Evaluation Framework for Transfer Learning between Closely Related Lects: A Case Study of Lemko

Ilia Afanasev
University of Vienna
ilia.afanasev.1997@gmail.com

Abstract

The creation of a robust evaluation methodology is one of the pivotal issues for transfer learning between closely related lects. The current study proposes to resolve this issue by concisely implementing a group of evaluation methods that enable a more systematic qualitative analysis of errata (for instance, string similarity measures to assess lemmatisation more effectively). The paper introduces a robustness score, a metric that aims to assess the stability of model performance across different datasets.

The case study is a morphosyntactic tagging of a small historical (beginning of the twentieth century) corpus of Lemko (Slavic clade, Transcarpathian area). It presents a diversity of cross-dependent tasks, made rather complex by the rich Lemko morphology, highly influenced by areal convergence processes. The tagger is a pre-trained Stanza. The study uses modern standard Ukrainian as the source language, as it is the closest to the Lemko high-resource lect.

The analysis reveals that linguistically-aware metrics improve the speed and accuracy of analysis of the errata, especially those caused by the differences between source and target lects. The key data contribution is the open-source dataset of Lemko, obtained during the tagging tasks. Future research directions include a larger-scale test that applies more models to a more extensive material.

1 Introduction

The study addresses the lack of robust methodology for evaluation of transfer learning between closely related lects[1] that can foster linguistics-aware enhancement of the models (Section 2 refers to this along with the other theoretical issues, present at study of transfer learning from high-resource to low-resource lects). The case study is the morphosyntactic tagging of Lemko[2], a group of small East Slavic territorial lects of the Transcarpathian region (Section 3 characterises the material of the study in more detail). The main research question is to what degree fine-grained linguistics-aware evaluation assists in the errata[3] analysis.

1.1 Contribution

To answer the research question, the paper devises a framework of fine-grained linguistics-aware evaluation, adapted to transfer learning between closely related lects, described in section 4. Besides using and modifying existing metrics, including string similarity measures for lemmatisation and complete predications number for dependency parsing, the study implements robustness score, a new measure that assesses the quality of performance on the original dataset, the quality of performance on the new dataset, and the quantitative difference between them to evaluate the stability of the model. Section 5 describes the application of the pipeline to the dataset, while section 6 delves into the prospects of further enhancements to the methodology.

The other pivotal contribution is a new open-access Lemko data set with gold morphosyntactic tagging. This dataset enables further inquiries into both the lect and the method. At the same time, it provides additional information about the unique culture of Lemko at the beginning of the twentieth century that could be useful to the modern Lemko

[1] The word *lect* here denotes any group of individual linguistic repertoires (including a single one, idiolect) that could undergo approximation to a single system of features that form phonetic, grammatical, and lexical systems (Campbell, 2013, p. 191). The most frequently discussed types of lects are sociolects (lects of particular social groups), dialects (small territorial lects), and standard lect (nation-wide, codified, often top-down imposed lects).

[2] Lemko is a denotation preferred by native speakers compared to Lemkian: https://uwr.edu.pl/en/lemkos-who-are-they/ (date of access: February 11, 2026)

[3] The term *error* here denotes a disagreement between a gold (manually checked by a human expert) tag and an automatically assigned tag. A tag can be either a label (part-of-speech, morphological category) or a sequence (lemma).

community.

2 Related Work

This section consists of three parts. It begins with an overview of current methods of low-resource lects processing. The following part provides an outline of existing computational studies of Lemko. The section ends with a discussion of existing fine-grained evaluation techniques.

2.1 Processing Low-Resource Lects

The main challenge of processing low-resource lects is the lack of training material. This is when manual tagging takes too long to carry out (for example, when financing places restrictions on a research project), but models are still unable to generalise well enough from the existing data (de Graaf et al., 2022; Bloom Ström et al., 2023; Rao and Gopinath, 2023). There are two ways to approach this problem: transfer learning and rule-based processing.

Rule-based processing has been present in low-resource NLP for a long time, and it was the most efficient before the emergence of the huge corpora that allowed statistics-based approaches to gain the edge. (Mills, 1998; Chrupa la, 2006; Plisson et al., 2008; Gesmundo and Samardžić, 2012; Radziszewski, 2013). However, in low-resource settings, the rule-based approach remains a strong contender against statistical methods (Sharipov and Sobirov, 2022; Lendvai et al., 2025). In lemmatisation tasks it is possible to train the model on the corpus of higher-resource lect to detect classes of word inflection (which do not change as significantly), and then to implement new rules for a lower-resource setting, saving the pre-trained model (de Graaf et al., 2022).

Transfer learning is a process of transferring the in-domain knowledge of the model to tagging the out-of-domain data (Bengoetxea et al., 2025, p.210). The most frequent way of doing transfer learning is to perform a task zero-shot (Tebbifakhr et al., 2020), not implementing any kind of fine-tuning for the target dataset (Kim et al., 2020, p. 72). Transfer learning may include normalisation as a part of the pipeline (van der Goot et al., 2017). This allows to reduce differences between analysed lect and the high-resource neighbour, simplifying the tagging. Sometimes, small territorial lect corpora adopt this ideology completely, erasing the presentation of the features for the end user as well (von Waldenfels et al., 2014). Currently, one of the ways to perform transfer learning is to utilise LLMs even for extremely low-resource languages (Faisal and Anastasopoulos, 2024; Liang and Levow, 2025).

2.2 Lemko in Computational Linguistics

Lemko (as most of the small territorial lects) is poorly represented in terms of computational studies. Existing works, including corpora (Rabus and Šymon, 2015), generally do not treat it as a full-fledged system, but as a part of greater continua (Scherrer and Rabus, 2017; Rabus, 2018; Scherrer and Rabus, 2019). This is a continuation of an earlier trend in linguistics in general: the preferred way of studying small territorial lects for a long time was top-down. The scholars tended to consider them primarily as subdivisions of a larger group or a standard lect, studying mostly those features that differ them from an umbrella language (Kuparinen et al., 2023). This was a *differentiating approach* (Kriuchkova, 2007, p. 31).

