DOI: https://doi.org/10.66260/SeS2026IA

SCRIPTA
& e- SCRIPTA

e-SCRIPTA

26 / 2026

Heatmap-based visualisation of the linguistic polymorphism in Transcarpathian East Slavic

Ilia Afanasev

Abstract: The study examines the complex phenomenon of language polymorphism (an umbrella term necessary in cases where there is no clear understanding of the source of variation and/or change: an occasional error, the influence of the native lect of the data collector, or other, more prototypical causes) through the lens of morphosyntactic tagging. It introduces a new visualisation method, based on heatmaps, that facilitates both a bird's-eye view of the data and closer inspection.

The main dataset of the research is TransVar, a historical corpus of Transcarpathian East Slavic (Ukrainian small territorial lects[1] recorded in the early twentieth century: Bojko, Central Transcarpathian, Hutsul, and Lemko[2]). The results of its morphosyntactic tagging with Stanza constitute the case study.

The analysis, conducted with the help of visualisation methods, identifies the primary source of tagger errors: copular clauses. It demonstrates how a single non-standard structure may significantly reduce tagger performance across all morphosyntactic levels. The prospects for future research include the collection of additional material and the development of more fine-grained annotation techniques.

Keywords: visualisation; Ukrainian; Transcarpathian; Lemko; heatmap; language variation and change

[1] The neutral term *lect* here denotes a single variety of a language, represented by a group of individual repertoires that share a sufficiently large number of features to constitute a single entity based on personal (idiolectal), social (sociolectal), territorial (dialectal), or political (standard) criteria (Campbell 2013: 191).

[2] Bojko, Hutsul and Lemko are denotations preferred by native speakers both in the modern day and historically: <https://uwr.edu.pl/en/lemkos-who-are-they/> (accessed 04.06.2025). There are on-going attempts to propose a single codified variant for all of these lects (Rusyn) and the existence of their single and exclusive ancestor (a sister of Proto-Ukrainian, Proto-Russian and Proto-Belarusian); for the discussion, refer to Del Gaudio (2017: 103–105).

1. Introduction

Corpus-based studies have typically used transformed data as the main visualisation method, presenting statistics of the corpus rather than the corpus itself (Radchankava 2025: 115–120). This work aims to amend that practice by presenting a visualisation of the continuous linguistic signal, with skewings detected within it, following the radical functional approach of the Columbia School of Linguistics (Diver 2012: 451–452).

The research takes inspiration from biological studies, which have a long tradition of visualising character sequences according to the rate of change within them. These studies represent generalised DNA sequences of the studied population as heatmaps, with regions containing differences shared by a significant proportion (>10%) of population members highlighted in colour (the greater the variation, the closer the colour is to red; for an example of a biological study that utilises heatmaps, see, among others, Yirgu et al. (2023: 3)).

The presented research concentrates on linguistic skewings that constitute significant differences between closely related lects. The article proposes for them the term *language polymorphism markers*, i.e. signals of linguistic variation and change that include not only reflections of regular processes in a speech community, but also non-standard structures otherwise deemed to be errors, and influences that arise in the course of textual transmission. The term *polymorphism* is chosen in analogy with the term *single nucleotide polymorphism markers* (SNP markers or SNPs), which designate parts of the genome that represent differences present in a substantial proportion of members of a population (Yirgu et al. 2023: 2).

The study employs TransVar, a morphosyntactically tagged corpus of Transcarpathian East Slavic. There is a high degree of variation and change between these lects, within them, and in relation to other Ukrainian lects (Del Gaudio 2017: 64–85), which makes the corpus particularly suitable for applying the proposed methods. In addition, the study aims to increase the visibility of this group in Natural Language Processing (NLP), where it is generally underrepresented, with some notable but rare exceptions (Rabus and Šymon 2015; Kyslyi et al. 2026).

