Dialects, Topic Models, and Border Effects: The Rusyn Case

Achim Rabus [1] **and Yves Scherrer** [2,3]
[1] University of Freiburg, Germany
[2] Department of Digital Humanities, University of Helsinki, Finland
[3] Department of Informatics, University of Oslo, Norway
achim.rabus@slavistik.uni-freiburg.de yves.scherrer@ifi.uio.no

Abstract

In this contribution, we present, discuss, and apply a data-driven approach for analyzing varieties of the Slavic minority language Carpathian Rusyn spoken in different countries in the Carpathian region. Using topic modeling, a method originally developed for text mining, we show that the Rusyn varieties are subject to border effects, i.e., vertical convergence and horizontal divergence, due to language contacts with their respective umbrella languages Polish, Slovak and Standard Ukrainian. Additionally, we show that the method is suitable for uncovering fieldworker isoglosses, i.e., different transcription principles in an otherwise homogeneous dataset.

1 Introduction

This contribution is devoted to applying and evaluating data-driven approaches for the analysis of (Carpathian) Rusyn. (Carpathian) Rusyn is a Slavic minority language spoken in the Carpathians, most notably in Poland (where it is usually called Lemko), Slovakia, Hungary, and Ukraine. From the viewpoint of both historical phonology and culture, it belongs to the East Slavic branch with the closest related standard language being Standard Ukrainian, while language contacts over the years have made them very close to West Slavic languages such as Slovak. The status of Rusyn is somewhat contested. Although traditional Ukrainian dialectology regards Rusyn varieties as dialects of Ukrainian (Skrypnyk, 2013), there is a strong movement that maintains that Rusyn is a language of its own, independent from Ukrainian (Plishkova, 2009).

Nowadays, the traditional Rusyn dialect continuum (Gerovskij, 1995) is divided by multiple state borders, resulting in distinct sociolinguistic situations on each side. Because of that, it is justified to assume that so-called **border effects** (Woolhiser, 2005) occur, i.e., horizontal divergence within an old dialect continuum due to intense linguistic contacts with the respective umbrella languages Polish, Slovak, and Standard Ukrainian. This is in line with qualitative (Vašíček, 2020) and quantitative (Rabus, 2019) studies focusing on selected features.

Our paper is structured as follows: First, we discuss related work dealing with data-driven, machine-learning-oriented approaches to dialectometry. Subsequently, we present the data used for our analysis and elaborate on our methodological approach. We then present and discuss our results, and end our paper with a conclusion and an outlook on future research perspectives.

2 Related Work

Using corpus-driven methods to infer dialect areas has become more popular within the last years. For example, Wolk and Szmrecsanyi (2018) provide a classification of British English dialects on the basis of morphosyntactic features extracted from a dialect corpus, and Lameli et al. (2020) use Levenshtein distance of parallel dialect transcriptions to infer dialectal partitions of German-speaking Switzerland. Hovy and Purschke (2018) jointly learn vector-space representations ("embeddings") for words and cities in a georeferenced corpus of social media data.

Kuparinen and Scherrer (2024) propose to apply topic modeling, a method generally used for text mining purposes, to dialect corpora. They show that topic models reliably infer major dialect areas and the corresponding lexical, morphological and phonological specificities. Their experiments focus on three non-Slavic linguistic varieties, namely Norwegian, Finnish and Swiss German.

For Slavic, different methods for variant classification have been proposed. In von Waldenfels (2014), Neighbor Net graphs to visualize the respective distance of the Slavic languages regarding specific features are used. The R package Stylo

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 38–43

Area	Documents	Utterances	Tokens
LEM	46	12 510	149 713
(legacy)	29	9 291	115 155
(non-legacy)	17	3 219	34 558
SLO	20	4 093	34 407
TRA	23	2 629	24 284
Total	89	19 232	208 404

Table 1: Corpus statistics.

(Eder et al., 2016) can be used for, among others, cluster analysis. Moreover, the NSC algorithm implemented in Stylo allows for zooming in and identifying individual features for subsequent quantitative analysis (Lahjouji-Seppälä et al., 2022).

