Quantitative Lect Description: A Case Study of Lemko from the Field Data of 1920s-1930s

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

Abstract

While qualitative descriptions (in the form of reference grammars) and benchmarks for low-resource languages are becoming increasingly widespread, computational linguists do not often use quantitative methods to describe a new lect rather than a new model. This paper intends to close this lacuna.

The case study is a Lemko text transcribed at the beginning of the twentieth century. Using morphosyntactic tagging and topic modelling, the study demonstrates areal influences and archaic features of the lect. Fine-grained evaluation significantly assists in identifying subtle patterns that are not readily apparent through traditional metrics such as accuracy score.

The results highlight the necessity of a more detailed analysis of model performance, which may yield more linguistically significant results than a purely manual check. This information is present in the resulting dataset, which can be used for further investigation into the structural features of the Lemko lect.

1 Introduction

The goal of this study is to provide an example of a quantitative characterisation of a low-resource lect that can enhance (but certainly not substitute for) the work of a field researcher through distributional analysis based on neighbouring lects. Quantitative characterisation is something that current studies of low-resource lects lack: while toolkits and benchmarks are crucial for language preservation, they do not in themselves provide a means of scientific description.

This often leaves researchers relying on qualitative methods for interpreting the errata produced by toolkits for benchmarks, and even more so in the age of Large Language Models (LLMs), which often become a single tool in a toolkit and remain barely explainable from any perspective, including a linguistic one (Rakotonirina et al., 2025). A more detailed investigation of this problem is presented in Section 2.

The distributions of particular features in raw texts of the lects remain in a blind spot between quantitative metrics that assess the number of correct outputs, qualitative interpretations of inconsistencies, and preprocessed datasets in variationist sociolinguistics. Section 4 proposes a way to close this gap, while Section 5 demonstrates a practical implementation of the algorithm. Section 6 provides an overview of the study and outlines its prospects.

The case study for this research is the Lemko[1] lects of the western part of the Transcarpathian region. Section 3 describes this dataset in detail. Using it, the paper demonstrates the benefits of taking a distant, *a posteriori* perspective on a new lect that is not heavily influenced by presupposed categories.

2 Related Work

The following section begins with a presentation of existing frameworks for quantitative and qualitative description. The next part focuses on quantitative studies of variation within corpora. The section ends with an overview of existing NLP studies of Lemko.

2.1 Frameworks of language description

Language documentation and description were among the earliest linguistic endeavours, especially in the study of newly documented lects. Prominent examples include the first grammars of South American languages (Domingo de Santo Tomás, 1560; Ruiz de Montoya, 1640; Adam and Henry, 1880). However, for a long time a much more widespread practice (illustrated below with examples from indigenous languages of Asia and

[1] Lemko is a denotation preferred by native speakers compared to Lemkian: https://uwr.edu.pl/en/lemkos-who-are-they/ (date of access: February 12, 2026)

Africa) was the collection of basic vocabulary lists (Dobrotvorskii, 1875; Munkácsi, 1894; Zukowsky, 1924) or text corpora (Bleek and Lloyd, 1911; Bleek, 1929). Grammars, enhanced by collections of texts, have become a standard means of describing a language only relatively recently; see, among others, Wiedemann (1884); Nakonečna and Rudnyc'kyj (1940).

In modern linguistics, the most traditional form of language description is the production of a reference grammar; some publishers dedicate entire series to this purpose (Lehman, 1989; Zigmond et al., 1991; Kimball, 1991; Osada, 1992; Brindle, 2017; Daniel et al., 2019; Namyalo et al., 2021). This is an extremely valuable and well-developed genre of linguistic literature that covers the historical development, phonetics, morphology, and syntax of a given lect. It recognises and critically assesses the *a priori* nature of many linguistic categories with which it must operate (Terhart, 2024, pp.113–114). Still, it rarely focuses on quantitative characteristics and distributions within corpora that represent actual linguistic practices of communities.

Natural Language Processing (NLP) studies, in contrast, focus on developing toolkits (Bolt et al., 2019; Tolmachev et al., 2018; Pauli et al., 2021; Rennes et al., 2022) and designing benchmarks (Shavrina et al., 2020; Aparovich et al., 2025; Chirkin et al., 2025; Umbet et al., 2025). These toolkits and benchmarks, while indispensable for increasing the presence of low-resource lects, are almost always based on existing descriptions and understandings of a given lect, employing either expert knowledge or reference grammars. Their use in model assessment is vital for understanding linguistic capabilities, but it reveals little about the language itself. Some works attempt to bring evaluation closer to linguistic exploration (Bindi, 2025; Neumann, 2025), but they still primarily guide researchers through tools, concentrating attention on the qualitative evaluation of quantitative methods.

2.2 Variation in corpora

Quantitative studies of variation are almost impossible without corpora (for an overview, see Tagliamonte (2025, pp. 1–15)). However, most rely on extracting units belonging to pre-determined categories from the material. The closest analogue to a quantitative description would be *a posteriori* studies that do not assume pre-given language categories (Otheguy, 2002) and instead operate with distributional skewings that mark the presence of semantic substance expressed through specific signals (García, 1989; Diver, 2012). When fully integrated into quantitative linguistics, this approach allows for the identification and explanation of less trivial regularities and irregularities in the data.

2.3 NLP and Transcarpathian lects

Despite its significance for studies of the historical development of the Slavic clade (Ševel'ov, 1979, p. 37), the Transcarpathian group remains underrepresented in NLP, as do most Slavic territorial lects. There has been an effort to create a corpus representing their modern state[2], which resulted in the development of some language-specific tools (Scherrer and Rabus, 2019), as well as computational research (Rabus and Scherrer, 2017; Rabus, 2018; Lahjouji-Seppälä and Rabus, 2021). However, these studies operate at a regional scale, and smaller lects, such as Lemko, have not received full-scale attention.

3 Data

The description of the research data contains two subsections. The first delves into the overall characterisation of the lect under consideration; the second reports on the subcorpus used in the current study.