2. TransVar: a Corpus to Study Language Variation and Change

The purpose of the TransVar corpus is to provide a solid basis for the study of language polymorphism in Transcarpathian East Slavic (in the early twentieth c.), based on the material of Nakonečna and Rudnyc'kyj (1940), Kuraszkiewicz (1963), and Zheguc (2001). Transcarpathian East Slavic consists of four key subunits: Bojko, Central Transcarpathian, Hutsul, and Lemko. All these are generally considered to

constitute an areal group of large territorial lects within the Ukrainian continuum, with the most likely common ancestor being Proto-Ukrainian (Kuraszkiewicz 1963: 67–72; Zilyns'kyj 1933: 8–10; Ševel'ov 1979: 37; Del Gaudio 2017: 64–85).

Currently, the geographical distribution of these lects is much more limited than historically, and their areal unity has largely been destroyed. During the twentieth c., their speakers experienced policies of forced resettlement (Magocsi 2015: 336–338). The material collected by scholars in the 1920s and 1930s is therefore even more valuable now, as it makes it possible to observe Transcarpathian lects in their form before extralinguistic circumstances disrupted the continuum. Figure 1 shows the distribution of the lects in the early twentieth c.

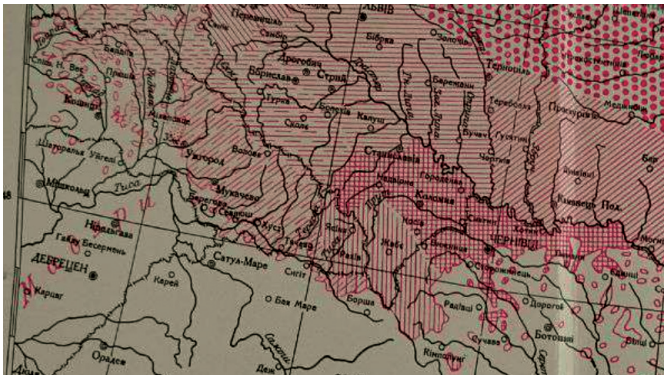

Figure 1. Transcarpathian lects in the early twentieth century (Zilyns'kyj 1933). Lemko is the westernmost lect (sparse horizontal shading), Bojko is located closer to the centre of the area (denser horizontal shading), the Central Transcarpathian lects lie to the south (diagonal shading), and the easternmost area corresponds to the Hutsul territory (dense vertical shading).

The corpus has the following tagging schema:

1. At least three layers of raw text representation are provided for each sentence. The first two layers consist of an International Phonetic Alphabet (IPA) transcription and a standardised transcription that uses Modern Standard Ukrainian orthography but reflects the key features of the studied lects (for instance, the lack of the *o* > *i* change in newly closed syllables in Lemko (Nakonečna and Rudnyc'kyj 1940: 7–8); cf. *nócma* (IPA [pˈost-ɑ] 'fast-GEN.SG') instead of Modern Standard Ukrainian *nícma* (IPA [pjˈist-ɐ] 'id.')). The third layer consists of an English translation.

2. Morphosyntactic annotation in the Universal Dependencies format (de Marneffe et al. 2021): part-of-speech tags, morphological features, lemmatisation, and dependency parsing tags.

3. Information on specific lexical categories: a basic vocabulary list, named entities (including, among other categories, settlements, personal names, and geographical features), and topic words, i.e. words especially characteristic of a particular text.

4. Variation annotation: a modified version of the standardised transcription that contains information about language polymorphism. The format is closest to that of the UA-GEC (Syvokon et al. 2023) corpus. If polymorphism is not present, the annotation reproduces the text as is. If polymorphism is present, the relevant part of a token (or a group of tokens) receives the following annotation: {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 or morphological) and the specific subtype. For instance, the aforementioned *nócma* is annotated as follows: *n{ó=>í::variation_type=Phon~NewClosedStress}cma.* This tag denotes phonetic polymorphism in the position of stressed vowels in newly closed syllables.

