TransVar – the Corpus for Variation and Change Study of the Historical Transcarpathian lects

Anonymous submission

Abstract

The paper introduces TransVar – the corpus of the historical Transcarpathian lects (the first half of the XX century, the territories of modern Ukraine, Poland, Slovakia, and Romania). The corpus contains data from Lemkian, Bojkian and Huzulian small territorial lect groups. It is crucial for studies of the people of these territories, who witnessed forceful deportation from their homeland in the 1940s – 1950s, soon after the recordings were made (1920s – 1930s). The article also provides a brief overview of their linguistic properties, as evident in the material.

The corpus is morphosyntactically tagged. It contains data on part-of-speech, morphological features, lemmata and syntactical dependencies. The study stresses the crux of manual analysis of the errata made in an automatic tagging phase for further improvement. The supplementary information includes named entities encountered in the text and the basic vocabulary. All the texts are accompanied by metalinguistic information, required for the sociolinguistic study.

After the analysis of the current stage of the corpus creation, the article outlines further research prospects. Apart from more thorough manual annotation, one of the prospects is to add English translation with the purpose of making the material more accessible to scholars without a background in Slavic studies.

Keywords: Transcarpathian, sociolinguistics, Universal Dependencies, low-resourced language modelling, part-of-speech tagging, lemmatisation, dependency parsing, qualitative analysis

1. Introduction

Historical low-resourced small territorial lects[1] are probably one of the most understudied types of resources in computational linguistics (Miletić and Siewert, 2023). The material is quite low-resourced and highly non-standard, lacking normalisation. This makes it especially challenging for NLP tools in comparison even to the modern non-standard varieties (Piotrowski, 2012). Still, there is a significant amount of resources that contain texts produced by the speakers of these lects. This article considers the material gathered in the Transcarpathian region in the XX century (the territories of modern Ukraine, Poland, Slovakia, and Romania) by several groups of scholars and published either concurrently (Nakonečna and Rudnyc'kyj, 1940) or afterwards (Kuraszkiewicz, 1963; Zheguc, 2001), and presents a corpus of it. The overview of the resources used to gather the corpus is in section 4.

All the lects are East Slavic[2]. They belong to the three main groups, Bojkian, Lemkian, and Huzulian (section 3 describes their linguistic features and position within the Slavic clade). These lects have never been a subject of a quantitative variationist study due to the absence of an open-access corpus, the gap this study intends to close (section 2 contains a short overview of the variationist studies of the East Slavic languages, including the Transcarpathian ones, with the help of NLP tools).

The corpus contains part-of-speech and morphological tags, along with lemmatisation and dependency parsing. There are also specific tags for basic vocabulary items and named entities that facilitate historical study. Each text has a set of extralinguistic tags (or metadata tags) containing the information on the speaker and the dialectologist who performed the recording, as well as the resource itself. Section 5 outlines the corpus design, the tagging schema, and the annotation tools. It also discusses the importance of shared lexicons for NLP in low-resourced settings. Section 6 puts this analysis into the perspective of possible future improvements.

1.1. Contribution

The study creates a new resource for studying the historical Transcarpathian lects from both the quantitative and qualitative standpoints, a morphosyntactically tagged corpus in the format of

[1] The article denotes any given language variety (idiolect, doculect, dialect, regiolect, sociolect, standard) as *lect*. This choice enables the elimination of overly hierarchical dichotomies, namely, **X** `is a dialect of` **Y**, which imply some degree of **X** being a lesser version of **Y**, and the study of **X** thus can be only comparative, guided by differences from **Y**. Rejecting this differential approach is utmost for building a corpus, an integrative endeavour by definition (Goldin and Kryuchkova, 2011).

[2] Within the historical Slavic studies, there is no consensus on whether West Slavic, South Slavic and East Slavic are valid groups and that proto-West Slavic, proto-South Slavic, or proto-East Slavic actually existed (Nikolaev, 1988, p. 116; Krys'ko, 1998, p.85–89; Zaliznyak, 2004, p. 7–8; Matasović, 2008, p. 83; Saenko, 2020). This article uses East Slavic in a purely geographical sense to denote the eastern part of the Slavic continuum. This includes the territory from eastern Slovakia to the south-east to the coast of White Sea in the north-west.

Universal Dependencies (UD) (de Marneffe et al., 2021). It also provides a linguistic analysis of the Stanza model performance trained on the modern Ukrainian and starts developing guidelines for the automatic annotation of East Slavic small territorial lects.

2. Related Work

2.1. East Slavic small territorial lects corpora

While gathering the material of small territorial lects has been a pivotal part of dialectology studies from their wake (Wenker et al., 1889–1923) for the recent two hundred years (Kalnyn', 1973; Nazarova, 1977), the push for digitisation is relatively recent, especially for the digitisation of texts (Goldin, 1990). Most digitisation efforts still consider atlases (Trüb, 1989) that get transformed into digital maps (Marchenko et al., 2025). Nevertheless, the Russian National Corpus (Kachinskaya and Sichinava, 2015) and the Belarusian N-Corpus both (National Corpus of the Belarusian Language, 2018) have dialectological subcorpora. While the General Regionally Annotated Corpus of the Ukrainian Language includes some regional variation, there are almost no autochthonous small territorial lects within the material (Shvedova et al., 2017–2025). Aside from the large national-level corpora, there are other initiatives. The most notable are the TrimCo corpus (Wiemer and Seržant, 2020), containing small territorial lects of East Baltic, including multiple Belarusian and Russian small territorial lects, and the HSE collection of the East Slavic small territorial lect corpora (Daniel et al., 2013–2018; Garder et al., 2018; Ter-Avanesova et al., 2018, 2019; Ronko et al., 2019; Ryko and Spiricheva, 2020), each containing some hundred thousand tokens, enabling medium-scale variation research, including quantitative approaches. Aside from them, there are local initiatives dedicated to thorough investigation of a particular group of lects. The Saratov dialectological corpus comprises material from three small territorial lects, Northern Russian Megra, Middle Russian Belogornoje, and Middle Russian Zemlianyje Hutora, gathered over the last forty years (Goldin, 1990; Goldin and Kryuchkova, 2011). The Tomsk dialect corpus is a digitisation of materials from many different lects of the Tomsk Region and Western Siberia in general, adding up to 3.5 million tokens (Zemicheva et al., 2023).