3.1 Lemko (Lemkian)

Lemko is a group of small territorial East Slavic lects historically spread across the territories of southern Poland and north-western Slovakia, as Figure 1 shows. It is in this territory that scholars collected the material constituting the dataset (Nakonečna and Rudnyc'kyj, 1940, p. 15). Unfortunately, after the Second World War most Lemkos living in this territory lost their homes due to discriminatory policies (Magocsi, 2015, pp. 336–338), so the material gathered at the beginning of the twentieth century may well be the latest available fragment (Baglioni and Rigobianco, 2024, pp. 1–9) of this lect.

Lemko is part of the Carpathian group (Del Gaudio, 2017, p. 78), which itself is part of the Transcarpathian area (Ševel'ov, 1979, p. 37), incorporating Hutsul, Bojko, and Central and South Carpathian lects (Del Gaudio, 2017, p. 78). There are at least three standards that roof these lects: Polish, Slovak, and Ukrainian. Modern standard Ukrainian, which stems from the Central

[2] http://www.russinisch.de/VarchoLatin2/login.php (date of access: February 12, 2026)

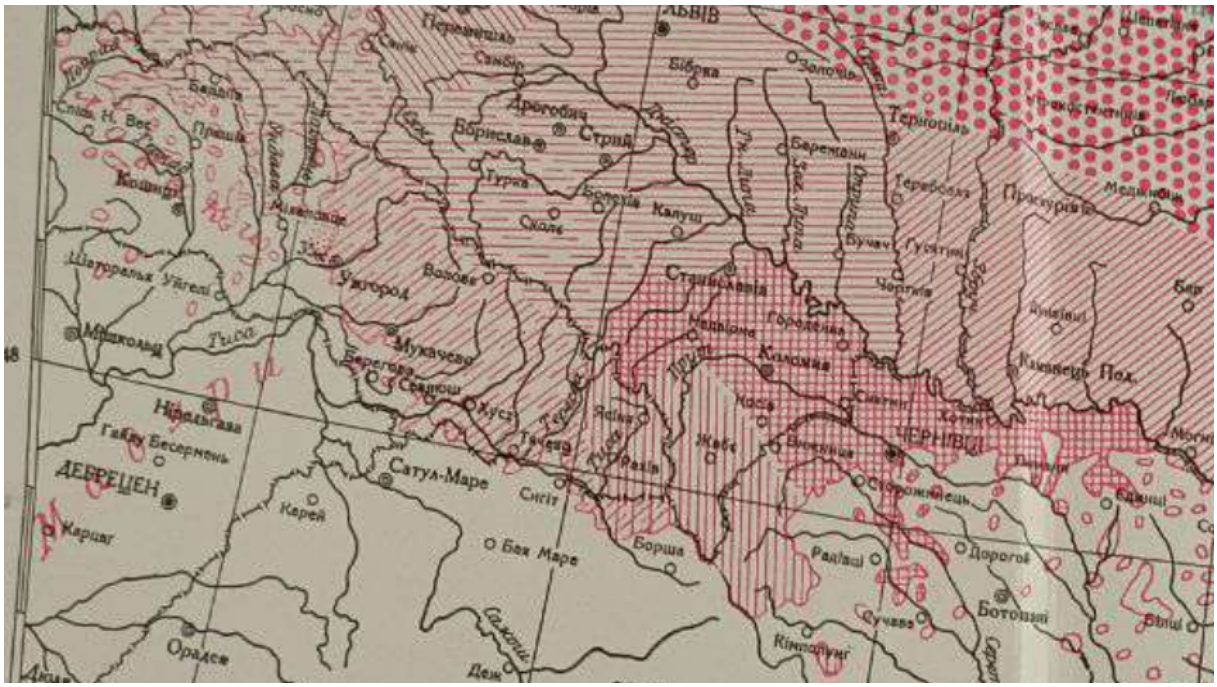

Figure 1: The territory of the Transcarpathian lects spread at the beginning of the twentieth century (Zilyns'kyj, 1933). Lemko is in the left corner, marked by rare horizontal strikes.

Dnipro group (Del Gaudio, 2017, p. 92), is the most closely related to Lemko among these standards. Nowadays, the Rusyn (micro)standard is also emerging (Dulichenko, 1981, p. 14; Magocsi, 2004). There has also been an attempt to standardise Lemko itself, resulting in the appearance of a grammar (Fontański and Chomiak, 2000). Most written or printed material in Lemko is in the Cyrillic script.

There are distinct characteristics that differentiate Lemko from neighbouring lects at all levels of the language system. For this paper, the most relevant are grammatical and lexical features.

3.1.1 Morphology

The inflection of nominal and pronominal forms in Lemko has significant peculiarities which, in the case of transfer learning, are likely to influence system performance. Among these are unique declension paradigms. For instance, the word denoting 'young girl' is *дівча* (*divča*, see Fontański and Chomiak (2000, pp. 79–80)), unlike Polish *dziewczyna* or Ukrainian *дівчина* (*divčyna*).

The instrumental singular feminine form of adjectives (and words undergoing similar inflection, such as some relative pronouns and nouns) has the *-ом* (*-om*) ending, cf. *к'отром* (*k'otr-om* 'which-FEM.INS.SG'[3]). For comparison, Polish has *-ą* (*ma l-ą* 'small-FEM.INS.SG' (Wróblewska, 2018)[4]), Ukrainian has *-ою* (*скороченою* (Kopp et al., 2023) (*skoročen-oju* 'shortened-FEM.INS.SG')), and Slovak has *-ou* (*z'ahrebsk-ou* (Zeman, 2017) 'Zagrebian-FEM.INS.SG').

The instrumental plural form is often *-ма* (*-ma*): *н'іма* (*n'i-ma* 'PRON.3PL-INS'). Slovak, Polish, and Ukrainian possess different forms: *ni-mi* (Slovak (Zeman, 2017); Polish (Wróblewska, 2018); Ukrainian *ними* (*ny-my* (Kopp et al., 2023))).

Another noteworthy feature of Lemko is the reduplicated form of determinative pronouns: *momo* (*toto*) instead of Ukrainian *це* (*ce*) (Kopp et al., 2023), Polish *tamto* (Brooks, 1975, p. 306), and Slovak *to* (Zeman, 2017).