5. Metadata: anonymised information about annotators, transcribers, and speakers.

The annotation schema is rather rich, and there is no format that includes all these tags in a form equally readable for both machines and humans. The solution is to use the CoNLL-U format of UD corpora, with lexical tags represented in the *misc* field and variation annotation constituting an additional meta-annotation layer. For metadata representation, the corpus includes a separate .csv file. An example of sentence annotation (for reasons of space, only the first token is provided) is given below.

sent_id = LA1407.2.9
IPA_transcription = ɑ ‿ nˈɑ ‿ nïx ça prïzɪrˈɑjut / nʲˈɑno mˈɑmɑ dʲˈido bˈɑbɑ / nevˈistï ˈɪ drˈuɦï z rodˈïnï //
standard_text = А на́ них ся призира́ют ня́ньо, ма́ма, ді́до, ба́ба, неві́сти і дру́ги з роді́ни.
variation_text = А на́ них ся призира́ют ня́ньо, ма́ма, ді́до, ба́ба, неві́сти і дру́{г=>г:::variation_type=Phon~G}и з роді́ни.
german_text = Und ihnen schauen zu: der Vater, die Mutter, der Großvater, die Großmutter, die Schwiegertöchter und andere Familienmitglieder.
english_text = And watching them are: the father, the mother, the grandfather, the grandmother, the daughters-in-law and other family members.
1 A a CCONJ _ _ 5 cc _ wf="A"|tf="ɑ"|PosRapidity=0|LemmaErrorSpots=_|TaggedLemma=a

The corpus is available in an open-access format. It is currently published as a dataset on Zenodo[3]. Table 1 summarises the key characteristics of TransVar.

Table 1. Key characteristics of the TransVar corpus.

Parameter	Value
Number of lects	3
Number of speakers	3
Number of texts	9
Number of sentences	90
Number of data collectors	2
Number of annotators	1
Token count	1949

The study suggests three possible ways to annotate polymorphism, each illustrating the degree of automation achievable at the current stage of dataset development. The first annotation model is language variation and change annotation. While it seems possible to utilise an architecture similar to those used in grammatical error detection[4], the low-resource nature of Transcarpathian lects currently makes this approach very difficult, given that neural architectures require a significant amount of data (Syvokon et al. 2023). An additional problem is regularity: signs of language variation and change are usually observed across different lects (including idiolects within a small territorial lect), so detecting them with a grammatical error detection tool may be problematic (Campbell 2013: 195–197). This is why, in TransVar, language variation and change annotation is performed manually.

The next two annotation types are derived from a comparison between manually checked morphosyntactic tags in TransVar and the raw output of Stanza (Qi et al. 2020). The first facilitates the assessment of part-of-speech and morphological tagging. The parameter *PosRapidity* (assigned in the *misc* field, the tenth column of a token row in the UD format) is a counter of errors in part-of-speech and morphological tagging. The more tags the model gets wrong, the higher the counter. The best-case scenario would be to identify the subtoken (or subtokens) responsible for the incorrect

[3] A link to the current version of the corpus used in the study: https://doi.org/10.5281/zenodo.19158681 (accessed 29.03.2025).

[4] As both represent some kind of language variation, not necessarily wrong, even if not coinciding with a standard.

tagging; however, this is rarely possible without extensive human analysis. Thus, the study simply records the number of errors for each token.

The last annotation type is fully automatic. It enables the investigation of differences in the structure of Modern Standard Ukrainian, on which the Stanza model was trained, and Transcarpathian lects through the analysis of lemmatiser errors. The annotation algorithm is as follows: the model searches for the longest common substring between the gold lemma and the predicted lemma and marks every symbol of the predicted lemma that is not part of this substring as erroneous. It then stores the results in the *LemmaErrorSpots* parameter of the *misc* field. The key difference between this type and the second type is that this type facilitates the precise localisation of errors at the subtoken level, while the second type cannot point to the exact subtoken that caused the error in tagging.

This list is far from exhaustive. There may be additional types of tagging that highlight further aspects of language polymorphism. This study, however, aims to present a case study rather than an exhaustive description and therefore restricts itself to three main approaches to polymorphism annotation.

3. Language Polymorphism Tagging and Visualisation

The tagger used is Stanza trained on Modern Standard Ukrainian, a neural pipeline that integrates different modules, from tokenisation to dependency parsing, and is designed for full morphosyntactic tagging of text (Qi et al. 2020). The creation of TransVar included manual tokenisation and manual checking of tagging results by a linguist with expertise in Slavic languages, in order to make subsequent processing stages more efficient. Thus, the lemmatiser used gold part-of-speech and morphological tags, and dependency parsing relied on gold morphological tags.