The trend in theoretical linguistics changed in the second half of the twentieth century. Technological advances drastically increased the volume of the recorded material; and the material on small territorial lects ceased to be fragmentary. It became possible to study these lects *integrally* (Kriuchkova, 2007, pp. 31–32) as full-fledged systems of their own (Goldin, 1990; Goldin and Kryuchkova, 2011; Hromko, 2020). This enabled bottom-up studies of the dialect continua (Kalnyn', 1973, 1992). However, Lemko has not yet been incorporated into this type of computational linguistics study. This is a research gap this article aims to close.

2.3 Fine-Grained Evaluation

The evaluation of morphosyntactic tagging conventionally incorporated simple metrics, such as the accuracy score for lemmatisation (Straka and Straková, 2017; Bergmanis and Goldwater, 2018; Anastasyev, 2020; Kanerva et al., 2021; Torre Alonso, 2022). Rare exceptions include the evaluation of part-of-speech tagger performance on out-of-vocabulary items (Scherrer and Rabus, 2017, pp. 86–87), implementing string similarity measures to assess the lemmatisers' performance (Afanasev and Lyashevskaya, 2024), and scoring the correct dependencies of the predicate (Plank et al., 2015, p. 316). Overall, the evaluation focus is not on how good of a model a tool is, but on how

well the results of its use align with gold labels. Combining qualitative and quantitative evaluation is also relatively limited (Avramidis et al., 2018, pp. 245–246) to more theoretical works (Chung and Chou, 2025).

3 Data

The section starts with a short linguistic overview of Lemko, followed by a description of the most characteristic features of the dataset itself. As the lect is low-resource and relatively poorly known, and the dataset is newly compiled, both require additional introduction.

3.1 Lemko

Lemko group of small territorial lects is in the eastern part of the greater Slavic continuum, with its closest relatives being Bojko, East Carpathian, and Upper San groups (Del Gaudio, 2017, p. 78). The historical development of Slavic lects in the Carpathian region is relatively poorly understood, and their current close grouping is geography-based rather than linguistics-based (Ševel'ov, 1979, p. 37).

The main distribution area of Lemko for a long time has been the bordering region between southern Poland, northern Slovakia, and south-western Ukraine. However, during the Soviet era, many Lemkos became victims of deportation (alongside other oppressive practices) (Magocsi, 2015, pp. 336–338). The current spread of this lect does not cover its historical territory of distribution. During the twentieth century, when researchers collected the material, the Lemkos attested that their lects and their closest relatives were spoken in the villages and towns from the westernmost border of Ukraine (river Tisa) and to the Poland-Slovakia border (river Poprad in the region of Spiš) (Nakonečna and Rudnyc'kyj, 1940, p. 31). Figure 1 shows the map, reconstructed from the data provided in the interview. This reconstruction is far from being full: the interviewee clearly lived at the north of Slovakia, thus he had only a vague understanding of Lemko dispersal to the south-west. Most probably, the north-eastern part with the settlements mentioned by the interview was the territory of Lemkos, while the land to the south-west belonged to the speakers of closely related lects (Zilyns'kyj, 1933).

Linguistically, several key features separate Lemko from neighbouring small territorial lects.

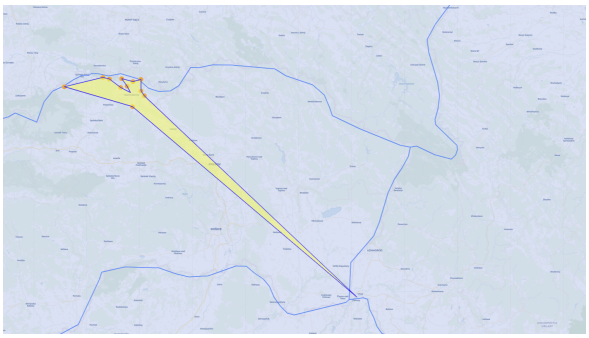

Figure 1: Map of the Transcarpathian (yellow) distribution in the beginning of the twentieth century. The blue line at the top shows the modern borders of Poland (to the north) with Slovakia (at the centre) and Ukraine (to the west), the blue line at the middle shows the modern Slovakia-Ukraine border, the blue line below shows the border that separates Hungary (to the south) from Slovakia and Ukraine. The orange dots point to the Lemko settlements, mentioned in the interviews of speakers.

These features are present at the levels of phonetics, morphology, lexis, and syntax.

As this study explores the effects of linguistic differences on transfer learning, among the features in focus are those that diverge the most from modern standard Ukrainian, because the models trained on this lect are going to be the main subject of evaluation. Thus, it is necessary to predict the possible errata to improve the interpretation of the evaluation. Nakonečna and Rudnyc'kyj (1940, pp. 23–30), Ševel'ov (1979, p. 37) and Del Gaudio (2017, pp. 76–85) provide a more detailed and general overview.

3.1.1 Phonetics

Due to the form of representation (see below in paragraph 3.2.2) the phonetic differences are relatively non-essential, and most features that are going to influence the model are in the domain of morphosyntax and lexis. However, traces of phonetic changes that separated Lemko from the other Slavic lects might still impact the behaviour of the taggers.

The most significant feature in the dataset is a rather specific pronunciation of the verb 'to speak'. The form here is *гвáрити* instead of expected *говорить* (Kopp et al., 2023)[4]. This is an areal, and, seemingly, almost exclusively East Slavic Transcarpathian[5] trait: Slovak has *hovorit'* (Ze-

[4]From here and onwards, for getting Ukrainian data the study relies on *UD_Ukrainian-ParlaMint corpus* (Kopp et al., 2023).

[5]Compare, for instance, other corpora (Rabus and Šymon,

man, 2017)[6], and Polish has a completely different lexical unit, *mówić* (Wróblewska, 2018)[7]. While seemingly not crucial, this peculiarity may significantly influence morphological tagging (if there are no character/subtoken-based embeddings, it is going to be an OOV item and the model is going to tag it unpredictably) and lemmatisation (if morphological tagging is incorrect, the model of lemma generation is very likely to be incorrect as well).

Another issue is the differences in infinitive endings. Nakonečna and Rudnyc'kyj (1940) provide a rather inconsistent account of *-mu/-mi* alternations. However, the speaker uses exclusively *-mi*, which is different from the standard Ukrainian *-mu*. This is going to affect the accuracy score, but should not crucially damage the quality of predictions.