3.1.2 Lexis

As the lect is part of an intense contact area in the Carpathian mountains, Lemko material demonstrates a significant presence of borrowings, mostly from Slovak and Polish. Slovak borrowings include words such as *вельо* (*vel'o* 'many') and *кед* (*ked* 'when, if') (Nakonečna and Rudnyc'kyj, 1940, p. 30). Material borrowed from Polish includes, among other items, *барз* (*barz* 'very') and *мераз* (*teraz* 'now') (Nakonečna and Rudnyc'kyj, 1940, p.

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

[4]See Appendix A for more detailed information on the sources of examples.

30). There are also Hungarian and German borrowings, for instance *кіральфій* (*kiral'fij* 'prince, son of king' (< Hungarian *kir'alyfi* 'id.' (Nakonečna and Rudnyc'kyj, 1940, p. 30))), but these are mostly long-integrated nouns, and any incorrect description by the model is more likely to be due to morphological reasons than to out-of-vocabulary status.

3.1.3 Syntax

Unlike Ukrainian, and like Polish and Slovak, Lemko tends to make heavy use of copular clauses of a very specific type. The subject is a determinative pronoun (in all of these languages in neuter gender form), the predicate is a noun, and the link between them is an auxiliary verb 'to be'. Examples from all of these languages are given below; the relevant parts of the sentences are in bold.

(1) Lemko (Nakonečna and Rudnyc'kyj, 1940, p. 31)

То				**сут**
T-o				**sut**
DET-NEUT.NOM.SG				**be.PRES.3PL**
вши́тки				лемки́вски
všýtk-y				lemkývsk-y
all-NOM.PL				Lemko-NOM.PL
се́ла	,	де	по	лемки́вски
sél-a	,	de	po	lemkývsky
village-NOM.PL	,	where	in	Lemko
гва́рят	.			
hvárj-at	.			
speak-PRES.3PL	.			

'**These are** all Lemko **villages**, where one speaks Lemko.'

(2) Polish (Wróblewska, 2018)

Co	**t-o**		**są**
What	**DET-NEUT.NOM.SG**		**be.PRES.3PL**
mechanizm-y		obronn-e	?
mechanism-NOM.PL		defence-NOM.PL	?

'**Are these** defence **mechanisms**?'

(3) Slovak (Zeman, 2017)

Je-∅			**t-o**
be.PRES-3SG			**DET-NEUT.NOM.SG**
naj-bežn-ějš-ia			
SUP-common-CMPR-FEM.NOM.SG			
kukuric-a	pre	priam-u	
corn-NOM.SG	for	direct-FEM.ACC.SG	
ľudsk-ú		spotreb-u	
human-FEM.ACC.SG		consumption-ACC.SG	
.			
.			

'**This is** the most common **corn** for direct human consumption.'

(4) Ukrainian (Kopp et al., 2023)

це	дуже
c-e	duže
DET-NEUT.NOM.SG	very
важлива	
važlyv-a	
important-FEM.NOM.SG	
поправка	
popravk-a	
amendment-NOM.SG	

'This is a very important amendment'

As can be seen from the examples, Lemko and Slovak use the auxiliary 'to be' (in this case in the present tense, third-person singular or plural form), whereas modern standard Ukrainian does not. Given the relatively high degree of descriptiveness in Nakonečna and Rudnyc'kyj (1940), which entails a high frequency of identificational constructions (roughly translated into English as *This is*), this discrepancy may create severe issues for the syntactic parser module, which is not trained for this type of sentence. Apart from this, however, there are no syntactic features that would characterise Lemko as a specific part of the Carpathian area.

3.2 Dataset

The case study is *LA1407*, a set of three texts in Lemko written sometime between the 1930s and the 1940s by a person who learned the lect in the settlement of Kamienka (Lemko *Камюнка*, modern Prešov Region of northern Slovakia) during childhood. The first text provides a general metacharacterisation of the lect, the second discusses traditional social gatherings, and the third is a folklore story about a devil who intended to destroy Stará Ľubovňa Castle.

The overall size of *LA1407* is 609 tokens, split into 34 predications[5], but the metainformation for each sentence is quite extensive. Nakonečna and Rudnyc'kyj (1940) provides detailed phonetic transcription, standardised transcription, and a German translation, which greatly aid modern scholarship. The format of the existing digitised version is CoNLL-U. The phonetic transcription (converted to IPA) and the German translation are provided as metadata fields for each sentence. Table 1 shows an example.

[5]For some utterances, Nakonečna and Rudnyc'kyj (1940) preserved them as sentences; others were chunked into smaller units, likely due to the need to include as much information as possible about each of them on a given page.

sent_id = LA1407.3.3
IPA_transcription = tɕʲ’ort maw prïkaz’ano
do dvan’atsʲatoj fiodïnɨ̈ v‿n’otɕʲɨ̈ rozb’ïtı z’amök //
standard_text = Чорт мав пріказа́но до дванаця́той годи́ни в но́чи розби́ті за́мок.
german_text = Es wurde dem Teufel befohlen, bis zwölf Uhr nachts das Schloß zu zerstören.

1 Чорт _ _ _ _ _ _ _ wf="Чорт"|tf="tɕʲ’ort"

(...)

Table 1: The initial digitisation of LA1407.3.3. Underscores denote the fields, obligatory for CoNLL-U format, but not yet filled with morphosyntactic tags. As the table is an illustration, it shows (for brevity considerations) only the first token. The translation of the example is *The devil had an order: before the clock strikes midnight, he should destroy the castle*.

During this study, the existing digitisation (Nakonetschna et al., 2025), which also contains information on named entities and basic vocabulary items, underwent an additional round of checks for consistency and correctness. While the tagging remained mostly intact, some normalisation was necessary. In this process, both instances of *староlюбов’ентскій* (*staroljubov’entskij* ‘of Stará Ľubovňa-ADJ’ in the standardised transcription) received the proper representation of the epenthetic *m* (*t*): in the original rendering, one instance lacked it, while the other represented *mc* (*ts*) as *ц* (*c*). This adjustment was required for an accurate visual representation of the linguistic variation.