3 Data

For our analysis, we used the plain textual data available in the *Corpus of Spoken Rusyn* (Rabus and Šymon, 2015)[1]. The corpus contains recordings and corresponding transcriptions of interviews and interactions with numerous speakers of different varieties of Rusyn in Slovakia, Poland, Zakarpattia Ukraine, and Hungary. Most of the recordings were made in the years 2015 and 2016 specifically for the corpus. Additionally, some data gathered for other projects were integrated, especially for the Lemko variety of Rusyn. Unlike the rest of the data, these data were initially transcribed in the Latin script, but were converted to the Cyrillic script to better align with the rest of the dataset. For this study, we restricted ourselves to the Lemko (LEM), Slovak (SLO), and Transcarpathian (TRA) data. As our research is primarily concerned with computational dialectology and Rusyn writing conventions or written standards are a separate issue, we refrained from using other available data sources such as the Rusyn Wikipedia.

4 Method

We apply the topic modeling method introduced by Kuparinen and Scherrer (2024) to the Rusyn data.[2]

[1] https://russinisch.uni-freiburg.de/corpus

[2] Our code and experimental results are available at https://github.com/achimrabus/Rusyn_Topic_Modelling. It is based on the original code of Kuparinen and Scherrer (2024), which uses the *scikit-learn* library (Pedregosa et al., 2011) and is available at https://github.com/Helsinki-NLP/dialect-topic-model.

4.1 Topic models

Topic models are statistical models that aim to discover underlying similarities in a collection of documents based on co-occurring items. Formally, topic models take a term-document matrix W (one document per row, one term/word per column) and decompose it into two matrices, Z and H, where Z contains the distribution of topics (also called components) over documents, and H contains the distribution of terms over topics. The number of topics is a parameter that has to be chosen manually.

There exist several topic modeling algorithms that differ in the exact way of building W and deriving Z and H from it. Kuparinen and Scherrer (2024) propose to use non-negative matrix factorization (NMF), as well as an alternative probabilistic approach, latent Dirichlet allocation (LDA). In preliminary experiments, we have found NMF to provide better performance (in terms of the evaluation metrics presented in Section 4.3) and therefore focus on NMF here.

Topic models are generally used to identify documents with similar content, e.g., newspaper articles referring to sports, politics or culture. Documents are not assigned a single topic, but a probability distribution over all topics; an article can thus be characterized as 10% sports, 70% politics and 20% culture, for example.

4.2 Data processing and tokenization

In traditional applications of topic models, morphological variation is generally reduced by lemmatization or stemming, and function words are removed because they are not assumed to contribute to the content of a document. In contrast, we are interested in inferring variation patterns in the linguistic form, not in the content. We therefore take the transcriptions as they are, without any normalization or stopword removal. The only data preprocessing steps involve removing punctuation signs and lowercasing all text. The data is tokenized into whitespace-separated words, and we run experiments with single words and word bigrams.

We train NMF topic models with 2–5 components, using different partitions of the data. Words appearing in only one document were excluded from the modeling, but otherwise there were no limits on input. To summarize, we train topic models with the following parameters:

- all data vs. without legacy Lemko transcriptions,

Tokenization	Topics	All data			Without legacy Lemko data		
		Homogeneity	Completeness	V-measure	Homogeneity	Completeness	V-measure
Single words	2	0.5499	**0.8161**	0.6837	0.4729	0.8264	0.6362
	3	0.6019	0.5901	0.5948	0.8370	0.8358	0.8363
	4	0.9115	0.6844	0.7602	0.8292	0.6670	0.7236
	5	0.8668	0.5767	0.6659	0.8292	0.5843	0.6626
Word bigrams	2	0.5251	0.7792	0.6528	0.4487	0.7842	0.6037
	3	0.6049	0.5901	0.5959	**0.9410**	**0.9374**	**0.9388**
	4	**0.9577**	0.7177	**0.7976**	0.8770	0.7224	0.7772
	5	**0.9577**	0.6469	0.7434	0.8881	0.6322	0.7146

Table 2: Evaluation results with homogeneity, completeness and V-measure scores.

- single words vs. word bigrams,
- 2–5 components.