Overall, there is a significant number of corpora for modern East Slavic territorial lects. Yet, historical variations are clearly underrepresented, and there are no corpora for the western part of the continuum representing the small territorial lects and not the more modern regiolects (for the whole territory of the standard Ukrainian distribution and further to the West into the territories of Poland and Slovakia).

2.2. East Slavic language variation and NLP tools

There is a rich tradition of variationist studies of Slavic languages in general, and East Slavic in particular. There are works dedicated to the variation in phonetics (Moroz, 2024), morphology (Ryko, 2024), lexicon (Zemicheva, 2020), and syntax (Moroz, 2016). Some take a wider typological approach (Wiemer et al., 2017), others restrain the comparison to the smaller areas (Ryko and Spiricheva, 2022).

However, there is one gap in this body of research. Despite the widening application of computational methods in variationist studies of East Slavic languages (Koile and Moroz, 2024), there is a relatively small number of studies that consider the role of variation in building language resources, specifically oriented towards East Slavic in general and Transcarpathian in particular (Scherrer and Rabus, 2017; Rabus and Scherrer, 2017), when contrasted to the other Slavic groups (Ondrejová and Šuppa, 2024; Lendvai et al., 2025). This article aims to outline the project of a corpus that facilitates closing the gap.

3. Transcarpathian lects

This section gives a short overview of all of the Transcarpathian lects used to build the corpus from both the diachronic and synchronic perspectives.

3.1. The general description history of the clade

The *Transcarpathian* lects are a group of East Slavic small territorial lects spoken in the geographical region of Transcarpathia. There are three main groups, for which the researchers collected the databases: Bojkian, Lemkian, and Huzulian (Kuraszkiewicz, 1963, pp. 67–72; Zilyns'kyj, 1933, pp. 8–10; Del Gaudio, 2017, pp. 64–85). The first two are Slavic lects that are the closest to being autochthonous (by the terminology of Barannikova (2005, p. 193)) on these territories. The Huzulian lects are late settlement (by the terminology of Barannikova (2005, p. 193)). The map 1 from (Zilyns'kyj, 1933) shows the geographic distribution of the lects in the 1920s – 1930s.

3.2. Morphology

The scholars (Kuraszkiewicz, 1963; Myholynec', 2004; Del Gaudio, 2017) report on a set of common

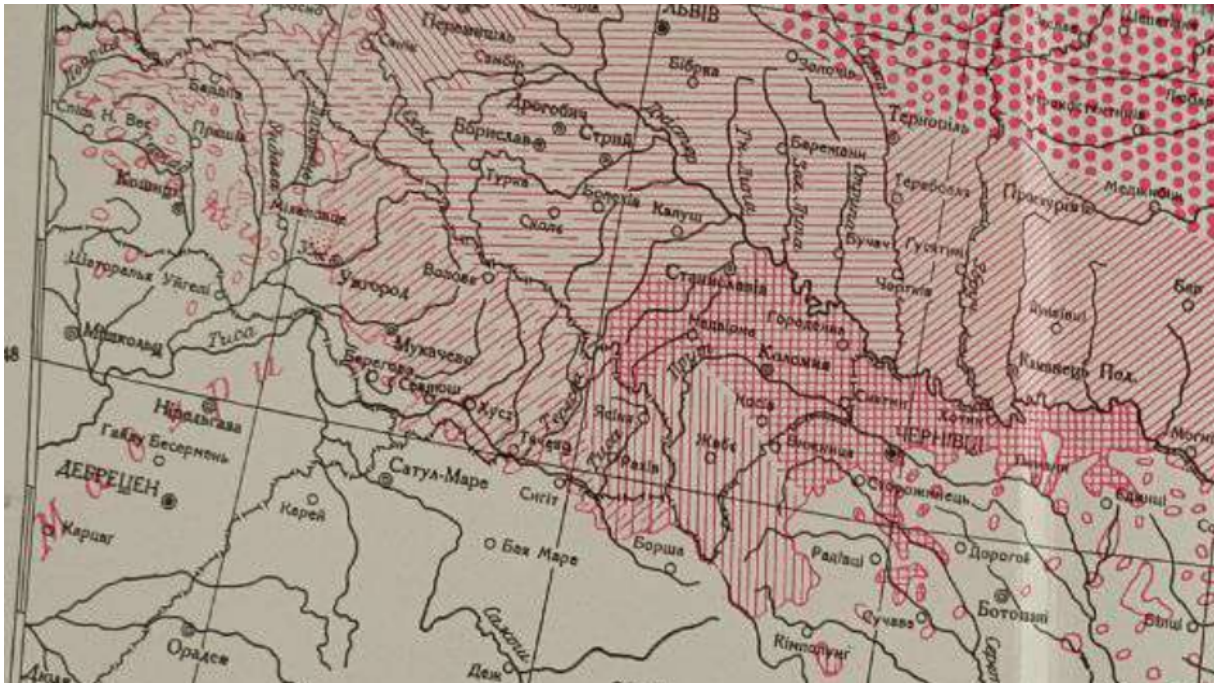

Figure 1: The map of Transcarpathian lects distribution in the beginning of the XX century (Zilyns'kyj, 1933). Rare horizontal strikes denote Lemkian (in the left corner), more dense horizontal strikes (closer to the center) denote Bojkian, dense vertical strikes (closer to the right) denote Huzulian.