For this study, the dataset underwent additional manual preprocessing by a linguist specialising in East Slavic languages. Some graphic variation in the standardised transcription, such as *хт’овди/хд’овди* (*ht’ovdy/hd’ovdy* ‘then’), was normalised to a single form corresponding to the phonetic transcription. Predications that were originally part of a single sentence were merged to restore the original structure and ensure correct dependency parsing. The sentences were provided with an English translation in addition to the original German one.

The final step consisted of tagging linguistic variation using a schema similar to UA-GEC (Syvokon et al., 2023). If no variation is present, the schema reproduces the sentence unchanged. Where variation occurs, the relevant part of the token receives the following tag: `{SOURCE=>INVARIANT:::variation_type=GROUP~TYPE}`, where `SOURCE` is the original segment of the token, `INVARIANT` is its normalised standard equivalent, and `GROUP/TYPE` is a variation label indicating the broader category (phonetic/morphological) and the specific subtype.

For instance, the sentence in Table 2 shows three distinct types of phonetic (`Phon`) variation: `G` (fricative/plosive velar), `NTSK` (presence or absence of the epenthetic *m* (*t*) between *н* (*n*) and *ск* (*sk*)), and `Edn` (presence or absence of prothetic *v* before *єдн* (*jedn-* ‘one-’)). This tagging is stored as a separate metadata field for each sentence in the dataset..

4 Method

This section consists of two parts. The first outlines the general methodological considerations of the current study, forming its theoretical backbone. The second provides the workflow for data annotation and the subsequent experiments that constitute the application of the theoretical principles developed in the first subsection.

4.1 How to describe a lect quantitatively?

The goal of quantitative description is to produce a model or a set of models that describe a lect as accurately as possible. The most appropriate strategy for performing this description for low-resource lects is to utilise models (pre-)trained on neighbouring higher-resource lects. In this type of exploration, it is crucial to treat the studied lect as an independent system rather than as an offshoot of a neighbouring higher-resource lect. Dialectologists have identified such biases even in qualitative research (Saenko, 2018) and have cautioned against them, arguing for describing a smaller lect as a self-contained system (*integral approach*) rather than as a deviation from a roofing standard (*differentiating approach*) (Goldin and Kryuchkova, 2011; Otheguy and Stern, 2011; Hromko, 2020).

The purpose of the subsequent analysis is to explain how the distribution of the grammatical features within a lect affects model performance and

За́мок старолюбове́{нтскі=>нскі:::variation_type=Phon~NTSK}й сто́їт ко́ло Попра́да бли́зко місто́чка Старолюбо́вні і Камю́нки, а є ма́етком {єдн=>єдн:::variation_type=Phon~Edn}о́{г=>г:::variation_type=Phon~G}о польско́{г=>г:::variation_type=Phon~G}о {ґ=>г:::variation_type=Phon~G}ро́фа.

Table 2: The example of tagging of grammatical variation in LA1407. The English translation is *The castle of Stará Ľubovňa stands on the river Poprad, near the Stará Ľubovňa city and the Kamienka village; it is a property of a Polish count.*

what this reveals about the distributional properties of this lect. A necessary component of the analysis is the combination of close reading, which selectively and thoroughly examines sections of the material to capture the full scale of variation, with distant reading, which traces a single feature across a larger body of data.

4.2 Experiment outline

As *LA1407* is a single digitised Lemko text, the current study primarily focuses on providing baseline morphosyntactic tagging. The analysis section discusses both its automatic and manual evaluation.

4.2.1 Tagging

For morphosyntactic tagging, the paper uses a modern standard Ukrainian-trained model within Stanza (Qi et al., 2020), an NLP toolkit covering a wide range of languages. The study implements only one model, as its focus is not on comparing the performance of different tools on a single dataset, but rather on examining what this specific model reveals. This is also why the paper does not employ modern generative AI models as tools, as their use introduces additional methodological variables that fall outside the scope of the present study.

The tagging process consists of three stages (PoS/morphological tagging, lemmatisation, dependency parsing), applied sequentially with intervening manual correction. This workflow reduces the number of errors propagated to subsequent stages. After preprocessing is complete, the study uses Latent Dirichlet Allocation (Blei et al., 2003) to generate topics for the text set, thereby adding a lexical layer to the analysis.

4.2.2 Evaluation

The study relies heavily on fine-grained evaluation. For part-of-speech and morphological tagging, it reports the number of errors for each tag. During the lemmatisation stage, in addition to the traditional accuracy score, the article implements string similarity measures: Levenshtein distance (Levenshtein, 1966) and Jaro–Winkler distance (Winkler, 1990). In addition to Unlabelled Attachment Score (UAS) and Labelled Attachment Score (LAS), dependency parsing utilises Unlabelled Complete Predication (UCP) and Labelled Complete Predication (LCP), which indicate whether the model identified all dependencies of the root verb and whether these dependencies received correct labels (Plank et al., 2015, p. 315).

This study slightly modifies UCP and LCP to account for cases in which the model assigns more labels than required (the original metric tests only the presence or absence of gold labels among the predicted ones). The modified version relaxes the metrics by scoring the proportion of correctly identified dependencies for each verb, rather than using a binary judgement of complete success or failure. An additional metric is the average tree edit distance (TED), which measures the number of edits required to transform a predicted tree into the gold tree (Plank et al., 2015, p. 315).

4.2.3 Qualitative exploration

Fine-grained evaluation highlights distributional skews present in the data itself or in comparison with other datasets (for instance, the initial input data). Close reading, by contrast, examines the tagging of specific forms. Its purpose is to illustrate the structural properties of Lemko, showing how different linguistic levels interact to create a distinctive combination of Transcarpathian features.

5 Experiments and Analysis

This section briefly discusses the tagging experiment results, presents the quantitative evaluation, and then explores the identified patterns. The tagging code is available in open access (Afanasev, 2026).

5.1 Tagged dataset

After the combination of manual (Section 3.2) and automatic (Section 4.2.1) tagging, the dataset

becomes substantially richer in linguistic annotation. The representation of sentence LA1407.3.3 is shown in Table 3.