4.3 Evaluation

Topic model training is unsupervised and only relies on the linguistic material in the transcriptions. Given the assumed border effects, we expect the inferred topics to reflect national borders. Following Kuparinen and Scherrer (2024), we pick the dominant topic of each data point (i.e., the topic with the highest probability per transcription) and compute completeness, homogeneity and V-measure scores. These scores tell us how well the dominant topics coincide with the national borders:

- An experiment obtains maximum **homogeneity** (1.0) if all dominant topics only contain data points with the same variety label.
- An experiment obtains maximum **completeness** (1.0) if all data points with a given variety label show the same dominant topic.
- **V-measure** is the harmonic mean of homogeneity and completeness.

Table 2 shows the results. When using all data, the best V-measure is obtained with 4 or 5 topics, and word bigrams provide slightly higher scores than single words. When removing the legacy data, the best solution is clearly the one with 3 topics and bigram tokenization.

4.4 Visualization

We visualize a trained topic model as a map, where each document (an interview with a Rusyn speaker) is represented by a pie diagram depicting the distribution of topics. The legend shows which features (i.e., words or word bigrams, depending on the model) are most characteristic for each topic.

5 Results

5.1 Fieldworker Isoglosses

Figure 1 is based on all data encoded as word bigrams, with five topics. It shows that each variety – Lemko, Zakarpattia Rusyn, and Slovak Rusyn – is represented by numerous data points, with the most for Lemko. While there is relative homogeneity for the data points in Zakarpattia Ukraine and East Slovakia – each region is represented by one topic, Topic 2 for Zakarpattia Rusyn and Topic 3 for Slovak Rusyn –, it is striking that the Lemko data points are distributed across as many as three topics, meaning that the Lemko data accounts for 60% of the variation in the data according to the model shown here. Upon closer inspection of the different topics, it can be seen that each of the topics in the Lemko area contains one orthographic variant of the bigram *no i* 'but also'. It is written with the graphemes <i>, <ï>, and <и>, respectively. These orthographic differences do not correspond to actual dialect differences or different pronunciation habits, but are merely due to different transcription principles. This means that the method applied here is specifically suitable for uncovering different transcription principles in the dataset used for analysis.

In Figure 2, we only show Lemko data in a single-word model with two topics. While it seems that the orange dots in the west and in the east represent different homogenous dialect zones, the center of the Lemko dialect region around Gorlice show a confusing and not obvious pattern. This is due to the fact that, as mentioned before, the LEM dataset consists of both data specifically gathered for the *Corpus of Spoken Rusyn* and legacy data originally collected for other purposes. Even though the legacy data were converted to the Cyrillic script to better match

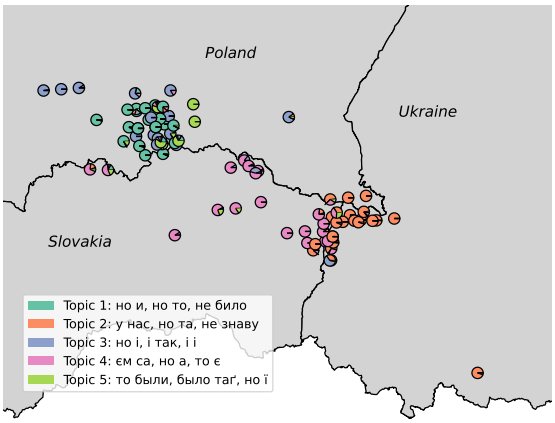

Figure 1: Distribution of five topics across all data with word bigrams. The Lemko dialect area is represented by three different topics (1, 3 and 5).

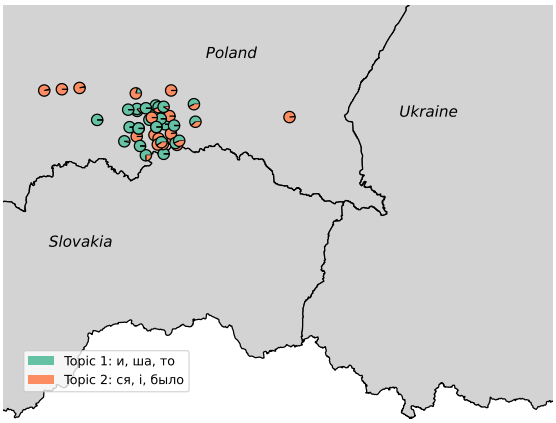

Figure 2: Two-topic solution of the Lemko-only data, clearly showing the different transcription principles of new and legacy data.

the rest of the data, the different transcription principles lead to data points from the same region being assigned to different topics.