Transcarpathian features that divide this area from the neighbouring East Slavic languages. The most prominent features include:

- Transcarpathian lects retain short forms of personal pronouns, for instance, ç'a 'self-ACC.SG' that in some cases can undergo root reduplication: seç'a 'self-ACC.SG'. The frozen form ç'a 'self-ACC.SG' denoting the reflexivity of the verb is not restricted in its distribution within the clause, in contrast (Kuraszkiewicz, 1963, p.71).

- In Transcarpathian lects INS.PL of names has -ma, a former INS.DU ending, cf. ps'oma 'dog-INS.PL'. The standard Ukrainian has form ps'ami 'id.'.

- Present tense third person singular and plural, along with imperative, have -t ending, in contrast to standard Ukrainian ∅ or -t^{j}: Transcarpathian xod'i-t 'walk-PRES.3SG' – standard Ukrainian xod'i-t^{j} 'id.' (Myholynec', 2004, p. 18).

To summarise, there are many common archaisms within the Transcarpathian area, which alternate the texts enough for the NLP tools to struggle, presenting a well-suited material for variation study research.

3.3. Lexicon

The lexicon of Transcarpathian lects is both highly innovative and highly archaic. There are many borrowings from the neighbouring languages across all the Transcarpathian lects (Nakonečna and Rudnyc'kyj, 1940, p. 71). At the same time, there are a lot of archaisms (Del Gaudio, 2017, p. 85). One of the most prominent features is the word for 'one'. In the Transcarpathian lects it is jed'en, a form closer to West and South Slavic, while in standard it is the usual East Slavic od'in (Kuraszkiewicz, 1963, p.72). This may cause some issues, similar to the out-of-domain problem (Afanasev and Lyashevskaya, 2023), but as the Transcarpathian lects contributed a lot to the standard Ukrainian vocabulary (Del Gaudio, 2017, p. 85), not crucial ones.

3.4. Syntax

There are several peculiar constructions in the Transcarpathian lects not present in standard Ukrainian, including:

- The Transcarpathian lects use prepositions na 'on' and k 'to' to express the direction instead of the standard Ukrainian do 'towards'.

- The Transcarpathian lects use conjunction fii in comparative constructions.

Some of these, like the accusative possessed construction, are less likely to cause issues with

syntactic tagging. The others, like the causative construction, seem to present a bigger problem.

4. Resources

The texts that comprise the corpus are mostly a product of the collective effort of a single group of scholars, who, during the 1920 – 1930s, documented various East Slavic lects spread mostly in the Transcarpathian geographical region from Central Slovakia to Western Ukraine and from Southern Poland to Northern Romania. More texts are available via the efforts of the 1950s researchers, but these are relatively more scarce. The following paragraphs describe each of the books that came out of the efforts in terms of the presented texts, their description, and metalinguistic information.

4.1. Nakonečna and Rudnyc'kyj (1940)

Nakonečna and Rudnyc'kyj (1940) contains texts, a dictionary and general information on three southwesternmost East Slavic small territorial lects of Transcarpathian territories. The book consists of three parts: the general introduction, the description of the material lect by lect, and the short differential[3] dictionary of the lexical items from all the lects. Of these three, all are relevant for the digitisation effort and corpus creation.

The introduction starts with the general information on the linguistic features of the analyzed lects, which are common for all of the analysed lects, when they are compared to other East Slavic lects and especially Ukrainian (Nakonečna and Rudnyc'kyj, 1940, pp. 7–8).

The introduction continues by providing metainformation on the speakers of all three lects, mainly stating their linguistic and social background (date of birth, education, exposure to standard Ukrainian). The introduction also outlines the research methodology. The crucial part is an extremely detailed table of phonetic transcription (Nakonečna and Rudnyc'kyj, 1940, pp. 17–18) providing key information for digitising the texts in a more common format.

The next section of Nakonečna and Rudnyc'kyj (1940) consists of three parts, each dedicated to one particular idiolect from Lemkian (Nakonečna and Rudnyc'kyj, 1940, pp.23–37), Bojkian (Nakonečna and Rudnyc'kyj, 1940, pp.49–62) and Huzulian (Nakonečna and Rudnyc'kyj, 1940, pp. 65–82) parts of the continuum. Each of these parts outlines the phonetic, morphological, and lexical peculiarities of the lect, as well as provides the texts in these lect. The main issue is that the comparative research does not represent small territorial groups of Lemkian, Bojkian and Huzulian, but a single speaker out of each of these groups, which can complicate a variation study (including a historical phonology one). Metainformation on the texts (there is no split between the texts, only between – it seems – the recordings) and the speakers is not present in these parts, which is especially problematic, given that it is already significantly restricted in the introductory part of the book.

to sut fʂ'ɯtkɣ lemk'ʉ̈wskɣ s'elɑ
de po lemk'ʉ̈wskɣ fiw'ɑrjɑt
То сут вши́тки лемки́вски се́ла,
де по лемки́вски гва́рят
Das sind alles lemkische Dörfer,
wo lemkisch gesprochen wird.

Table 1: An example of a sentence from Nakonečna and Rudnyc'kyj (1940, p. 31). At the top, there is phonetic representation and in the middle is a standard-based transcription. The German translation is at the bottom. The translation of the sentence is *These all are Lemkian villages, where one speaks Lemkian.*

The representation of texts is rather detailed, with three forms: the phonetic transcription[4], the standardised[5] transcription, and the German translation[6]. Table 1 shows an example.

The fine-grained transcription is, on the one hand, a serious advantage for the reconstruction of the phonetic system. However, there is no understanding of how reliable it actually is, especially without surviving recordings. One more issue is that there is almost no possibility to use either original transcription or its IPA rendering for automatic tagging. For the latter purposes, the study is going to use a standard-based transcription, also provided in Nakonečna and Rudnyc'kyj (1940).