3.1.2 Morphology

Among the many morphological peculiarities of Lemko, the dataset in question mostly contains unique (as compared to the other Slavic lects) nominal inflection features. They span across nouns, adjectives, and pronouns alike. Among these, the following are the most frequent in the dataset.

- The feminine instrumental singular ending is *-ом* (for instance, *ко́тром* 'which-FEM.INS.SG'[8]), and not *-ою*, present in modern standard Ukrainian (see *незалежною* 'independent-FEM.INS.SG').

- The instrumental plural ending is *-ма*. The example here is *ни́ма* 'they-INS.PL'. This form is also present in declension systems of other Slavic languages, cf. Belarusian *двума* 'two-INS.PL' (Shishkina and Lyashevskaya, 2021)[9]. However, only in Lemko it is consistent and not alternating with other endings.

- Determinative pronouns have long partially reduplicated forms, for instance, *мо́мо* 'DET-N.NOM.SG'. Modern standard Ukrainian lacks this feature.

3.1.3 Lexis

Most lexical peculiarities within the dataset that differ it from modern standard Ukrainian are borrowings. They come mainly from neighbouring Slavic languages and mostly include functional words: *ве́льо* 'many' < Slovak *veĺa* and *ба́рз* < Polish *bardzo* (Nakonečna and Rudnyc'kyj, 1940, p. 30). However, there are also borrowings from other languages, for instance, Hungarian: *кіральфі́я* 'prince-GEN.SG' (< Hungarian *királyfi* 'id.' (Nakonečna and Rudnyc'kyj, 1940, p. 30)).

3.1.4 Syntax

The key feature of the Lemko syntax is the copular structures (Fontański and Chomiak, 2000, p. 141). Here, the example would be але то єст єдна бесі́да. 'but this.SHORT-N.NOM.SG AUX one-FEM.NOM.SG language-NOM.SG'. This structure is present in the neighbouring Polish language: *To jest mój debiut* 'this.-N.NOM.SG AUX my-MASC.NOM.SG debut-NOM.SG'. Modern standard Ukrainian lacks this kind of copular structures; therefore, at all stages of morphosyntactic tagging such sentences may present issues to the pre-trained model.

3.2 Dataset

This subsection describes the fragment (Baglioni and Rigobianco, 2024, pp. 1–9) of Lemko that constitutes the dataset. It discusses its source, some basic quantitative characteristics, along with the form of representation.

3.2.1 Source

The original dataset is LA1407 (Nakonečna and Rudnyc'kyj, 1940, pp. 31–35), a set of three texts, recorded in the late 1930s from a single Lemko speaker who grew in the village of Kamienka (Lemko *Камюнка*; now a part of the Prešov Region in the north of Slovakia) and transcribed as a part of more general Transcarpathian lects survey. The data presents three layers: detailed phonetic transcription, standard-like transcription, and German translation. The unit of data is a predication (some sentences are rather long, which led to them being cut into chunks). Overall, the dataset estimates to 609 tokens. Of these 609 tokens, 5.58% are adverbs, 7.88% are adjectives, 11.99% are verbs, 15.93% are punctuation marks, 22.66% are proper and common names, and 35.96% are function words.

2015)

[6]From here and onwards, for getting the Slovak data the study relies on *UD_Slovak-SNK corpus*.

[7]From here and onwards, for the Polish data the study relies on *UD_Polish-PDB corpus*.

[8]Glosses given according to Comrie et al. (2008)

[9]From here and onwards, for getting the Belarusian data the study relies on *UD_Belarusian-HSE corpus*.

3.2.2 Digital Representation

The dataset is represented digitally as a .conllu-file. It contains texts in all three forms of their representation: phonetic transcription, German text, and standard-like transcription. It also brings three key modifications.

Phonetic transcription uses the IPA rendering of the original system instead of copying the latter. The metadata now provide the English translation alongside the German. For better dependency parsing, the study merges predication units into clauses, where possible. These changes facilitate a better understanding of the material by both scholars and tools.

4 Method

This section starts with an experiment workflow outline. The subsequent part describes the model used for dataset tagging. The section ends with a discussion of the fine-grained evaluation metrics.

4.1 Experiment Workflow

The study represents three consecutive stages of tagging the dataset (Afanasev, 2026): part-of-speech and morphological tagging, lemmatisation, and dependency parsing. Between each stage, the research uses a *human-in-the-loop* (Jiang et al., 2024; Umphrey et al., 2024; Verma et al., 2025) approach. After conducting each stage of tagging and before each stage of evaluation, a human scholar checks the result of the tagging stage to create a gold version of the data. This approach facilitates the evaluation and, crucially, a more efficient tagging during the next stage (Anastasyev, 2020; Milintsevich and Sirts, 2021).

The part-of-speech and morphological tagging study is based on two experiments. The first involves tagging a pre-tokenized LA1407 corpus with Stanza, the model that undergoes the evaluation through all of the tasks (see subsection 4.2), with the primary aim of obtaining labels for the dataset. The second experiment compares different models on different materials and is designed to evaluate the newly introduced metrics (see subsection 4.3). For lemmatisation and dependency parsing, the study conducts two further experiments: the first uses manually corrected tagging from the preceding stages, and the second starts from raw text.

4.2 Model

The key utilised model is Stanza (Qi et al., 2020) pre-trained on the Ukrainian language (in terms of standard lects, probably the closest relative of Lemko (Ševel'ov, 1979, p. 37)). It is computationally effective, running even on relatively slow CPUs, and achieving over 90% accuracy on its test dataset. As the study focuses on evaluation and analysis, it is the only model implemented for all tasks. While Generative AI models could have provided better results, their zero-shot use is not always as effective; and research risks becoming irreproducible after a very short time (de Wynter, 2025).