Topic 1 features, among others, ша, apparently for the reflexive particle, while Topic 2 shows ся for the same feature. These are also merely different transcription principles as they both approximate [ɕa]. The data-driven approach applied here, thus, shows fieldworker isoglosses for the Lemko data, i.e., clusters that do not reflect any actual linguistic differences in the data, but rather differences in transcription conventions, which is in line with other data-driven approaches applied to this dataset (Rabus and Lahjouji-Seppälä, 2023).

While our method has proven to be effective for uncovering such fieldworker isoglosses, the main goal of the study is to evaluate the effectiveness of topic modeling for uncovering real dialect differences in a dataset, which is why we removed the legacy dataset for our remaining experiments.

5.2 Border Effects

As soon as the Lemko legacy data are excluded, the performance of the models significantly increases (see Table 2). The best models are those with three topics, and in particular the bigram model, shown in Figure 3. Here, we see an almost perfect distribution according to the three regions LEM, SLO and TRA. One exception is the outlier right at the Ukrainian side of the Ukrainian-Slovak border featuring Topic 2 (orange) instead of Topic 1 (green). Upon closer inspection, it turned out that this individual data point corresponds to an interview transcribed with a different transcription standard. Once again, this highlights the method's capability

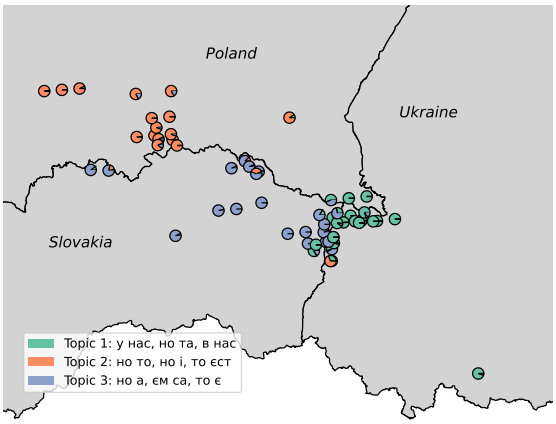

Figure 3: Word bigrams model with three topics, excluding the legacy Lemko data.

to uncover noisy data and fieldworker isoglosses.

The word bigrams that constitute these topics are linguistically highly plausible: Topic 1 features both у нас and в нас 'with us'. These are two orthographic variants of the East Slavic indirect *habeo*-construction. In the East Slavic languages, possession is predominantly expressed not by using a *habeo*-verb, i.e., some continuant of the Common Slavic verb **iměti* 'to have' and a direct object, but rather with an adessive construction ("with-me-there-is") and the nominative. Conversely, West Slavic languages exclusively use the construction with a *habeo*-verb, specifically *mieć* in Polish and *mať* in Slovak. In the Rusyn dialects, both variants are, in principle, possible. However, it becomes clear from the bigrams in Topic 1 that the Rusyn variety spoken in Zakarpattia Ukraine (TRA) adopted the adessive construction frequent in Standard Ukrainian, the umbrella language relevant for

TRA. The other topics for LEM (Topic 2) and SLO (Topic 3) do not include this feature, which is a clear sign of border effects between the Rusyn varieties roofed by West Slavic languages (LEM, SLO) and the variety roofed by an East Slavic language (TRA). One might wonder why there is no element with a *habeo*-verb – the other possible variable realization – in one of the other topics. This is because the model analyzed here considers word bigrams, and the bigrams with the *habeo*-verb exhibit significant variation due to the diverse lexemes used as objects.

Another opposition showing border effects is the bigram то єст 'that is' (Topic 2) versus то є (Topic 3). Here, there is also clear evidence of the influence of the respective umbrella languages, since the corresponding Polish bigram is *to jest* 'that is', while the Slovak one is *to je* 'that is'. The inflected Rusyn equivalents to English 'to be' in LEM and SLO (*jest* and *je*, respectively), thus, follow the Polish and Slovak patterns, respectively.