After a short conclusion (Nakonečna and Rudnyc'kyj, 1940, pp. 83–85), the book provides a dictionary of the words, specific for the small territorial lects under study. While not providing the full information required for a lexical study, this is especially helpful for lemmatisation. Overall, Nakonečna and Rudnyc'kyj (1940) provides a material of high quality for the time given, though lacking clarity and transparency.

4.2. Zheguc (2001)

Zheguc (2001) is a collection of texts, collected in

[3]The one that contains only the items that differ from the standard language, in this case, Ukrainian, as opposed to the full, which contains all the lexical items from the studied lect.

[4]I provide the IPA version, as it is more well-known and easy to use than the original transcription

[5]Using the same set of graphemes as standard Ukrainian of the time

[6]German seems to be L2 for authors, as preivous research has claimed (Afanasev, 2025)

different time periods (1920s – 1930s and 1990s) in the Huzulian part of the Transcarpathian region. For chronological uniformity, the corpus takes only the texts that come from the 1920s – 1930s. While this restricts the overall quantity of the material and does not allow for a proper language change study, it greatly assists in balancing the corpus. The modern material is the result of continuous field trips; thus, it is going to reduce the historical material to residuals if put in the same corpus. This is not to mention that due to the Soviet era deportations the linguistic landscape has probably drastically changed between the 1920s – 1930s and the 1990s, therefore it is hardly suitable for a proper comparison.

The transcription system of texts is clear. It is standardised phonetic/phonematic transcription, based on standard Ukrainian, but depicting the key phonetic and lexical features of the Transcarpathian lects, relevant for the purposes of the study. It does not represent all the phonetic nuances and interidiolect variation, but preserves the crucial characteristics and acts as a compromise solution, necessary for the texts for which there is no recording available. The established rules of the transcription system allow it to be unified with the standardized transcription system of Nakonečna and Rudnyc'kyj (1940). Pivotally, Zheguc (2001, p. 8) provides the description of the phonetic features of the lects, which assists in unifying the transcription.

From the sociolinguistic point of view, Zheguc (2001) provides the best possible ground for tagging metadata. It contains standard sociolinguistic information about the speakers (age, occupation, level of education, name) and the dialectologist who had initially recorded the texts, Ivan Paňkevyč. This not only enables providing the research with more data for the variationist study, but possibly facilitates insight into how a particular scholar chooses to represent a non-standard variety (Saenko, 2018).

Thus, while the representation of the texts is not the most detailed among the other resources, the quality of the material is probably the best. The transcription transparency and the relative abundance of sociolinguistic information make the data useful for sociophonetic, sociolexical, and sociogrammatical studies alike and necessitate including this resource in the corpus.

4.3. Kuraszkiewicz (1963)

Kuraszkiewicz (1963) contains a short description of each of the East Slavic small territorial lect groups, as well as texts from the lects of these groups. Among these are several texts from the Transcarpathian lects, recorded and transcribed mostly at the beginning of the XX century (Kuraszkiewicz, 1963, pp. 124–128).

The possible digitisation of the texts from Kuraszkiewicz (1963) faces some significant issues. The transcription system here, on the contrary to Nakonečna and Rudnyc'kyj (1940), is not explained, and it is rather hard to make a proper correspondence between it and IPA, or even standard Ukrainian. There are also no signs of either a standard-based additional transcription or a translation into any other language. Given the lack of audio recordings, this is quite problematic. The crucial issue, however, is the lack of metadata, aside from the group (Lemkian/Bojkian/Huzulian) and the source of transcription, including the dialectologist's name. Kuraszkiewicz (1963) also generally provides the particular area where the recording took place. This facilitates some understanding on how different researchers represented different varieties; however, studying the varieties themselves is going to be significantly complicated.

While not exactly rich with metalinguistic and linguistic information, Kuraszkiewicz (1963) possesses one crucial advantage over all the other sources, namely, the geographical distribution of the lects. It covers at least two lects from each group, bringing much-needed diversity to the dataset.

5. Corpus creation

5.1. Corpus design

The main purpose of the corpus is to provide insights into the processes of variation and change that were going on in the Transcarpathian lects, as illustrated by the morphosyntactic properties of the texts. The best format to represent these properties is `.conllu`, so this is the main format of the corpus.

The transcriptions within the different sources are drastically different, so the research opts to unify them. The main transcription system is IPA. The main representation system for morphological tagging, however, is a standard Ukrainian phonetic/phonematic rendering, applied by both Nakonečna and Rudnyc'kyj (1940) and modern small territorial lects corpora of East Slavic languages, for instance, Goldin and Kryuchkova (2011) or Ryko and Spiricheva (2020). This system renders non-standard texts in something more resembling a neighbouring standard, but with faithful depiction of key phonetic, morphological, lexical and syntactical features (von Waldenfels et al., 2014). This helps to present the corpus to the audience unfamiliar with the IPA conventions. Crucially, this system also facilitates more effective automatic processing. When possible, the corpus also provides translations into other languages.

The corpus has two layers of tagging: mor-

phosyntactic for each token (part-of-speech, morphological tags, lemma, syntactic features), and an additional linguistic one. The latter denotes the elements of the Automatic Similarity Judgement Program (ASJP) basic vocabulary list (Holman et al., 2008, pp. 336–337) and named entities. The main motivation to do so is to provide additional information for the study of language change (ASJP basic vocabulary) and the reconstruction of the social landscape that surrounded the speakers (named entities).

As the target audience of the corpus is linguists who lean into variation studies, the additional file includes sociolinguistic information for both the speakers and the scholars who performed the recording and/or transcription. This is a `.csv`-file, with columns representing parameters (among others, name of the speaker, year of birth of the speaker, notes of the dialectologist on the speaker's background), and rows representing texts. This is due to the possibility of documents containing the texts from more than one speaker, and the speakers not being mentioned in some of the resources Kuraszkiewicz (1963).

5.2. Manual Digitisation

Due to the specific system of transcription for the given lects, using OCR techniques of e-Scriptorium (Kiessling et al., 2019) or Transkribus (Kahle et al., 2017) proved to be impossible. Therefore, the performance of digitisation process is manual.