As the study introduces a new metric, it is crucial to implement data and method cross-evaluation techniques (Afanasev, 2024). The model used for comparison is *stanzatagger*[10], a modification of Stanza that employs character-based embeddings and is better adapted to lower-resource settings (Scherrer, 2021). The training data are, as with Stanza, *UD_Ukrainian-IU*[11]. The training configuration follows the guideline from the project repository. The material used to assess the efficiency of the trained model includes the test subset of *UD_Ukrainian-IU*, the test subset of *UD_Ukrainian-ParlaMint* (Kopp et al., 2023), and LA1407. This setup ensures diversity of material and enables a more robust evaluation of both the model in comparison with Stanza and the newly introduced metric.

4.3 Metrics

The study implements three types of metrics. The first group facilitates a more effective comparison between the model performance on the source lect (standard Ukrainian) and the target lect (Lemko). The purpose of the second type is to provide a more linguistically-aware picture that enables the assessment of transfer learning. The third type implements fine-grained measurements, used for specific subcategories within the data.

4.3.1 Standard Metrics

To evaluate part-of-speech and morphological tagging, the study uses macro-F1 score, a traditional metric in the field (Scherrer and Rabus, 2019; Qi et al., 2020). It also implements exact match

[10]The code is available at the GitHub repository.

[11]The corpus is available at the Universal Dependencies repository.

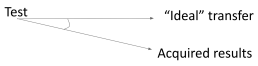

Figure 2: Geometric value of robustness between an ideal (top) and a result (bottom) vectors.

(overall accuracy score of part-of-speech and morphological tagging). For assessing the lemmatisation quality, the traditional metrics is the accuracy score. To estimate the efficiency of dependency parsing, the study uses Unlabelled Attachment Score (UAS; how many elements have a correct head, i. e., the element from which they depend syntactically) and Labelled Attachment Score (LAS; how many elements have a correct head and a correct dependency type).

4.3.2 Quality of Transfer

The key metric to interpret the quality of transfer learning is robustness. The robustness in its core is a cosine similarity measure between the vectors of ideal transfer and result. For both vectors, the starting point is (0; in-domain accuracy score). In case of Ukrainian-trained Stanza in the task of part-of-speech tagging it is (0; 97.52). The end point of an ideal vector is (END; in-domain accuracy score). END here is a researcher-chosen number that is higher than 0. This study uses 25[12]. The end point of the result vector is (END; out-of-domain accuracy score). If the out-of-domain accuracy score is higher or equal to the in-domain accuracy score, the value of robustness is positive, otherwise it is negative. The study uses robustness instead of simple comparison, because it gives better scores to the models that support stable and high level of performance in both in-domain and out-of-domain datasets. Figure 2 represents the robustness graphically.

When performing transfer learning, it is crucial to recognise not only how accurate the model is, but also how well it grasps the concept of what it has to do, and how dramatic its fall in quality on out-of-domain data is. The metrics for this task are string similarity measures: Levenshtein distance (Levenshtein, 1966) and Jaro-Winkler distance (Winkler, 1990). These metrics are very helpful in analysing lemmatisation of non-standard lects (Afanasev and Lyashevskaya, 2024), providing a clearer picture in case of non-standard inflection in the dataset.

[12]Experiments showed that if END is significantly smaller than 25, the metric loses sensitivity, if the fall in accuracy is too big, while if the END is significantly higher than 25, the metric loses sensitivity, if the fall in accuracy is rather small

For evaluating dependency parsing, the study adds three metrics: tree edit distance, Unlabelled and Labelled Complete Predications (TED, UCP and LCP). TED is a number of edits required to get a gold tree from a predicted one. UCP is a share of correctly detected dependencies of the verbal predicate. LCP is its stricter version that scores a share of correctly detected relations of the verbal predicate. This study slightly modifies the metric by Plank et al. (2015, p. 316). Unlike the original one, it fines the model for the generated dependencies that are not part of the gold dataset.

4.3.3 Fine-Grained Metrics

To get a more detailed view of the results, it is possible to evaluate the performance of the model by groups of labels. In case of part-of-speech and morphological tagging, these are parts of speech. Dependency parsing may use relationship labels. As Stanza uses a sequence-to-sequence model for lemmatisation, the word inflection classes are unavailable; the study opts for using parts of speech.

5 Results and Discussion

The section begins with a cross-evaluation of the robustness score. It provides an initial overview of the results and addresses the issues that arise during part-of-speech and morphological tagging, drawing on fine-grained metrics. The following subsection discusses the advantages of using string similarity measures as evaluation metrics for lemmatisation. Finally, based on the dependency parsing results, the study delves into linguistic aspects of transfer learning.

5.1 Cross-Evaluation of Robustness Score

Before analysing the performance of Stanza, it is crucial to place the robustness score in context. Table 1 presents the values of this metric for different datasets and compares them with a simple difference measure, defined as the subtraction of the model performance on the new dataset from its performance on the test subset of the training dataset.

Robustness score does not produce substantially different results from the simple subtraction measure. Both metrics highlight the key contrast between the two models: while *stanzatagger* not only maintains but slightly improves its tagging quality from *UD_Ukrainian-IU* to the test subset of Kopp et al. (2023), performance on LA1407 remains comparably poor (*stanzatagger* lags slightly

Dataset (model)	PoS	Exact match	Robustness (PoS)	Difference (PoS)	Robustness (Exact match)	Difference (Exact match)
UD_Ukrainian-IU (Stanza)	**97.52**	**92.07**	–	–	–	–
UD_Ukrainian-ParlaMint (Stanza)	89.38	83.59	-0.95	-8.14	-0.95	-8.48
LA1407 (Stanza)	66.78	55.99	-0.63	-30.74	-0.56	-36.08
UD_Ukrainian-IU (stanzatagger)	95.54	76.58	–	–	–	–
UD_Ukrainian-ParlaMint (stanzatagger)	96.06	76.59	**1**	**0.52**	**1**	**0.01**
LA1407 (stanzatagger)	62.89	40.56	-0.61	-33.17	-0.57	-36.02

Table 1: The results of tagging the test subsets of *UD_Ukrainian-IU*, Kopp et al. (2023) and LA1407. All the metrics, except for robustness and difference performance of the model on standard Ukrainian and Lemko, are in per cent values. Rounding is to the second digit after zero. Best values are in **bold**.

behind in PoS, whereas Stanza marginally underperforms in exact match). The main difference is that the robustness score smooths smaller discrepancies between the models, providing a clearer comparative picture than the simple difference.