The key idea of the proposed visualisation technique is to obtain a bird's-eye view of the errors made by the tagger and, through this, to identify the linguistic peculiarities of the lect under study. The visualisation takes the form of a heatmap, with each sentence constituting a single row and each cell corresponding to a symbol in the sentence, coloured according to its heat level. The heat level is a visual representation of how many markers of language polymorphism a particular cell and its closest neighbours contain. The motivation for including closest neighbours in the heat range was that much linguistic variation and change, especially at the phonetic level, is contextual (Campbell 2013: 23). Thus, it is crucial to understand not only which characters tend to change, but also which contexts (in a technical sense, neighbouring characters to the left and right) tend to trigger these changes more frequently than others.

Consider, for example, sentence LA1407.3.9 (Lemko lect): *За́мок старолюбове́{нтскі=>нскі::variation_type=Phon~NTSK}й сто́їт ко́ло Попра́да бли́зко місто́чка Старолюбо́вні і Камю́нки, а є маєтком {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* (IPA [zˈɑmok stˌɑrolʲubovˈentskɪj stˈojit kˌolo poprˈɑdɑ blʲˈɪsko // mistˈo͡ʂʲkɑ stˌɑrolʲubˈownʲi // ɪ kɑmjˈuŋkɤ̯ / ɑ je mɑjˈetkom jedˈnoɦo polʲskˈoɦo grˈofɑ //] ‘Starolubovnja Castle is located next to Poprad, near the town of Starolubovńa and Kamjunka, and is owned by a Polish count.’). It exhibits three types of polymorphism: the epenthetic [t] between [n] and [s], which may be present or absent; lexicalised differences between the voiced velar plosive [g] and the voiced uvular fricative [ɦ]; and the prothetic consonant in the word for ‘one’ (NOM.SG *jˈeden*; texts may contain either forms with [j] as prothesis or [w]). Figure 2 visualises all three types simultaneously.

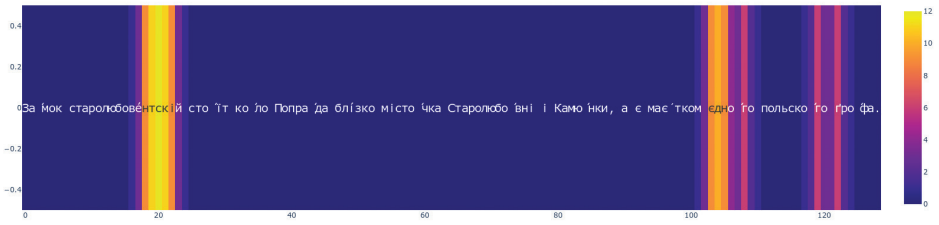

Figure 2. Heatmap-based visualisation of language polymorphism.

Here, the visualisation highlights the locations in the text where language polymorphism is most densely concentrated. These constitute evidence of contact between Lemko and West Slavic varieties, as well as toponyms and loanwords. A preliminary conclusion could be that their adaptation in Lemko has not yet fully stabilised. Although a qualitative study could yield similar results, visual representation facilitates faster and more efficient pattern recognition, as well as better generalisation.

4. Results and Discussion

This section focuses on errors produced by the morphological tagger and the lemmatisation module for document LA1407 (Lemko). The code and data used are openly accessible. Figure 3 illustrates errors in part-of-speech tagging, while Figure 4 highlights errors produced by the lemmatisation module.