The workflow was the following:

- Manually type the material into a machine-readable form.
- In cases of Kuraszkiewicz (1963) and Zheguc (2001) it is necessary to add a standardised layer.
- Split the whole material into documents (especially relevant for Kuraszkiewicz (1963) and Zheguc (2001); Nakonečna and Rudnyc'kyj (1940) already contains splits by document.
- Split the document into texts (especially relevant for Nakonečna and Rudnyc'kyj (1940), which does not provide splits by texts within the documents).
- Split the texts into sentences (mostly done by the scholars).
- Perform manual word tokenisation.
- Put data in `.conllu` format.
- When required, provide additional information about the ASJP basic vocabulary list and named entities to the *misc* section of the token. In the *misc* section, there are also WF and TF fields, the latter providing the IPA-based transcription (if the original data contains phonetic transcription), and the former – normalised token (standard-based representation with removed diacritics).

This prepared the corpus for the next phases of tagging. These stages included automatic annotation, manual correction, and providing sociolinguistic information about the speakers.

5.3. Automatic Annotation

The code used for automatic annotation is in a demo repository[7]. The annotation process uses its separate file (`code/silver_tagging_depparse.py`).

The main tool of automatic annotation is Stanza (Qi et al., 2020), a well-known set of pre-trained models designed to perform basic NLP tasks, providing output in the Universal Dependencies (UD) format (de Marneffe et al., 2021). Among others, the ones especially relevant for this study are: part-of-speech tagging, morphological tagging, lemmatisation, and dependency parsing. The automatic annotation pipeline includes running the Stanza model, trained on the standard Ukrainian corpus from UD (Nivre et al., 2020). This is due to the similarity of the selected representation system and the standard Ukrainian graphic system: it allows for the best possible results.

When having performed tagging, the script deletes XPOS: these are language-specific tags for standard Ukrainian. While, due to the high degree of similarity, some of them still can be applicable to the Transcarpathian lects, the differences within the grammatical systems of even rather closely related varieties are generally still significant enough to cause issues in the XPOS schema (Shishkina and Lyashevskaya, 2021). Thus, the decision was to temporarily remove this feature. After XPOS deletion, the data are ready for manual correction.

5.3.1. Note on the use of GenAI

While Generative AI (GenAI) is extremely useful in low-resourced settings, when compared to the more traditional models (Baturova et al., 2025), its zero-shot application, even when there are

[7] Demo material, anonymised for review purposes, available at https://osf.io/528zy/overview?view_only=3c4c72fd0e3d44ebb2b732d191409193 (last accessed: October 25, 2025). The code and data excerpts, provided in the repository, are purely for reproducibility and transparency purposes. The paper is reviewable on its own in terms of the resource I am presenting.

high-resource closely related languages in the pretraining dataset, may still be problematic (Umbet et al., 2025). At this stage, there are not enough resources in the corpus to provide examples for GenAI prompts. For the further stages of the research, when the first texts for each Bojkian, Huzulian and Lemkian are fully digitised and cross-checked, the experiments with GenAI, as compared to the traditional tools, are necessary.

5.4. Manual Correction and Preliminary Analysis

The aim of manual correction stage is to edit the incorrectly assigned lemmata, part-of-speech, morphology and syntactic relationship tags. The crucial stage of manual correction is error analysis. As the automatic tagging stage uses Stanza cross-linguistically, there are many mistakes, especially caused by the errata in the previous phases of tagging. Below are examples for each category.

5.4.1. Part-of-speech/morphological tagging

The errata in part-of-speech tagging and, subsequently, morphological tagging emerge heavily from the differences in distributions of some items between standard Ukrainian and the Transcarpathian lects. In this fashion, Stanza tags То 'this' in LA1407.1.4 as a particle (PART). While in standard Ukrainian то is indeed a particle, то in the analysed text is rather a demonstrative pronoun (DET), more akin to standard Ukrainian це.

This error also causes a subsequent chain of errors, the most prominent being the complete absence of the morphological tags required for the demonstrative pronoun. This underscores one of the crucial issues of Stanza-based pipelines: the simultaneous tagging of part-of-speech and morphology.

In other cases, the part-of-speech tag is correct, but some of the morphological tags are not. For instance, the model tags Фóльварк [8]'Folvark village' as NameType=Sur (family name). This is especially frequent with some village names that get confused with family names or given names, which underscores the issues of out-of-domain tagging (Lyashevskaya and Afanasev, 2021).

5.4.2. Lemmatisation

The lemmatisation errata often stem from the incorrect part-of-speech tags. This is the case, for instance, of the model transforming вши́тки 'every-NOM.PL' to вшиток as a consequence of NOUN part-of-speech tag. In fact, вши́тки is an adjective, and therefore should get lemma вшиткий. The other cases include, for instance, lemmatising ме́нджі 'between' to мендж, triggered by NOUN tag.

The other type of error is the combination of lexical differences and the inability of the models to account for phonetic or morphological properties. Thus, Русна́ці 'Rusnak-NOM.PL' becomes Руснац instead of expected Руснак. The phonetic changes that led to this ц/к alternation (Zhovtobriukh et al., 1979, pp. 119–120) are common for the Transcarpathian lects and standard Ukrainian, cf. році 'year-NOM.PL' – рік. It is clear that the model does not grasp this kind of alternation.

In some cases, the combination of the lack of training material and the grammatical differences may also cause lemmatisation errata. For instance, ні́ма 'they-INS.PL' is a form analogous to standard Ukrainian ними 'id.' (see Section 3.2. Its lemma is він. The model, however, picks Кіма, which is a clear generation error, caused by the absence of both ними and the words ending with ма and being in INS.PL in the training dataset. This shows that Ukrainian is still a low-resourced language in terms of the UD corpora.