5.2 Stanza Results

The overall results of the Stanza performance are rather poor. Taggers lose significantly in their accuracy, when compared to the initial dataset. Table 2 shows the results.

Lemmatisation and UAS demonstrate the best robustness scores, while exact match demonstrates the lowest one. Lemmatisation witnesses a relatively small fall; while the performance of UAS is surprisingly stable, given the initial low score of the system. Overall, robustness shows the difference between the easier tasks in the pipeline (lemmatisation with gold tags and head marking) and the more complicated ones (producing an exact match of part-of-speech and morphological tags), highlighting the spots for further enhancements. The simple difference metric between model performance on standard Ukrainian and on Lemko captures the overall drop in accuracy scores but misses important nuances: for instance, it fails to show how well the lemmatiser and part-of-speech tagger perform on the test dataset.

5.3 Fine-Grained Evaluation of Morphological Tagging

The robustness of morphological tagging placed it among the hardest tasks to transfer. To further discover the issue, the study performs a by-part-of-speech evaluation. Figure 3 shows the results.

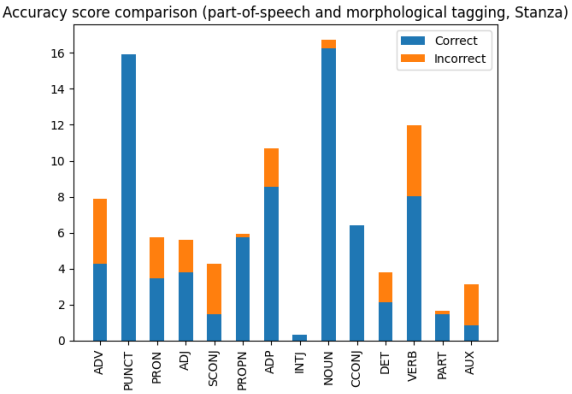

Figure 3: A by-part-of-speech evaluation of exact match results. Y-axis shows the share of part-of-speech in the dataset, the blue part of bars – the share of correctly tagged items of this part-of-speech, the orange part – the share of incorrectly tagged items of this part-of-speech. All the values are in %, rounded to the second digit.

As the figure demonstrates, the model struggles heavily with what in theoretical linguistics is sometimes called *Satellite Cluster B*(Reid, 2011, pp. 1107–1108) – verb and its adjacents. The results are especially problematic for the verb itself, adverbs, and auxiliaries. Here, the main issue is syntactical: the presence of copular sentences confuses the model. This leads to incorrect tagging of the auxiliaries. In addition, a lot of function words, and especially adverbs, like бараз 'very' are loanwords from other Slavic languages. They get an X tag from the model, which may be true for modern

Metric	Standard Ukrainian	Lemko	Robustness	Difference
PoS macro-F1	**97.52**	66.78	-0.63	-30.74
Feats macro-F1	93.11	71.00	-0.74	-22.11
Exact match	92.07	55.99	-0.56	-36.08
Lemmatisation accuracy score (gold tagging)	96.72	**75.7**	*-0.77*	-21.02
UAS (gold tagging)	85.87	67.65	*-0.77*	**-18.22**
LAS (gold tagging)	82.77	52.71	-0.6	-30.06

Table 2: The results of tagging *LA1407* with Stanza (Qi et al., 2020). All the metrics (except for robustness and difference between performance of the model on standard Ukrainian and Lemko) are in per cent values. Rounding is to the second digit after zero. Best values are in **bold**, matched best values are in *italics*.

standard Ukrainian, but most certainly is incorrect for Lemko. Verbs are the most enigmatic case, as most errors in their tagging come from Aspect and Tense. The model often confuses the imperfective verbs with the perfective ones, as the imperfectives often took on iterative sense, cf. *полю́бят* 'love.IPFV-PRES.3PL'.

5.4 Lemmatisation

Lemmatisation was a comparatively easier task where Stanza demonstrated robustness. To determine the principal contribution, it is crucial to perform a more in-detail view. Figure 3 shows the results.

Gold tagging seems to boost the accuracy score, but efficiency of tagging almost does not increase for string similarity measures, especially for Jaro-Winkler distance. This shows that the model generally grasps the concept of lemmatisation rather well. Still, the decrease in accuracy score is significant and requires a more thorough investigation, demonstrated in Figure 4.

Once again, the most complicated task for the model is to process verbs; the accuracy score is close to zero. On the contrary, string similarity measures show very good results (Levenshtein distance of 1.15 means the average error of 1.15 symbols, and Jaro-Winkler distance is 0.92). The inspection of the results shows that the most frequent (and almost the only) error is the wrong ending, *-mu* instead of *-mi*. This shows the value of fine-tuning or class-based lemmatisation: with a good rule system, this would not be an issue. However, the lemmatiser of Stanza is a sequence-to-sequence model, unable to precisely transfer its understanding of lemmatisation to new material.

One equally difficult but less problematic case is the auxiliaries (accuracy score of 15.79%, average Levenshtein distance of 2.88 and average Jaro-

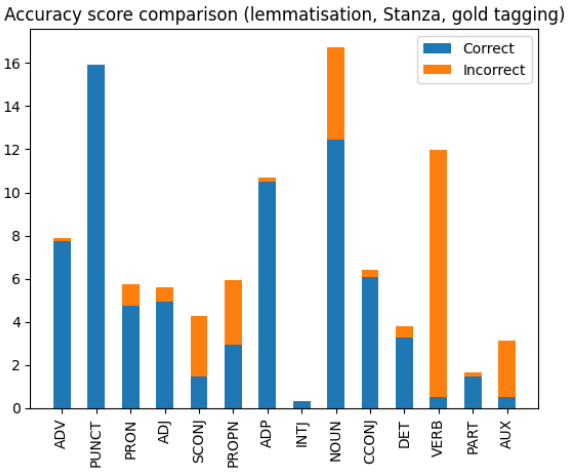

Figure 4: A by-part-of-speech evaluation of lemmatisation results. Y-axis shows the share of part-of-speech in the dataset, the blue part of bars – the share of correctly lemmatised items of this part-of-speech, the orange part – the share of incorrectly lemmatised items of this part-of-speech. All the values are in %, rounded to the second digit.