The tokens also contain information on the errors that occurred during the tagging phases. The miscellaneous field `PosRapidity` denotes the quantitative measurement of errors made by the model for both part-of-speech and morphological tags (in this case, 0), which in subsequent visualisation studies is converted into a heat rate. The miscellaneous field `LemmaErrorSpots` indicates (in this case, absent) the differences between the gold lemma and the predicted lemma, while `TaggedLemma` preserves the predicted lemma.

5.2 Evaluation results

5.2.1 Morphological tagging

In evaluation terms, the model shows a moderate decline relative to the results it achieved on the test subset of the standard Ukrainian corpus on which it was trained [6]. For part-of-speech tagging, the macro F1-score is 66.78%, with recall of 68.71% and relatively higher precision of 75.88%. The exact match rate for morphological tags is 55.99%.

Table 4 shows the fine-grained evaluation results, excluding punctuation, interjections, and coordinating conjunctions (as there were no errors in these categories).

As can be seen, the categories that affect the model most are adverbs (`ADV`, which account for almost ten percent of the dataset), subordinate conjunctions (`SCONJ`, almost five percent), and verbs (`VERB` and `AUX`, more than ten percent). These errors concentrate around Satellite Cluster B (Reid, 2011, pp. 1107–1108), the verb, while the effect on the nominal group, Satellite Cluster A (adjectives `ADJ`, nouns `NOUN`, proper nouns `PROPN`, and adpositions `ADP`), is significantly smaller. The key factors here are lexical and syntactic differences. Many verbs, such as гварят (*hv'ar-jat* 'speak-`IPFV.PRES.3PL`'), are unknown to the model (standard Ukrainian would yield говорят (*hovor-jat*)), which leads to incorrect tagging. This, in turn, propagates errors down the syntactic tree, as the model misinterprets the remaining signals. This suggests that the model is unable to rely on character-level sequences within tokens that might otherwise signal a particular meaning.

The very low score for `AUX` is likely due to the absence of copular constructions in the training dataset, as mentioned in Section 3.1.3. Because the model was not adjusted to this type of signal grouping, it fails to distinguish its constituents, especially auxiliary verbs.

One of the greatest difficulties for the model is identifying tense (32.50%) and aspect (39.56%) in the verbs. Explaining these failures is more complex, but it is clear that the stanza-uk model does not adequately capture the semantic distinctions encoded in the verbal inflectional system. For instance, *жи́ют* (*žý-jut*'live-`IPFV.PRES.3SG`') receives the tag `Aspect=Perf`.

5.2.2 Lemmatisation

The model performs better in lemmatisation: the accuracy score is 75.69%. The Levenshtein (1.44) and Jaro-Winkler (87.40%) distances, however, indicate even stronger performance. This suggests that the concept of lemmatisation does not differ substantially cross-linguistically, or at least not between standard Ukrainian and Lemko. When the model makes an error, it rarely exceeds two characters and typically reflects an incorrect inflection rather than a random substitution.

The most problematic categories are, once again, verbs and auxiliaries. The accuracy score for verbs is 4.11% and for auxiliaries is 15.79%. It is noteworthy, however, that string similarity metrics for verbs are much better: a Levenshtein distance of 1.15 and a Jaro-Winkler distance of 92.04%. Errors are almost absent; the only significant issue is that the model predicts the infinitive ending as *-mu* (*-ty*), whereas in Lemko it is generally -ти (*-ti*) with some exceptions.

For auxiliaries, by contrast, all metrics indicate poor performance: the Levenshtein distance is 2.88, and the Jaro–Winkler distance is 56.19%. The model handles this class very poorly, which reflects its substantially different usage in the training dataset (see Section 3.1.3).

5.2.3 Dependency parsing

The syntax is seemingly the weakest spot of the model, as shown in Table 5.

On the surface level, the performance is acceptable. UAS and TED could have been considerably worse (given that there are sentences of length 60, an average error of 9 is substantial, but not critical). LAS and both CP metrics, however, indicate

[6] See the results of the model on the test subset here (date of access: February 12, 2026). The work uses results from the test subset as comparative material in accordance with common practice in NLP for closely related varieties (Bhatia et al., 2021; Blaschke et al., 2023; Pugh and Tyers, 2024).

sent_id = LA1407.3.3
IPA_transcription = tɕʲ’ort maw prïkaz’ano
do dvan’atsʲatoj fiodïnɨ̈ v‿n’otɕʲɨ̈ rozb’ïtɪ z’amök //
variation_text = Чорт мав пріказа́но до дванаця́той {г=>г:::variation_type=Phon~G}оди́ни в но́чи розби́{ті=>ті:::variation_type=Morph~Infinitive} за́мок.
standard_text = Чорт мав пріказа́но до дванаця́той годи́ни в но́чи розби́ті за́мок.
german_text = Es wurde dem Teufel befohlen, bis zwölf Uhr nachts das Schloß zu zerstören.
english_text = The devil had an order to destroy the castle before midnight.

1 Чорт чорт NOUN _ Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing
3 nsubj:pass _ wf="Чорт"|tf="tɕʲ’ort"|PosRapidity=0|LemmaErrorSpots=____|TaggedLemma=так

(...)

Table 3: The final digitisation of LA1407.3.3. The underscore denotes an empty field of language-specific morphosyntactic tagging (XPOS): such tagging requires additional effort that is out of the scope for this research. As the table is an illustration, it shows (for brevity considerations) only the first token.

PoS	ADV	PRON	ADJ	SCONJ	PROPN	ADP	NOUN	DET	VERB	PART	AUX
A	54.17	60.00	67.65	34.62	97.22	80.00	97.06	56.52	67.12	90.00	26.32
S	7.88	5.75	5.58	4.27	5.91	10.67	16.75	3.78	11.99	1.64	3.12

Table 4: The accuracy scores (A), % and shares (S), % (rounded to the second digit) of PoS within the dataset (excluding 100% results).

that the model does not infer syntactic categories reliably. One of the most problematic categories is reflexivity (`expl:pv`); the model never predicts it correctly. This is due to the reflexive short pronoun *ся* (*sja* ‘REFL’) behaving more freely than, for example, in standard Ukrainian, in a manner closer to Slovak or Polish.