5.4.3. Dependency parsing

The dependency parsing errata are multiple, but mostly have a single cause: incorrect part-of-speech tag. Thus, the aforementioned То gets DISCOURSE tag, while in fact it is NSUBJ. In case it were tagged as DET (as it should have been), there is a high chance that the assigned syntactic tag would have been correct.

5.4.4. Discussion and final representation

Overall, the errata made by the model stem not from significant differences in the morphological or syntactical structure of standard Ukrainian and the Transcarpathian lects, but rather from their lexical differences, low training material (Ukrainian corpus in UD 2.12, on which Stanza for Ukrainian was trained, has only 114 000 tokens), and running Stanza as a pipeline without manual checks in-between. Still, the manual check significantly reduced the errata of the model. Table 2 shows an example of a data piece after the manual check.

5.5. Metadata

The tagging of metadata, while not pivotal in terms of, for instance, automatic processing, is a critical part of both language variation studies in general and corpus-based dialectology in particular (Tagliamonte, 2025, pp. 109–111). The texts within the corpus are thus going to receive the metadata tag

[8] I provide the form with stress for illustrative purposes. However, the forms that the model tagged underwent normalisation to exclude this factor of errata.

sent_id = LA1407.1.4
IPA_transcription = to sut fʂ'ɯtkɣ lemk'ӧwskɣ s'elɑ de po lemk'ӧwskɣ fiw'ɑrjɑt
standard_text = То сут вши́тки лемки́вски се́ла, де по лемки́вски гва́рят
german_text = Das sind alles lemkische Dörfer, wo lemkisch gesprochen wird.

1 То то PRON _ Animacy=Inan|Case=Nom|Gender=Neut|Number=Sing|PronType=Dem
2 nsubj 2:nsubj wf="То"|tf="то"

(...)

Table 2: The manually checked annotation of LA1407.1.4. As the table is an illustration, it shows (for brevity considerations) only the first token.

of belonging to one of these three groups, according to their descriptions within the sources, as well as the discovered phonetic features of the lects.

Where it is possible, the texts receive the tags of information about the speakers themselves. The first name and the last name undergo encryption via being transliterated into the Roman script, and afterwards abbreviated to the format "first letter of the first name + second name + the last three digits of the year of birth, otherwise 000". Thus, Василь Кабалюк[9], born in 1870, has id *vkabaluk870*, while Гафія Прумптула[10], information on whose year of birth is not in the data, has id *hprumptula000*. This is due to the variationist studies conventions and the consensus about the anonymity of the speakers[11] (Tagliamonte, 2025, pp. 48–50). Other features of the speakers include year of birth, place of birth (if known), education and occupation, where available.

More information is available from the recordings themselves. The corpus takes date and place of original publishing, genre and form (there are some songs in the dataset, so it is necessary to distinguish between poetry and prose). This enables the basic discourse analysis.

5.6. Representation and Access

The corpus is going to be available on Zenodo as a separate language resource. The demonstration part of the material is available at the anonymous Open Science Framework repository[12].

[9]Fictionalised name, not attested in the data

[10]Fictionalised name, not attested in the data

[11]While the original resources are available for research purposes and have a permissive license, they have not been widely accessed before the publication. Thus, it is better to adapt additional measures for privacy safety.

[12]Demo material, anonymised for review purposes, available at https://osf.io/528zy/overview?view_only=3c4c72fd0e3d44ebb2b732d191409193 (last accessed: October 25, 2025). The data files are in the `data` folder of the repository.

6. Current state: Conclusion

The article presented the design for the corpus of the historical Transcarpathian lects. The study outlined the resources used for the creation of this corpus. It demonstrated the research pipeline, utilised in building this corpus, from gathering resources and manual digitisation to the manual correction of the errata made by the Stanza model. The research showed the advantages of using stage-by-stage (part-of-speech tagging followed by lemmatisation, followed by dependency parsing) automatic processing of the texts, interrupted by manual check, when contrasted to the whole pipeline run. The main contribution is the open-access texts of the Lemkian part of the corpus.

The study provided the analysis of the errata made by the Stanza model. It provided evidence that the quality of performance by part-of-speech and morphological models is pivotal to all the subsequent stages. The analysis also demonstrated that lexical differences are one of the most crucial factors of model performance, and that the Stanza is not able (at least, on 110 000 tokens of training material) to catch some phonetic change patterns present in the dataset.

The next stage is continuing to fill the corpus with the material. Now, when part of the material is manually checked, the next step is to use GenAI as either an aid to the Stanza kit or the main tagger. The corpus also needs English translations of the sentences for better accessibility and, possibly, TEI representation for the a more qualitative study.

7. Ethical Considerations

The data has been published in the printed form and available for research purposes for fifty to ninety years by the time this article was written. Still, I anonymise the metadata, where possible, masking the names of the speakers, to compensate for possible ethics violations that could have happened at the time of the material collection.

The data themselves can contain slight mentions of xenophobic behaviour and religious (mostly,

Christian) imagery. Discretion is advised.

8. Limitations

The corpus is currently in its development phase, which means that the digitisation and tagging processes are not fully finished. The tagging is mostly silver, the estimated F1 for this degree of relationship between lects is 0.8 (Lyashevskaya and Afanasev, 2021), which can be acceptable for proof-of-concept, but not for a more detailed study.

9. 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.

10. Bibliographical References

References

Ilia Afanasev. 2025. Computer-assisted study of historical Lemkian (Transcarpathian) lects: Basic vocabulary approach. *Scripta & e-Scripta*, 25:11–24.

Ilia Afanasev and Olga Lyashevskaya. 2023. From web to dialects: how to enhance non-standard Russian lects lemmatisation? In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 167–175, Gothenburg, Sweden. Association for Computational Linguistics.