Winkler distance of 0.56). The issue at stake here is very similar to the part-of-speech and morphological tagging. The model just does not know what to do with these items, completely alien to the original dataset. They get a very random tagging that has nothing to do with their lemmatisation.

5.5 Dependency Parsing

The evaluation of dependency parsing indicates similar issues as the other types of tagging. Table 4 shows the results.

TED of 9 seems to indicate a significant issue, however, this is mostly due to the presence of large sentences that span 60 tokens. More problematic is the contrast between assignment scores (UAS and LAS) and complete predications (UCP and LCP). The model, once again, fails to correctly identify

Metric	Gold tagging	Silver tagging
Accuracy score	75.7	70.11
Levenshtein distance	1.44	1.6
Jaro-Winkler distance	0.87	0.87

Table 3: Lemmatisation evaluation results, rounded to the second digit.

Metric	Result
UAS	67.65
LAS	52.71
TED	9
UCP	47.18
LCP	35.57

Table 4: Dependency parsing evaluation results, rounded to the second digit. UAS, LAS, UCP and LCP values are percentages, TED value is an absolute value.

some Satellite Clusters B, dragging down its performance. This highlights the necessity of introducing copular structures (probably, by introducing the neighbouring languages material (Scherrer and Rabus, 2019)) into the dataset for further tagging.

5.6 Discussion

All metrics identify a single issue at the core of all the Stanza downfalls: clause structure, specifically, copular sentences, strongly characteristic of Lemko, and quite frequent in Carpathian linguistic area in general but very rarely represented in modern standard Ukrainian. The model is unable to identify the relations between tokens and, therefore, perceives the tokens themselves wrongly at the morphological level.

Lemmatisation suffers the least, because it is relatively syntax-independent, especially with lects being relatively close in terms of vocabulary. In addition, it probably is the biggest beneficiary of human-in-the-loop: correctly assigned morphology tags aid the lemmatisation the most. For instance, *ні́ма* 'PRON.3PL.INS' without gold tagging gets **Кіма* as lemma, instead of intended *він* 'PRON.MASC.3SG.NOM'. Gold morphological tagging resolves the issue.

Still, the predications are clearly the weakest point (as very low values of UCP and LCP show), even with gold morphological tagging and lemmata. The key issue here is the poor identification of the lexical verb, the attractor of Satellite Cluster B. Further study needs to address the issue.

6 Conclusion

The study demonstrates that more fine-grained evaluation techniques are better at providing a concise summary of the errata that models make. This includes explaining anomalies; for instance, string similarity measures highlight the overall high quality of lemmatisation despite a low accuracy score. The metrics also help to identify key features that separate a target lect from a source lect, and thus to prepare better post-processing or fine-tuning techniques, fostering a more robust transfer learning.

The paper contributes to the study and representation of low-resource small territorial lects by making an open-access dataset of Lemko from the beginning of the twentieth century. This dataset contains more than 600 tokens with UPOS, UFeats, lemmata, and dependency parsing tags, and it is possible to use it for further experiments with other models.

Future research directions include both enriching the material and developing the metrics. There is a clear need to extend the data to other Lemko material and other Transcarpathian lects that possess their own unique features. For a more transparent demonstration of the robust evaluation advantages, it is utmost to include more models to the comparison, including region-specific ones (Scherrer and Rabus, 2019). The crucial step would be to develop a computationally effective language-aware model (Chung and Chou, 2025).

Limitations

LA1407 (Nakonečna and Rudnyc'kyj, 1940, 31–37), the main material of the research, does not represent Lemko (and Transcarpathian lects in general) of the 1920s-1930s in its entirety; a lot of material is in the process of digitisation. In addition, LA1407 represents only one speaker of Lemko, which can affect the distributions.

Ethical Considerations

The data had been published in printed form and available for research purposes for fifty to ninety

years by the time this article was written. Still, I anonymise the metatagging, where possible, masking the names of the speakers, to compensate for possible ethics violations that could have happened at the time of material collection.

The data themselves can contain slight mentions of xenophobic behaviour and religious (mostly, Christian) imagery. Discretion is advised.

Disclosure of Generative AI use

This study does not use Generative AI (in the modern colloquial meaning: the decoder models with more than a billion parameters trained on high-resource corpora; Stanza (Qi et al., 2020) is also a generative AI, but it performs on a much lesser scale, locally run and reproducible) in the research process. During the process of editing the authors utilised Generative AI (Grammarly) to polish the phrasing of the parts of the work where the authors felt that their non-native knowledge of English was not sufficient to produce grammatically and/or stylistically correct sentences. However, the written text is not a product of AI generation.

Acknowledgements

I would like to acknowledge all of the speakers, whose speech presents the recordings of the studied small territorial lects, the scholars who produced the initial transcription, as well as the research groups who produced the revised transcriptions. I also owe special thanks to Olga Fedorivna Mygolynets (ukr. Ольга Федорівна Миголинець, University of Uzhhorod), who greatly helped me with the understanding of transcription systems and phonetics of the lects analysed.

References

Ilia Afanasev. 2024. *The Cross-Evaluation Crux for Computational Phylogenetic Linguistics*, pages 75–89. Springer, Cham.

Ilia Afanasev. 2026. Evaluation framework for transfer learning between closely related lects: A case study of lemko – supplementary material.

Ilia Afanasev and Olga Lyashevskaya. 2024. *Chapter 2 String Similarity Measures for Evaluating the Lemmatisation in Old Church Slavonic*, pages 13 – 35. Brill, Leiden, The Netherlands.

Dan Anastasyev. 2020. Exploring pretrained models for joint morpho-syntactic parsing of Russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020"*, pages 1–12. Moscow.

Eleftherios Avramidis, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2018. Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 243–248, Boston, MA. Association for Machine Translation in the Americas.

Daniele Baglioni and Luca Rigobianco. 2024. *Chapter 1 Rethinking Fragmentariness and Reconstruction: An Introduction*, pages 1 – 25. Brill, Leiden, The Netherlands.