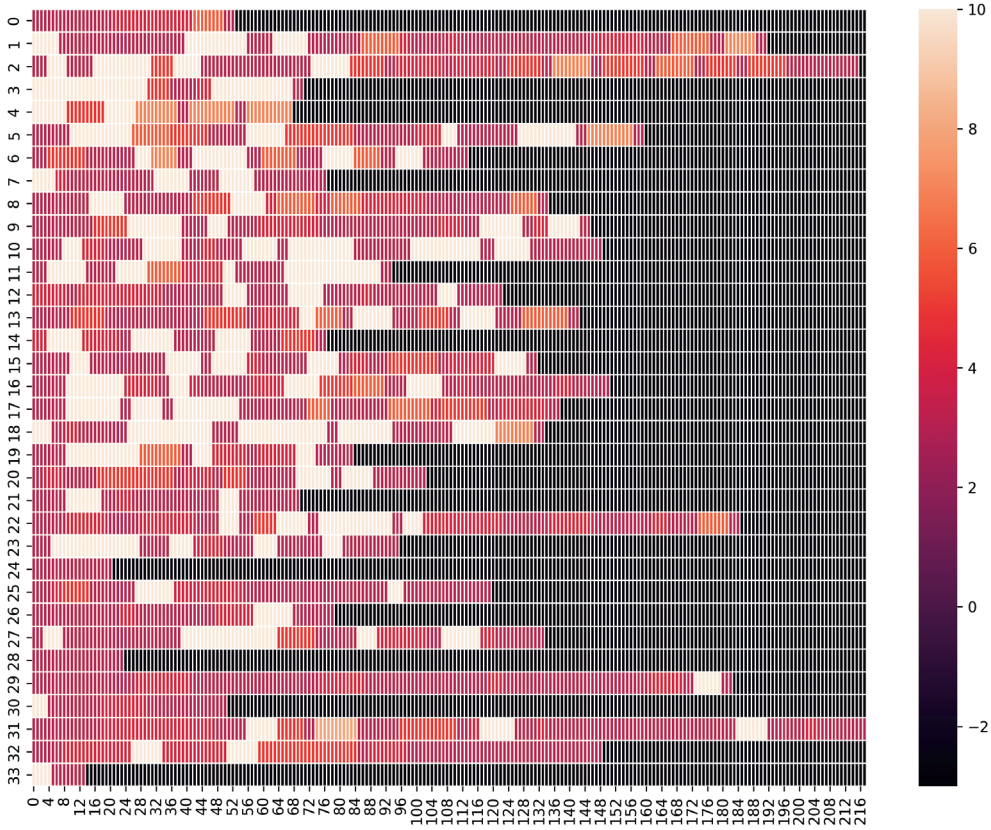

Figure 3. Visualisation of errors produced by the Stanza morphological tagging module (text LA1407).

Figures 3 and 4 represent the text in slightly different ways. Figure 3 shows large heat spots that start and end abruptly, while Figure 4 highlights longer spans of heat that are less intense. This is due to the annotation schema: morphological errors accumulate the heat parameter for the entire word, whereas lemmatisation errors highlight only the erroneously generated parts of lemmas.

The overall pattern of error distribution, however, is rather similar. The middle part of the document, along with the initial parts of sentences at the beginning, is the brightest in both cases. A closer examination of these regions reveals a key factor: verbal morphology that differs substantially between Lemko and Modern Standard Ukrainian, which serves as training data for the system. Lemmatisation accuracy is strongly affected because the infinitive ending in Lemko is <-ті> rather than <-ти>. However, this is not the only instance of polymorphism. The centres of