L. I. Barannikova. 2005. Govory territorij pozdnego zaselenija i problema ih klassifikacii [dialects of late-settled territories and the problem of their classification]. In V. Goldin and O. Kryuchkova, editors, *Barannikova L. I. Obshhee jazykoznanie: izbrannye raboty [L. I. Barannikova. General linguistics: Selected works]*, pages 192–203. KomKniga.

Dari Baturova, Sarana Abidueva, Dmitrii Lichko, and Ivan Bondarenko. 2025. Low-resource buryat-Russian neural machine translation. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 85–93, Vienna, Austria. Association for Computational Linguistics.

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

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

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.

E. Holman, S. Wichmann, C. Brown, V. Velupillai, A. Müller, and D. Bakker. 2008. Explorations in automated language classification. *Folia Linguistica*, 42:331–354.

I. B. Kachinskaya and D. V. Sichinava. 2015. Dialektnyj korpus segodnja [the dialect corpus today]. *Trudy Instituta russkogo jazyka im. V.V. Vinogradova*, 6:142–163.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus - a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24.

Ljudmila Ė. Kalnyn'. 1973. *Opyt modelirovanija sistemy ukrainskogo dialektnogo jazyka: fonologicheskaja sistema [An attempt at modeling the system of the Ukrainian dialectal language: The phonological system]*. Nauka, Moscow.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. escriptorium: An open source platform for historical document

analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19.

E. Koile and G. Moroz. 2024. Detecting linguistic variation with geographic sampling. *Journal of Linguistic Geography*, 12(1):24–31.

V. B. Krys'ko. 1998. Drevnij novgorodsko-pskovskij dialekt na obshcheslavjanskom fone [the Old Novgorod-Pskov dialect against a common Slavic background]. *Voprosy Jazykoznanija*, 3:74–93.

Władysław Kuraszkiewicz. 1963. *Zarys dialektologii wschodniosłowiańskiej z wyborem tekstów gwarowych*, wyd. 2., zmien. i rozsz. edition. Państwowe Wydawn. Naukowe, Warszawa.

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.

Olga Lyashevskaya and Ilia Afanasev. 2021. An hmm-based pos tagger for old church slavonic. *Journal of Linguistics/Jazykovedný casopis*, 72(2):556–567.

Ranko Matasović. 2008. *Poredbenopovijesna gramatika hrvatskoga jezika [The historical-comparative grammar of the Croatian language]*. Matica hrvatska, Zagreb. Biblioteka Theoria.

Aleksandra Miletić and Janine Siewert. 2023. Lemmatization experiments on two low-resourced languages: Low Saxon and Occitan. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 163–173, Dubrovnik, Croatia. Association for Computational Linguistics.

G. A. Moroz. 2016. Adverbial'nye konstrukcii vremennoj distribucii v balto-slavjanskih jazykah: areal'noe i korpusnoe issledovanie [adverbial constructions of temporal distribution in the balto-slavic languages: An areal and corpus study]. In N. Kazanskij and D. V. Gerasimov, editors, *Acta Linguistica Petropolitana. Trudy Instituta lingvisticheskih issledovanij RAN (Tom XII, chast' 1) [Acta Linguistica Petropolitana. Proceedings of the Institute for Linguistic Studies of the Russian Academy of Sciences (Volume XII, Part 1)]*, volume XII/1, pages 151–167. Nauka, Saint Petersburg.

G. A. Moroz. 2024. Skorost' russkoj rechi na osnove bilingval'nyh i dialektnyh ustnyh korpusov [the speed of russian speech based on bilingual and dialectal oral corpora]. In N. A. Korotaev and N. R. Sumbatova, editors, *Sostav nauki: Sbornik statej k jubileju Very Isaakovny Podleskoj [The Composition of Science: A Collection of Articles for the Anniversary of Vera Isaakovna Podleskaya]*, pages 366–378. Buki Vedi, Moscow.

Ol'ha F Myholynec'. 2004. *Ukraïns'ki zakarpats'ki hovirky : teksty*. Lira, Užhorod.

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.

National Corpus of the Belarusian Language. 2018. Nacyjanalny korpus bielaruskaj movy ŭ kantekscie corpusnaj linhvistyki slavianskich krain [the national corpus of the belarusian language in the context of corpus linguistics of slavic countries]. In *XVI International Congress of Slavists*, Belgrade.

Tetjana V. Nazarova. 1977. *Hovory ukraïns'koï movy: zbirnyk tekstiv [Dialects of the Ukrainian language: A collection of texts]*. Naukova Dumka, Kyiv.

S. L. Nikolaev. 1988. Sledy osobennostei vostochnoslavianskikh plemennykh dialektov v sovremennykh velikorusskikh govorakh. 1. Krivichi [Traces of Features of East Slavic Tribal dialects in Modern Great Russian Dialects. 1. Krivichi]. In *Baltoslavianskie issledovaniia 1986 [Balto-Slavic Investigations 1986]*, pages 115–154. Nauka, Moscow.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4034–4043, Marseille, France.

Viktória Ondrejová and Marek Šuppa. 2024. Can LLMs handle low-resource dialects? a case study on translation and common sense reasoning in šariš. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 130–139, Mexico City, Mexico. Association for Computational Linguistics.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Morgan & Claypool Publishers.

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.

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.

A. I. Ryko and M. V. Spiricheva. 2022. The degree of preservation of dialectal features in different generations (khislavichi district of the smolensk region). *RSUH/RGGU Bulletin: "Literary Theory. Linguistics. Cultural Studies" Series*, 5:121–141. (In Russian).

Anastasiia Ryko. 2024. A study of russian-belarusian border dialects: The use of the genitive case ending -u in the khislavichi dialect. *Zeitschrift für Slawistik*, 69(4):747–765.

M. N. Saenko. 2018. Netochnosti v opisanii semantiki, vyzvannye vosprijatiem dialektnoj leksiki skvoz' prizmu literaturnogo jazyka: neskol'ko primerov [inaccuracies in the description of semantics caused by the perception of dialectal vocabulary through the prism of the literary language: Several examples]. In *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.