Jaione Bengoetxea, Mikel Zubillaga, Ekhi Azurmendi, Maite Heredia, Julen Etxaniz, Markel Ferro, and Jeremy Barnes. 2025. HiTZ at VarDial 2025 NorSID: Overcoming data scarcity with language transfer and automatic data annotation. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 209–219, Abu Dhabi, UAE. Association for Computational Linguistics.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Eva-Marie Bloom Ström, Onelisa Slater, Aron Zahran, Aleksandrs Berdicevskis, and Anne Schumacher. 2023. Preparing a corpus of spoken Xhosa. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 62–67, Gothenburg, Sweden. Association for Computational Linguistics.

Lyle Campbell. 2013. *Historical Linguistics: An Introduction*, ned - new edition, 3 edition. Edinburgh University Press.

Grzegorz Chrupa la. 2006. Simple data-driven context-sensitive lemmatization. *Proces. del Leng. Natural*, 37:121–137.

Meng-Hsuan Chung and Chao-Ting Tim Chou. 2025. Climbing towards the nlu of the universal reading of shei 'who'. *Concentric*, 51(2):303–348.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. https://www.eva.mpg.de/lingua/resources/glossing-rules.php. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig.

Evelien de Graaf, Silvia Stopponi, Jasper K. Bos, Saskia Peels-Matthey, and Malvina Nissim. 2022. AGILe: The first lemmatizer for Ancient Greek inscriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5334–5344, Marseille, France. European Language Resources Association.

Adrian de Wynter. 2025. Awes, laws, and flaws from today's LLM research. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12834–12854, Vienna, Austria. Association for Computational Linguistics.

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

Fahim Faisal and Antonios Anastasopoulos. 2024. Data-augmentation-based dialectal adaptation for LLMs. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 197–208, Mexico City, Mexico. Association for Computational Linguistics.

Henryk Fontański and Miros lawa Chomiak. 2000. *Gramatyka języka lemkowskiego*. Śląsk, Katowice.

Andrea Gesmundo and Tanja Samardžić. 2012. Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 368–372, Jeju Island, Korea. Association for Computational Linguistics.

V. E. Goldin. 1990. K proektu tekstovogo dialektologicheskogo podfonda mashinnogo fonda russkogo jazyka [on the project of the textual dialectological sub-fund of the machine fund of the russian language]. In *Materialy III Vsesojuznoj konferencii po sozdaniju Mashinnogo fonda russkogo jazyka [Materials of the 3rd All-Union Conference on the Creation of the Machine Fund of the Russian Language]*, pages 92–103, Moscow. Izd-vo Moskovskogo universiteta.

V. E. Goldin and O. Yu. Kryuchkova. 2011. Korpus russkoi dialektnoi rechi: kontseptsiia i parametry otsenki [Corpus of Russian Dialectal Speech: Concept and Evaluation Parameters]. In *Komp'uternaia lingvistika i intellektual'nye tekhnologii : Materialy ezhegodnoi Mezhdunarodnoi konferentsii, Bekasovo, 25–29 maia 2011 goda [Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference, Bekasovo, May 25–29, 2011]*, volume 10, pages 359–367, Moscow. Russian State University for the Humanities.

Tetiana Vasylivna Hromko. 2020. Stanovlennia monohovirkovoi deskryptsii u vitchyznianomu movoznavstvi (kinets' xix – 40-i roky xx st.) [formation of monographic description in domestic linguistics (end of the xix – 40s of the xx century)]. In M. Pantiuk, A. Dushnyi, and I. Zymomria, editors, *Aktual'ni pytannia humanitarnykh nauk: mizhvuzivs'kyj zbirnyk naukovykh prats' molodykh vchenykh Drogobyts'koho derzhavnoho pedahohichnoho universytetu imeni Ivana Franka [Current Issues of the Humanities: Interuniversity Collection of Scientific Works of Young Scientists of the Drohobych Ivan Franko State Pedagogical University]*, Vypusk 34, Tom 2, pages 118–123. Vydavnychyi dim "Hel'vetyka", Drohobych.

Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex Pentland, Yoon Kim, Deb Roy, and Jad Kabbara. 2024. Leveraging large language models for learning complex legal concepts through storytelling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7194–7219, Bangkok, Thailand. Association for Computational Linguistics.

L. È. Kalnyn'. 1973. *Opyt modelirovanija sistemy ukrainskogo dialektnogo jazyka [An Attempt at Modeling the System of the Ukrainian Dialectal Language]*. Nauka, Moscow.

L. È. Kalnyn'. 1992. Foneticheskij stroj odnogo gucul'skogo govora [the phonetic system of a hutsul dialect]. In *Issledovanija po slavjanskoj dialektologii. [Vyp.] 1: Karpato-ukrainskie dialekty [Studies in Slavic Dialectology. [Issue] 1: Carpathian-Ukrainian Dialects]*. Nauka, Moscow.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2021. Universal Lemmatizer: A sequence-to-sequence model for lemmatizing Universal Dependencies treebanks. *Natural Language Engineering*, 27(5):545–574.

Hwichan Kim, Tosho Hirasawa, and Mamoru Komachi. 2020. Zero-shot North Korean to English neural machine translation by character tokenization and phoneme decomposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 72–78, Online. Association for Computational Linguistics.

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

O. Ju. Kriuchkova. 2007. Élektronnyj korpus russkoj dialektnoj rechi i principy ego razmetki [electronic corpus of russian dialect speech and the principles of its marking]. *Izvestija Saratovskogo universiteta. Seriâ: Filologiâ. Žurnalistika [Proceedings of Saratov University. Series: Philology. Journalism]*, 7(1):51–55.

Olli Kuparinen, Aleksandra Miletić, and Yves 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.

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2025. Retrieval of parallelizable texts across Church Slavic variants. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 105–114, Abu Dhabi, UAE. Association for Computational Linguistics.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37, Vienna, Austria. Association for Computational Linguistics.

Paul R Magocsi. 2015. *With their backs to the mountains : a history of Carpathian Rus' and Carpatho-Rusyns*. Central European University Press,, Budapest :.

Kirill Milintsevich and Kairit Sirts. 2021. Enhancing sequence-to-sequence neural lemmatization with external resources. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3112–3122, Online. Association for Computational Linguistics.