(5) Lemko (Nakonečna and Rudnyc’kyj, 1940)
А на́ них **ся**
A ná n-ych **sja**
And on PRON.3-ACC.PL **REFL**
призіра́ют ня́ньо ,
pryzirá-jut nján’-o ,
watch-PRES.3PL father-NOM.SG ,
ма́ма , ді́до
mám-a , díd-o
mother-NOM.SG , grandfather-NOM.SG
, ба́ба ,
, báb-a ,
, grandmother-NOM.SG ,
невісти і дру́ги
nevíst-y i drúh-y
daughter.in.law-NOM.PL and other-NOM.PL
з роди́ни .
z rodín-y .
from family-GEN.SG .
’And watching them are: the father, the mother, the grandfather, the grandmother, the daughters-in-law and other family members.’

The freer behaviour of the reflexive in Lemko results from its status as a clitic that most frequently occupies the position of the second phonetic word in a phrase (Kolaković et al., 2022, pp. 22–32). As Polish and Slovak, alongside other Slavic languages (Kolaković et al., 2022, pp. 22–32), preserve this pattern, the most likely explanation is a shared archaism inherited from Proto-Slavic. While Central Dnipro lects (and, subsequently, modern standard Ukrainian) innovated toward a tighter attachment of the reflexive to the verb, Lemko and other lects of the Transcarpathian area retained the older structure. In this respect, they are closer to historical Slavic lects such as Old Church Slavonic (Polivanova, 2013, pp. 462–464).

5.2.4 Thematic modelling

The hyperparameters for the LDA model are given in Appendix B. The extracted topics are *под* (*pod* ’under’), *познати* (*poznati* ’know’), *польский* (*pol’skyj* ’Polish’), *полюбити* (*poljubyti* ’love’), *посмотріті* (*posmotriti* ’take a look’), *пост* (*post* ’fasting-NOM.SG’), *походити* (*pohodyti* ’originate’), *през* (*prez* ’through’), *премінїті* (*preminiti* ’change’), *призірати* (*pryzirati* ’watch’). Together,

Metrics	UAS	LAS	TED	UCP	LCP
Value	67.65%	52.71%	9.00	47.18%	35.57%

Table 5: UAS, LAS, UCP, LCP % and tree-edit distance for dependency parsing of LA1407 with Stanza

they summarise the three texts with considerable clarity. For instance, *под* (*pod* 'under'), *польский* (*pol'skyj* 'Polish'), *походити* (*pohodyti* 'originate') and *през* (*prez* 'through') characterise the first text, which describes the speaker and the geographical distribution of the lect. Even the functional words are thematic here, as they emphasise spatial relations between the entities expressed through the content words. The item *польский* (*pol'skyj* 'Polish') highlights the role of neighbouring Slavic languages in the development of Lemko, already visible in its syntax.

The second group of words evokes the atmosphere of a traditional gathering: *познати* (*poznati* 'know'), *полюбити* (*poljubiti* 'love'), *посмотріті* (*posmotriti* 'take a look'), *пост* (*post* 'fasting'), *призірати* (*pryzirati* 'watch'). These items relate either to the social purpose of such meetings (courtship and marriage) or to their temporal setting (summer fasting). The word *преміністі* (*preminiti* 'change') criginates from the third text and does not characterise it as directly. However, it may function as a predicate summarising the story of the devil transforming into scale grease after failing to fulfil an order. In that sense, it remains thematically appropriate.

5.3 Discussion

Through its errors, the model facilitates a description of Satellite Cluster B (the lexical verb and its adjacent elements, excluding the subject and its adjacent morphemes, Satellite Cluster A), whose structure differs substantially between Lemko and modern standard Ukrainian. Some predication types are entirely unknown to the model, while those it recognises exhibit a divergent distribution (see paragraph 5.2.3).

The analysis not only supports the existing body of qualitative research, but also highlights the Transcarpathian features of Lemko by introducing a quantitative perspective. The distributional skewings present in data affect the clause types and, consequently, roots and affixes that structure semantic space through formal markers. This organisation diverges markedly from that of modern standard Ukrainian.

One of the key features is the copular construction, widely distributed across the Transcarpathian area but absent in modern standard Ukrainian. In this respect, the Lemko lect, as part of the Transcarpathian area, places relatively strong emphasis on temporal anchoredness of identificational constructions, expressed through the explicit marking of past, present, and future forms of the verb 'to be'. In addition, aspectual marking appears less grammaticalised, as evidenced by persistent tense/aspect tagging issues, and more semantically diffuse. This may point to a grammatical system that differs not only from modern standard Ukrainian but potentially also from other Transcarpathian lects.

6 Conclusion

The paper applies Stanza to the quantitative linguistic description of the Lemko lect (East Slavic, Transcarpathian region). It identifies areal distributions in syntactic structures and lexical items that complicate tagging by models trained on neighbouring lects, most notably copular predications of the type *short determinative + auxiliary verb + noun*. The research produces a morphologically tagged dataset of Lemko from the 1920s–1930s, enriched with variation annotation and topic modelling. The tagging underwent manual verification by a linguist specialising in East Slavic languages and can therefore be considered reliable.

Future work should prioritise extending the dataset to additional Lemko texts and, more broadly, to other Transcarpathian materials from the same period. At present, much of the description remains contrastive; only the topic modelling component allows the Kamienka lect to be examined on its own terms.

For a fully adequate quantitative description, it would be preferable to develop a model designed specifically for Lemko morphosyntax. Any future topic-modelling study should incorporate stopword removal in order to obtain a clearer thematic profile of the texts. A crucial long-term objective is the development of a model that represents the lect in a more transparent and linguistically interpretable manner (Chung and Chou, 2025).

Limitations

LA1407 (Nakonečna and Rudnyc'kyj, 1940, 31–37), the primary material of this study, does not represent Lemko (or Transcarpathian lects more generally) in their entirety; a large portion of the available material is still undergoing digitisation. Moreover, LA1407 reflects the speech of a single Lemko speaker, which may influence the observed distributions.

Ethical Considerations

The data were published in printed form and have been available for research purposes for fifty to ninety years at the time of writing. Nevertheless, the metatagging has been anonymised where possible, masking speaker names in order to mitigate potential ethical issues related to historical data collection practices.

The texts contain occasional references to xenophobic behaviour and religious (primarily Christian) imagery. Reader discretion is advised.