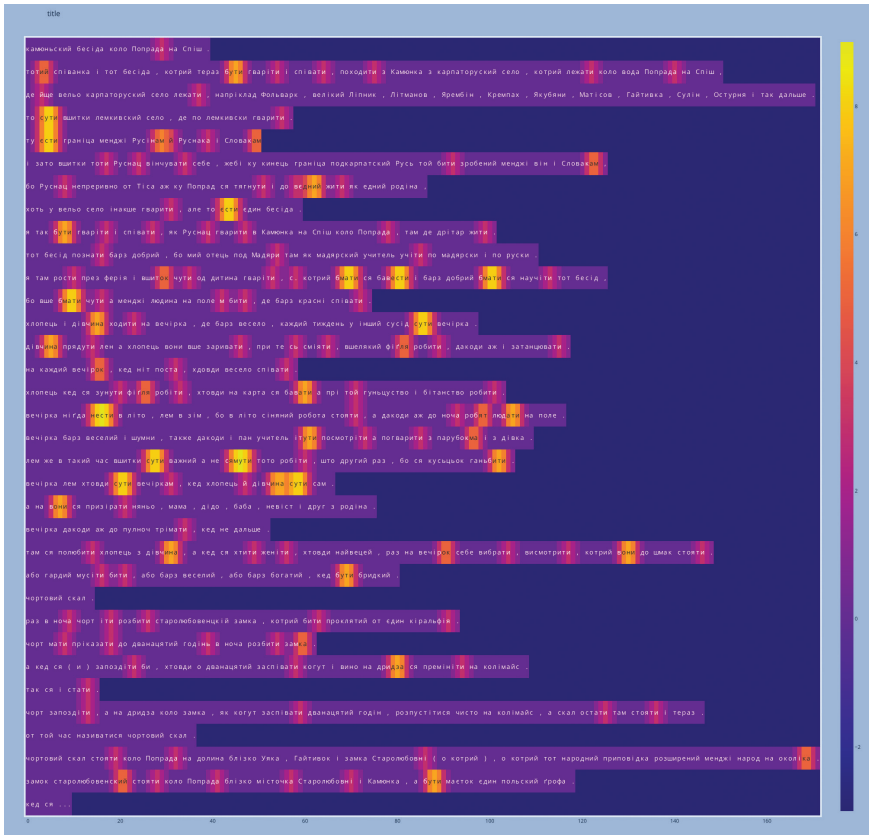

Figure 4. Visualisation of errors produced by the Stanza lemmatisation module (text LA1407).

heat concentration are the auxiliary forms *sut* 'be.PRES.3PL', and most of the heat covers copular clauses (see Moro (2017) for an introduction), for instance sentence LA1407.1.4: То сут вши́тки лемки́вски се́ла, де по лемки́вски гва́рят (IPA [to sut fʂˈɯtkɣ lemkˈöwskɣ sˈelɑ de po lemkˈöwskɣ fiwˈɑrjɑt//] 'These are all Lemko villages, where one speaks Lemko'). Statistical measures support the hypothesis. Preliminary measurements of Spearman's coefficient and Pearson's coefficient between the heat of part-of-speech tagging errors and the heat of lemmatisation errors, implemented using SciPy (Virtanen et al. 2020), show a statistically significant correlation, with p-values equal to *2.000925336400395e-11* (<0.001) and *4.298090487452403e-14* (<0.001) respectively.

The reason is the absence of corresponding structures in the UD_Ukrainian-IU corpus. The system simply does not recognise this type of token sequence and therefore

tags it incorrectly. At the morphological level, this leads to incorrect assignment of morphological categories. Since languages generally have only one copula, there is no direct analogy from which the lemmatiser can derive an appropriate model, which results in reduced performance at this level as well. The syntactic structure is likewise unfamiliar to the dependency parser, further affecting its performance.

Thus, the visualisation demonstrates the bidirectional effect of structures that differ between a higher-resource source language and a lower-resource target language when the latter is processed via transfer learning. It captures both top-down cascading effects, where non-standard syntactic structures affect morphological tagging and lemmatisation, and bottom-up accumulation of errors that negatively impact dependency parsing.

5. Conclusion

The study demonstrates the effectiveness of visualisation techniques for representing language polymorphism in Transcarpathian East Slavic lects. It enables faster detection of the structure that differs most between Modern Standard Ukrainian (the source lect for tagging) and Transcarpathian East Slavic (the target lect): copular clauses of the type *demonstrative pronoun + auxiliary verb + noun*. The study also shows that language polymorphism is not always reducible to a single factor and that tagging errors may provide insights into linguistic structure.

The visualisation of morphological tagging errors reveals remaining limitations of the method. The main drawback is that, without human annotation, it is not always possible to examine and adequately explain polymorphism in either the source or target data.

Future work includes extending the analysis to other lects in the corpus (Bojko, Hutsul, and Central Transcarpathian) using the same methods. The methodology itself should also be further developed by making the annotation schema for morphological tagging and language variation more transparent and more easily processed by automatic methods.

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 Myholynets' (ukr. Ольга Федорівна Миголинець, University of Uzhhorod) for her invaluable assistance with transcription systems and the phonetics of the analysed lects.