Jon Mills. 1998. Lemmatisation of the corpus of Cornish. In *Workshop on Language Resources for European Minority Languages, LREC First International Conference on Language Resources and Evaluation*, pages 1–6, Granada, Spain.

Hanna Nakonečna and Jaroslav Bohdan Rudnyc'kyj. 1940. *Ukrainische Mundarten : Südkarpatoukrainisch ; (Lemkisch, Bojkisch und Huzulisch) [Ukrainian dialects: South Carpathian Ukrainian; Lemkian, Bojkian and Huzulian]*. Arbeiten aus dem Institut für Lautforschung an der Universität Berlin ; 9. Otto Harrassowitz, Berlin.

Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkler, and Anders Søgaard. 2015. Do dependency parsing metrics correlate with human judgments? In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 315–320, Beijing, China. Association for Computational Linguistics.

Joël Plisson, Nada Lavrac, Dunja Mladenić, and Tomaž Erjavec. 2008. Ripple down rule learning for automated word lemmatisation. *AI Commun.*, 21:15–26.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

A. Rabus. 2018. Obrazovanie prošedšego vremeni v raznovidnostjach karpatorusinskogo: kvantitativnyj analiz [the formation of the past tense in varieties of carpathian rusyn: A quantitative analysis]. In Kvetoslava Koporova, editor, *20 rokiv vyšokoškol'skoj rusynistikŷ na Slovakiji [20 Years of University Rusyn Studies in Slovakia]*, pages 139–151. Prešovská univerzita v Prešove, Prešov.

A. Rabus and A. Šymon. 2015. Na novŷch putjach isslidovanja rusyns'kŷch dialektu. korpus rozhovornoho rusyns'koho jazŷka [on new paths of rusyn dialect research. corpus of spoken rusyn language]. In Kvetoslava Koporová, editor, *Rusyn'skyj literaturnyj jazyk na Slovakiji. 20 rokiv kodifikaciji [The Rusyn Literary Language in Slovakia. 20 Years of Codification]*, pages 40–54. Prešov University Publishing, Prešov.

Adam Radziszewski. 2013. Learning to lemmatise Polish noun phrases. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 701–709, Sofia, Bulgaria. Association for Computational Linguistics.

Amit Rao and Kanchi Gopinath. 2023. A Sanskrit grammar-based approach to identify and address gaps in Google Translate's Sanskrit-English zero-shot NMT. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 141–166, Gothenburg, Sweden. Association for Computational Linguistics.

Wallis Reid. 2011. The communicative function of English verb number. *Natural Language & Linguistic Theory*, 29(4):1087–1146.

Yves Scherrer. 2021. *Adaptation of Morphosyntactic Taggers*, page 138–166. Studies in Natural Language Processing. Cambridge University Press.

Yves Scherrer and Achim Rabus. 2017. Multi-source morphosyntactic tagging for spoken Rusyn. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 84–92, Valencia, Spain. Association for Computational Linguistics.

Yves Scherrer and Achim Rabus. 2019. Neural morphosyntactic tagging for rusyn. *Natural Language Engineering*, 25(5):633–650.

Maksud Sharipov and Ogabek Sobirov. 2022. Development of a rule-based lemmatization algorithm through finite state machine for uzbek language. *CoRR*, abs/2210.16006.

Yana Shishkina and Olga Lyashevskaya. 2021. Sculpting enhanced dependencies for belarusian. In *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, page 137–147, Berlin, Heidelberg. Springer-Verlag.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Amirhossein Tebbifakhr, Matteo Negri, and Marco Turchi. 2020. Machine-oriented NMT adaptation for zero-shot NLP tasks: Comparing the usefulness of close and distant languages. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 36–46, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Roberto Torre Alonso. 2022. Automatic lemmatization of Old English class III strong verbs (L-Y) with ALOEV3. *Journal of English Studies*, 20:237–266.

Ray Umphrey, Jesse Roberts, and Lindsey Roberts. 2024. Investigating expert-in-the-loop LLM discourse patterns for ancient intertextual analysis. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 31–40, Miami, USA. Association for Computational Linguistics.

Rob van der Goot, Barbara Plank, and Malvina Nissim. 2017. To normalize, or not to normalize: The impact of normalization on part-of-speech tagging. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 31–39, Copenhagen, Denmark. Association for Computational Linguistics.

Kanishk Verma, Sri Balaaji Natarajan Kalaivendan, Arefeh Kazemi, Joachim Wagner, Darragh McCashin, Isobel Walsh, Sayani Basak, Sinan Asci, Yelena Cherkasova, Alexandrous Poullis, James O'Higgins Norman, Rebecca Umbach, Tijana Milosevic, and Brian Davis. 2025. BullyBench: Youth & experts-in-the-loop framework for intrinsic and extrinsic cyberbullying NLP benchmarking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2172–2208, Suzhou (China). Association for Computational Linguistics.

Ruprecht von Waldenfels, Michael Daniel, and Nina Dobrushina. 2014. Why standard orthography? building the ustya river basin corpus, an online corpus of a russian dialect. In *Komp'juternaja lingvistika i intellektual'nye technologii: Po materialam ežegodnoj Meždunarodnoj konferencii «Dialog» (Bekasovo, 4 — 8 ijunja 2014 g.) [Computational Linguistics and Intellectual Technologies: Based on the materials of the Annual International Conference "Dialog" (Bekasovo, June 4-8, 2014)]*, volume 13 (20), Moscow. Izd-vo RGGU.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359.

Alina Wróblewska. 2018. Extended and enhanced polish dependency bank in universal dependencies format. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 173–182. Association for Computational Linguistics.

Daniel Zeman. 2017. Slovak Dependency Treebank in Universal Dependencies. *Jazykovedný časopis / Journal of Linguistics*, 68(2):385–395.

Ivan M Zilyns'kyj. 1933. *Karta ukraïns'kych hovoriv : z pojasnennjamy ; mirylo 1:4.000.000*. Praci Ukräins'koho Naukovoho Institutu 14. Ukraïns'kyj Naukovyj Instytut, Warszawa.

Jurij Volodymyrovyč Ševel'ov. 1979. *A historical phonology of the Ukrainian language*. Historical phonology of the Slavic languages ; 4. Winter, Heidelberg.