Disclosure of Generative AI use

This study does not employ generative AI in the research process. While Stanza (Qi et al., 2020) technically belongs to the broader class of generative models in the modern colloquial meaning (the decoder models with more than a billion parameters trained on high-resource corpora), it operates at a much smaller scale and is locally reproducible. During the editing stage, the author used generative AI tools (Grammarly and OpenAI) solely for language polishing where non-native proficiency might otherwise have limited grammatical or stylistic clarity. The intellectual content of the article is entirely human-authored.

Acknowledgements

I thank the anonymous reviewers for their insightful feedback, which substantially improved the article. I am also grateful to the speakers whose recorded speech preserves the studied lects, to the scholars responsible for the original transcriptions, and to the research teams who produced the revised versions. Special thanks are due to Olha Fedorivna Mygolynets (ukr. Ольга Федорівна Миголинець, University of Uzhhorod) for her invaluable assistance with transcription systems and the phonetics of the analysed lects.

References

Lucien Adam and Victor Henry. 1880. *Arte y vocabulario de la lengua chiquita. Con algunos textos traducidos y explicados compuestos sobre manuscritos inéditos del XVIII siglo*. Maisonneuve y Cia., Paris.

Ilia Afanasev. 2026. Quantitative lect description: A case study of lemko from the field data of 1920s-1930s - supplementary material.

Maksim Aparovich, Volha Harytskaya, Vladislav Poritski, Oksana Volchek, and Pavel Smrz. 2025. BelarusianGLUE: Towards a natural language understanding benchmark for Belarusian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–527, Vienna, Austria. Association for Computational Linguistics.

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

Kushagra Bhatia, Divyanshu Aggarwal, and Ashwini Vaidya. 2021. Fine-tuning distributional semantic models for closely-related languages. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 60–66, Kiyv, Ukraine. Association for Computational Linguistics.

Beatrice Bindi. 2025. Evaluating stanza and udpipe for morphosyntactic annotation of old russian: A case study on maximus the greek. *Scripta & e-Scripta*, 25:39–60. Pages: 22. Language: English. Published by: Institute for Literature, Bulgarian Academy of Sciences.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? a study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Dorothea Frances Bleek. 1929. *Bushman Folklore*. The African Review.

Wilhelm Heinrich Immanuel Bleek and Lucy Catherine Lloyd. 1911. *Specimens of Bushman Folklore*. George Allen & Company.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Thomas J. Bolt, Jeffrey H. Flynt, Pramit Chaudhuri, and Joseph P. Dexter. 2019. A stylometry toolkit for Latin literature. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 205–210, Hong Kong, China. Association for Computational Linguistics.

Jonathan Brindle. 2017. *A dictionary and grammatical outline of Chakali*. Number 2 in African Language Grammars and Dictionaries. Language Science Press, Berlin.

Maria Z. Brooks. 1975. *Polish Reference Grammar*. De Gruyter Mouton, Berlin, Boston.

Andrey Chirkin, Svetlana Kuznetsova, Maria Volina, and Anna Dengina. 2025. RusConText benchmark: A Russian language evaluation benchmark for understanding context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1158–1170, Vienna, Austria. Association for Computational Linguistics.

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.

Michael Daniel, Nina Dobrushina, and Dmitry Ganenkov, editors. 2019. *The Mehweb language*. Number 1 in Languages of the Caucasus. Language Science Press, Berlin.

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

Salvatore Del Gaudio. 2017. *An introduction to Ukrainian dialectology*. Wiener slawistischer Almanach. Linguistische Reihe Sonderband 94. Peter Lang, Frankfurt am Main Bern Wien.

William Diver. 2012. Theory. meaning as explanation: Advances in linguistic sign theory. In Alan Huffman and Joseph Davis, editors, *Language: Communication and Human Behavior: The Linguistic Essays of William Diver*, pages 445–519. Brill, Leiden/Boston. Revised and reprinted from the 1995 original: Contini-Morava, E., & Sussman-Goldberg, B. (Eds.). (1995). *Meaning as Explanation: Advances in Linguistic Sign Theory* (pp. 43–114). Mouton de Gruyter.

M. M. Dobrotvorskii. 1875. *Ainsko-russkij slovar' [Ainu-Russian Dictionary]*. Universitetskaya tipografiya[Kazan' University Typography], Kazan'.

Domingo de Santo Tomás. 1560. *Grammatica o Arte de la lengua general de los Indios de los Reynos del Peru*. Valladolid. First grammar of the Quechua language, printed in Valladolid. Often attributed to the Dominican friar Domingo de Santo Tomás.

A. D. Dulichenko. 1981. *Slavjanskie literaturnye mikrojazyki: voprosy formirovanija i razvitija [Slavic Literary Microlanguages: Questions of Formation and Development]*. Valgus, Tallinn.

Henryk Fontański and Mirosława Chomiak. 2000. *Gramatyka języka łemkowskiego*. Śląsk, Katowice.

Erica C. García. 1989. Quantitative aspects of diachronic evolution:: The synchronic alternation between o.sp. y, alli' there'. *Lingua*, 77(2):129–149.

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.

Geoffrey D. Kimball. 1991. *Koasati grammar*. Brill, Lincoln.

Zrinka Kolaković, Edyta Jurkiewicz-Rohrbacher, Björn Hansen, Dušica Filipović Đurđević, and Nataša Fritz. 2022. *Clitics in the wild*. Number 7 in Open Slavic Linguistics. Language Science Press, Berlin.

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

Z. Lahjouji-Seppälä and A. Rabus. 2021. A robust approach to variation in carpathian rusyn: Resampling-based methods for small data sets. *Jazykovedný časopis*, 72(2):603–617.

Thomas Lehman. 1989. *A grammar of modern Tamil*. Number 1 in Pondicherry Institute of Linguistics and Culture publications. Pondicherry Inst. of Ling. and Culture, Pondicherry.

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.

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

Paul Robert Magocsi. 2004. Jazykovyj vopros [the language question]. In Paul Robert Magocsi, editor, *Rusyns'kyj jazyk [The Rusyn Language]*, Najnowsze dzieje języków słowiańskich, pages 39–66. Uniwersytet Opolski, Opole.