REFERENCES

Sources

Kuraszkiewicz 1963: Kuraszkiewicz, Władysław. *Zarys dialektologii wschodniosłowiańskiej z wyborem tekstów gwarowych*. 2nd ed. Warszawa: Państwowe Wydawn. Naukowe, 1963.

Nakonečna and Rudnyc'kyj 1940: Nakonečna, Hanna, and Jaroslav Bohdan Rudnyc'kyj. *Ukrainische Mundarten: Südkarpatoukrainisch.* (Lemkisch, Bojkisch und Huzulisch). [*Ukrainian dialects: South Carpathian Ukrainian.* Lemko, Bojko and Hutsul]. Berlin: Otto Harrassowitz, 1940.

Zheguc 2001: Zheguc, Ivan. *Vybrani teksty z hucul's'koho hovoru v Zakarpatti* [*Selected texts from the Hutsul dialect in Transcarpathia*]. Munich, 2001.

Literature

Campbell 2013: Campbell, Lyle. *Historical Linguistics: An Introduction*. 3rd ed. Edinburgh: Edinburgh University Press, 2013.

Del Gaudio 2017: Del Gaudio, Salvatore. *An Introduction to Ukrainian Dialectology*. Frankfurt am Main: Peter Lang, 2017. (*Wiener Slawistischer Almanach. Linguistische Reihe Sonderband* 94).

de Marneffe et al. 2021: de Marneffe, Marie-Catherine, Christopher Manning, Joakim Nivre, and Daniel Zeman. "Universal Dependencies." *Computational Linguistics* 47/2 (2021), 255–308.

Diver 2012: Diver, William. "Theory." In *Language: Communication and Human Behavior: The Linguistic Essays of William Diver*. Leiden: Brill, 2012, 444–519.

Kyslyi et al. 2026: Kyslyi, Roman, Artem Orlovskyi, Pavlo Khomenko, Bohdan Onyshchenko, and Zakhar Guzii. "Building ASR Resources for the Hutsul Dialect of Ukrainian." In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*. Rabat: Association for Computational Linguistics, 2026, 186–195.

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

Moro 2017: Moro, Andrea. "Copular Sentences." In *The Wiley Blackwell Companion to Syntax*. Second Edition, 1–23. John Wiley & Sons, 2017.

Qi et al. 2020: Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. "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*. Edited by Asli Celikyilmaz and Tsung-Hsien Wen. Association for Computational Linguistics, 2020, 101–108.

Rabus and Šymon 2015: Rabus, A., and A. Šymon. "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 *Rusyn'skyj literaturnyj jazyk*

na Slovakiji. 20 rokiv kodifikaciji [*The Rusyn Literary Language in Slovakia. 20 Years of Codification*]. Edited by Kvetoslava Koporová, 40–54. Prešov: Prešov University Publishing, 2015.

Radchankava 2025: Radchankava, Liudmila. "Die thematisch-fokussierende komplexe Präposition v lice. Eine diachrone Analyse mithilfe von BERT." *SeS* 25 (2025), 101–126.

Ševel'ov 1979: Ševel'ov, Jurij Volodymyrovyč. *A historical phonology of the Ukrainian language*. Heidelberg: Winter, 1979.

Syvokon et al. 2023: Syvokon, Oleksiy, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. "UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language." In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*. Edited by Mariana Romanyshyn. Association for Computational Linguistics, 2023, 96–102.

Virtanen et al. 2020: Pauli Virtanen, and SciPy 1.0 Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17/3 (2020), 261–272. DOI: 10.1038/s41592-019-0686-2.

Yirgu et al. 2023: Yirgu, M., M. Kebede, T. Feyissa, et al. 2023. "Single nucleotide polymorphism (SNP) markers for genetic diversity and population structure study in Ethiopian barley (*Hordeum vulgare* L.) germplasm." *BMC Genom Data* 24/7 (2023).

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

About the Author …

Ilia Afanasev is a PhD student at the University of Vienna and research fellow in Columbia School of Linguistics. His research interests include Slavic dialectology, language variation and change, corpus linguistics, and computer-assisted language studies. Email: ilia.afanasev.1997@gmail.com; ORCID: 0000-0002-9169-6829.