M. N. Saenko. 2020. Taxonomy of Slavic languages, history of the. In M. L. Greenberg, editor, *Encyclopedia of Slavic Languages and Linguistics Online*. Brill.

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.

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.

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

Rudolf Trüb. 1989. Der Sprachatlas der deutschen Schweiz (SDS): ein Grossatlas für einen Kleinraum [the language atlas of german-speaking switzerland (sds): A large atlas for a small area]. In Werner H. Veith and Wolfgang Putschke, editors, *Sprachatlanten des Deutschen: laufende Projekte [Language Atlases of German: Ongoing Projects]*, pages 133–177. Niemeyer, Tübingen.

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.

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, Moscow. Izd-vo RGGU.

Björn Wiemer, Ilja Seržant, and Aksana Erker. 2017. Convergence in the Baltic-Slavic contact zone: Triangulation approach. In Juliane Besters-Dilger, Cynthia Dermarkar, Stefan Pfänder, and Achim Rabus, editors, *Congruence in Contact-Induced Language Change*, pages 15–42. De Gruyter.

Björn Wiemer and Ilja A. Seržant. 2020. East Slavic dialectology: Achievements and perspectives of areal linguistics. In I. A. Seržant and B. Wiemer, editors, *Contemporary Approaches to Dialectology: The Area of North, Northwest Russian and Belarusian Vernaculars*, volume 13 of *Slavica Bergensia*, pages 11–80. John Grieg AS, Bergen.

A. Zaliznyak. 2004. *Drevnenovgorodskij dialekt [The Old Novgorodian Dialect]*. Jazyki slavjanskoj kul'tury, Moscow.

S. S. Zemicheva. 2020. Ot abarma do jashhichishka: razrabotka leksikograficheskogo komponenta tomskogo dialektnogo korpusa [from *abarm* to *jashhichishek*: Development of the lexicographic component of the tomsk dialect corpus]. *Voprosy Leksikografii*, 18:98–117.

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

Mikhailo Andriiovych Zhovtobriukh, V. M. Rusanivs'kyi, and V. H. Skliarenko. 1979. *Istoriia ukraïns'koï movy. Fonetyka [History of the Ukrainian Language. Phonetics]*. Naukova dumka, Kyiv.

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.

Language Resource References

Daniel, Michael and Dobrushina, Nina and von Waldenfels, Ruprecht. 2013–2018. *The Language of the Ustja River Basin: A Corpus of North Russian Dialectal Speech*. Linguistic Convergence Laboratory, NRU HSE.

Garder, M. O. and Petrova, N. S. and Moroz, A. B. and Panova, A. B. and Dobrushina, N. R. 2018. *Corpus of Spiridonova Buda Dialect*. Linguistic Convergence Laboratory, HSE University. Accessed on 24.10.2025.

Marchenko, I. A. and Dolgov, O. N. and Azanova, A. S. and Zambrzhitskaya, M. S. and Zalivina, E. A. and Zemlyanskaya, S. A. and Mochul'skij, D. I. and Tsejtina, E. I. and Chistyakova, D. G. and Ron'ko, R. V. 2025. *Database of the Dialectological Atlas of the Russian Language*. Institute of the Russian Language RAS. Accessed October 24, 2025.

Ronko, Roman and Volf, Elena and Grebyonkina, Maria and Ershova, Maria and Okhapkina, Anna and Khadasevich, Anna and Morozova, Valeria. 2019. *Corpus of Opochetsky Dialects*. Linguistic Convergence Laboratory, HSE University; V.V. Vinogradov Russian Language Institute Russian Academy of Science. Accessed on 24.10.2025.

Ryko, A. I. and Spiricheva, M. V. 2020. *Corpus of the Russian Dialect Spoken in Khislavichi District*. Linguistic Convergence Laboratory, HSE University. Available online at `https://lingconlab.ru/khislavichi/`, accessed on 23.10.2025.

Shvedova, M. and von Waldenfels, R. and Yaryhin, S. and Rysin, A. and Starko, V. and Nikolaenko, T. and Lukashevskyi, A. and others. 2017–2025. *General'nyj rehional'no anotovanyj korpus ukraïns'koï movy (HRAK) [General Regionally Annotated Corpus of the Ukrainian Language (GRAC)]*. University of Jena.

Ter-Avanesova, A. V. and Balabin, F. A. and Dyachenko, S. V. and Malysheva, A. V. and Panova, A. B. and Morozova, V. A. 2019. *Corpus of the Malinino Dialect*. Linguistic Convergence Laboratory, NRU HSE; V.V. Vinogradov Russian Language Institute of the Russian Academy of Science. Accessed on 24.10.2025.

Ter-Avanesova, A. V. and Dyachenko, S. V. and Kolesnikova, E. V. and Malysheva, A. V. and Ignatenko, D. I. and Panova, A. B. and Dobrushina, N. R. 2018. *Corpus of Rogovatka Dialect*. Linguistic Convergence Laboratory, NRU HSE. Accessed on 24.10.2025.

Wenker, Georg and Maurmann, Emil and Wrede, Ferdinand. 1889–1923. *Sprachatlas des Deutschen Reichs [Language Atlas of the German Empire]*. Research Center Deutscher Sprachatlas. Original manuscript (1889–1923). Published as the Digital Wenker Atlas (DiWA).

Zemicheva, S. S. and Dubtsova, L. A. and Gromov, M. L. and Galanina, V. V. and Ugryumova, M. M. and Vasilchenko, A. A. and Parshina, A. V. and Popova, D. P. and Duminskaya, A. V. and Zyuzkova, N. A. and Bukhanova, E. D. 2023. *Tomskij dialektnyj korpus 2.0 [Tomsk Dialect Corpus 2.0]*. Laboratorija obshhej i sibirskoj leksikografii NI TGU [Laboratory of General and Siberian Lexicography, National Research Tomsk State University]. Access mode: free.