B. Munkácsi. 1894. *A vogul nyelvjárások szóragozásának ismertetve [Description of the Conjugation of Vogul Dialects]*. Budapest.

Hanna Nakonetschna, Jaroslau Rudnyćkyj, and Ilia Afanasev. 2025. Computer-assisted study of historical lemkian (transcarpathian ukraine) lects: basic vocabulary approach - supplementary material 1 (dataset).

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.

Saudah Namyalo, Alena Witzlack-Makarevich, Anatole Kiriggwajjo, Amos Atuhairwe, Zarina Molochieva, Ruth Gimbo Mukama, and Margaret Zellers. 2021. *A dictionary and grammatical sketch of Ruruuli-Lunyala*. Number 5 in African Language Grammars and Dictionaries. Language Science Press, Berlin.

Vladimir Neumann. 2025. Effektiver einsatz von nlp-methoden am beispiel des codex suprasliensis [effective use of nlp methods using the example of the codex suprasliensis]. *Scripta & e-Scripta*, 25:79–100. Pages: 22. Language: German. Published by: Institute for Literature, Bulgarian Academy of Sciences.

Toshiki Osada. 1992. *A reference grammar of Mundari*. Tokyo Univ. of Foreign Studies, Inst. for the Study of Languages and Cultures of Asia and Africa (ILCAA), Tokyo.

Ricardo Otheguy. 2002. *Saussurean Anti-Nomenclaturism in Grammatical Analysis: A Comparative Theoretical Perspective*, pages 373–403. John Benjamins Publishing Company.

Ricardo Otheguy and Nancy Stern. 2011. On so-called spanglish. *International Journal of Bilingualism*, 15(1):85–100.

Amalie Brogaard Pauli, Maria Barrett, Ophélie Lacroix, and Rasmus Hvingelby. 2021. DaNLP: An open-source toolkit for Danish natural language processing. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 460–466, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

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.

A. K. Polivanova. 2013. *Staroslavjanskij jazyk: Grammatika. Slovari [Old Church Slavonic Language: Grammar. Dictionaries]*. Universitet Dmitrija Pozharskogo, Moscow.

Robert Pugh and Francis Tyers. 2024. Experiments in multi-variant natural language processing for Nahuatl. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 140–151, Mexico City, Mexico. Association for Computational Linguistics.

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 rusynistikÿ na Slovakiji [20 Years of University Rusyn Studies in Slovakia]*, pages 139–151. Prešovská univerzita v Prešove, Prešov.

Achim Rabus and Yves Scherrer. 2017. Lexicon induction for spoken Rusyn – challenges and results. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 27–32, Valencia, Spain. Association for Computational Linguistics.

Nathanaël Carraz Rakotonirina, Corentin Kervadec, Francesca Franzon, and Marco Baroni. 2025. Evil twins are not that evil: Qualitative insights into machine-generated prompts. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 48–68, Suzhou, China. Association for Computational Linguistics.

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

Evelina Rennes, Marina Santini, and Arne Jönsson. 2022. The Swedish simplification toolkit: Designed with target audiences in mind. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 31–38, Marseille, France. European Language Resources Association.

Antonio Ruiz de Montoya. 1640. *Arte y vocabulario de la lengua guaraní*. Madrid, Madrid. Published as a quarto.

M. N. Saenko. 2018. Netochnosti v opisanii semantiki, vyzvannye vosprijatiem dialektnoj leksiki skvoz' prizmu literaturnogo jazyka: neskol'ko primerov [inaccuracies in the description of dialect lexis semantics, caused by literary language interference. a few examples from the east slavic dialect dictionaries]. In L. Ė. Kalnyn', editor, *Issledovanija po slavjanskoj dialektologii 19–20. Slavjanskie dialekty v sovremennoj jazykovoj situacii. Dialektnyj slovar' kak sposob issledovanija slavjanskix dialektov [Studies in Slavic dialectology 19–20. Slavic dialects in the modern language situation. Dialect dictionary as a method of studying Slavic dialects]*, pages 218–222. Institut slavjanovedenija RAN, Moscow.

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

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Sali A. Tagliamonte. 2025. *Analysing Sociolinguistic Variation*. Cambridge University Press.

Lena Terhart. 2024. *A grammar of Paunaka*. Number 7 in Comprehensive Grammar Library. Language Science Press, Berlin.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Sanzhar Umbet, Sanzhar Murzakhmetov, Beksultan Sagyndyk, Kirill Yakunin, Timur Akishev, and Pavel Zubitski. 2025. KazBench-KK: A cultural-knowledge benchmark for Kazakh. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 38–57, Vienna, Austria. Association for Computational Linguistics.

Ferdinand Johann Wiedemann. 1884. *Grammatik der Syrjänischen Sprache mit Berücksichtigung ihrer Dialekte und des Wotjakishen [Grammar of the Syrjän Language with Consideration of its Dialects and of the Wotjak]*. Russian Academy of Sciences, St. Petersburg.

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.

Maurice L. Zigmond, Curtis G. Booth, and Pamela Munro. 1991. *Kawaiisu : a grammar and dictionary, with texts*. Number 119 in Univ. of California publications. Linguistics. Univ. of California Press, Berkeley, CA.

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.

Ludwig Zukowsky. 1924. Beitrag zur kenntnis der säugetiere der nördlichen teile deutsch-südwestafrikas [contribution to the knowledge of the mammals of the northern parts of german south west africa]. *Archiv für Naturgeschichte*, 90(A, 8):43–139.

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

A Example sources

The modern standard Ukrainian examples are from *dev* branch of (Kopp et al., 2023), date of access: February 12, 2026. The Polish examples are from *dev* branch of Wróblewska (2018), date of access: February 12, 2026. The Slovak examples are from *dev* branch of (Zeman, 2017), date of access: February 12, 2026.

B LDA hyperparameters

Parameter	Value
seed	1590
num_topics	10
alpha	auto
epochs	300
passes	500
random_state	0
taken topics	2 – 9

Table 6: LDA model hyperparameters. Parameter "taken topics" denotes the topics selected from the result of topic modelling in the order that the model yielded.