. Extending this line of work to speech would also be valuable, given that dialects are first and foremost spoken varieties.

Limitations Our reference translations were constructed by sampling LLM outputs in round-robin fashion and manually correcting them. While this ensures balanced model representation, the evaluation metrics, particularly BLEU, may still be influenced by shared output patterns across the generating models, as surface-level phrasing choices could carry over even after manual editing. A more principled selection strategy could mitigate this, for instance by leveraging quality signals from model outputs rather than fixed rotation. For scaling to a larger parallel corpus, active learning and LLM-as-a-judge approaches could help prioritize which sentences require human correction, reducing annotation effort while improving reference quality. Additionally, all corrections were performed by a single annotator without involvement of dialectologists, introducing potential subjectivity. Furthermore, the dialect boundaries used in our classification largely coincide with administrative borders, which do not necessarily reflect natural linguistic boundaries. Finally, the language identification models were trained on a small dataset of approximately 500 lines per class, with some minority classes containing as few as 28 lines, which limits the generalizability of classification results.

Ethics Statement All data were sourced from publicly available academic publications and an open online platform. Speaker metadata present in the original sources (names, birth years) is not redistributed; only textual content is used. We note that dialect classification could in principle be misused for regional profiling and encourage responsible use of the released resources.

Acknowledgments

Sina Ahmadi gratefully thanks the support of the UZH Grant (reference number 269093).

7. Bibliographical References

S. Ahmadi, D. Jaff, M. M. I. Alam, and A. Anastasopoulos. 2024. Language and speech technology for Central Kurdish varieties. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10034–10045.

M. M. I. Alam, S. Ahmadi, and A. Anastasopoulos. 2024. CODET: A benchmark for contrastive dialectal evaluation of machine translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

L. Borin. 2025. Corpus and computational linguistic approaches to Germanic languages. In *Oxford Research Encyclopedia of Linguistics*.

S. Chatzikyriakidis, C. Qwaider, I. Kolokousis, C. Koula, D. Papadakis, and E. Sakellariou. 2023. Grdd: A dataset for Greek dialectal NLP. *arXiv preprint arXiv:2308.00802*.

S. Del Gaudio. 2017. *An Introduction to Ukrainian Dialectology*. Peter Lang Verlag, Berlin, Germany.

F. Faisal, O. Ahia, A. Srivastava, K. Ahuja, D. Chiang, Y. Tsvetkov, and A. Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

D. Goldhahn, T. Eckart, and U. Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

N. Hibeba and V. Lesnova, editors. 2019. *Treasures of Ukrainian Dialects: Texts about Borshch [Скарби українських говорів: тексти про борщ]*. Dialektolohichna skrynia. Instytut Ukrainoznavstva im. I. Krypiakevycha NAN Ukrainy, Lviv.

P. Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed. 2023. Mistral 7b.

A. Joshi, R. Dabre, D. Kanojia, Z. Li, H. Zhan, G. Haffari, and D. Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57(6):1–37.

A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

A. Kantharuban, I. Vulić, and A. Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

A. Keleg, S. Goldwater, and W. Magdy. 2025. Revisiting common assumptions about Arabic dialects in NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3327.

M. Kopp, A. Kryvenko, and A. Rii. 2023. Ukrainian parliamentary corpus ParlaMint-UA 4.0.1. Slovenian language resource repository CLARIN.SI.

N. Kovalenko and B. Kovalenko, editors. 2019. *Volhynian Dialects of Khmelnychchyna: A Collection of Dialect Texts [Волинські говірки Хмельниччини. Збірник діалектних текстів]*. TOV "Ruta", Kamyanets-Podilskyi.

O. Kuparinen, A. Miletić, and Y. Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP*

2023, pages 13814–13828, Singapore. Association for Computational Linguistics.

Roman Kyslyi, Yuliia Maksymiuk, and Ihor Pysmennyi. 2025. Vuyko Mistral: Adapting LLMs for low-resource dialectal translation. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 86–95, Vienna, Austria (online). Association for Computational Linguistics.

J. Lang, Z. Guo, and S. Huang. 2024. A comprehensive study on quantization techniques for large language models. In *2024 4th International conference on artificial intelligence, robotics, and communication (ICAIRC)*, pages 224–231. IEEE.

V. Lesnova. 2013. *Dialects of Eastern Slobozhanshchyna: A Collection of Dialect Texts [Говірки Східної Слобожанщини: збірник діалектних текстів]*. Vyd-vo DNU imeni Tarasa Shevchenka, Luhansk.

P. H. Martins, J. Alves, P. Fernandes, N. M. Guerreiro, R. Rei, A. Farajian, M. Klimaszewski, D. M. Alves, J. Pombal, N. Boizard, et al. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

H. Martynova. 2012. *Dialects of Western Poltava Region: A Collection of Dialect Texts [Говірки Західної Полтавщини: зб. діалектних текстів]*. Bohdan Khmelnytsky National University of Cherkasy, Cherkasy.

H. Martynova, T. Shcherbyna, and A. Taran. 2013. *Dialects of the Cherkasy Region: A Collection of Dialect Texts [Говірки Черкащини: збірник діалектних текстів]*. Bohdan Khmelnytsky National University of Cherkasy, Cherkasy.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

M. Plüss, J. Deriu, Y. Schraner, C. Paonessa, J. Hartmann, L. Schmidt, C. Scheller, M. Hürlimann, T. Samardžić, M. Vogel, and M. Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

A. Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

R. Rei, C. Stewart, A. C Farinha, and A. Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

T. Samardžić, Y. Scherrer, and E. Glaser. 2016. ArchiMob - a corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066, Portorož, Slovenia. European Language Resources Association (ELRA).

M. Shvedova, N. Prydvorova, and I. Skibina. 2022. Normalization of early modern Ukrainian in GRAC: The case of Lesia Ukrainka's works. In *COLINS*, pages 71–80.

M. Shvedova, R. von Waldenfels, S. Yarygin, A. Rysin, V. Starko, T. Nikolajenko, and others. 2017–2022. GRAC: General Regionally Annotated Corpus of Ukrainian. Electronic resource: Kyiv, Lviv, Jena. Available at `http://uacorpus.org`.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

C. Ziems, W. Held, J. Yang, J. Dhamala, R. Gupta, and D. Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768.