In Topic 3, there is one bigram єм ca, which is interesting both from a morphosyntactic and a phonetic viewpoint. Both the verb form єм 'I am' and the reflexive particle ca are clitic. Since ca follows єм directly, this means that it usually precedes the reflexive verb. This is typical for SLO and also possible in the West Slavic languages Polish and Slovak, while ся in Standard Ukrainian is a postfix and cannot precede the verb. According to Jabur et al. (2015, p. 311), the position of ся in the Slovak codification of Rusyn is identical to that of *sa* in the Slovak language. Apparently, the Rusyn dialects follow this pattern as well. Additionally, ca demonstrates the depalatalization of /s/, aligning with the Slovak example at the phonetic level.

6 Conclusion and Outlook

Our analysis of Rusyn dialects has shown that topic modeling is a promising novel method in computational dialectology that can be used for different purposes. It is data-driven and provides a bird's eye view for variant classification, but it also allows for zooming in to the levels of individual features in the different topics as well.

In-depth-analysis of the features of the individual topics has shown that – as opposed to typical use cases for topic modeling approaches – it is crucial *not* to exclude stopwords before analysis, since the most relevant linguistic differences between the individual topics are actually based on stopwords.

The experiments presented here show some method-inherent limitations that leave room for follow-up research. We discuss some perspectives below.

Reduce fieldworker isoglosses Since the method is sensitive to different transcription conventions, further research perspectives include conducting topic modeling analysis on normalized data and/or on data re-transcribed using state-of-the-art speech-to-text models.

Increase focus on morphology The experiments conducted here do not include any subword tokenization and consider whitespace-separated words as the minimal unit of analysis. This favors frequent word forms and neglects variation patterns that occur regularly, but as parts of different word forms, such as inflectional endings. Kuparinen and Scherrer (2024) experiment with character n-grams and unsupervised morphological segmentation to capture (concatenative) morphology. These extensions would be straightforward to apply to a morphologically rich language like Rusyn.

Neural topic models We used traditional topic modeling methods in order to easily experiment with different tokenization settings and to avoid any influence of external (pre-)training data. However, there is a wide range of neural approaches to topic modeling, some of which rely on embeddings from pretrained language models (for an overview, see e.g. Wu et al., 2024). As there are – to our knowledge – no such models specifically for Rusyn, it would be particularly instructive to assess the potential for cross-lingual transfer on the basis of multilingual language models trained on the closely related languages Polish, Slovak and Ukrainian. Multilingual embeddings could also be helpful for the automatic identification of corresponding n-grams from different topics.

More fine-grained evaluation At the moment, we use two metadata items provided in the *Corpus of Spoken Rusyn* for automatic evaluation: the country of the recording, and the project of origin (for distinguishing between legacy and new data). It would be interesting to assess the methods along other axes of variation and explore evaluation metrics that do not require ground truth labels, such as the silhouette coefficient (Rousseeuw, 1987).

Acknowledgements

This work is supported by the Research Council of Finland through project No. 342859 "CorCoDial – Corpus-based computational dialectology". Furthermore, it relies on work conducted in the project "Rusyn as a minority language across state borders: quantitative perspectives" funded by the German Research Foundation (RA 2212/2-2).

References

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 8(1):107–121.

Georgij Gerovskij. 1995. *Jazyk Podkarpatskoj Rusi*.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Vasyl' Jabur, Anna Pljiškova, and Kvetoslava Koporova. 2015. *Gramatika Rusyn'skoho Jazyka*, vydaňa perše edition. Vydavatel'stvo Prešovs'kej Univerzity, Prešov.

Olli Kuparinen and Yves Scherrer. 2024. Corpus-based dialectometry with topic models. *Journal of Linguistic Geography*, 12(1):1–12.

M. Zaidan Lahjouji-Seppälä, Achim Rabus, and Ruprecht von Waldenfels. 2022. Ukrainian standard variants in the 20th century: Stylometry to the rescue. *Russian Linguistics*, 46:217–232.

Alfred Lameli, Elvira Glaser, and Philipp Stöckle. 2020. Drawing areal information from a corpus of noisy dialect data. *Journal of Linguistic Geography*, 8(1):31–48.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Anna Plishkova. 2009. *Language and national identity: Rusyns south of Carpathians*, volume 14 of *Classics of Carpatho-Rusyn scholarship*. Columbia University Press and East European Monographs, New York.

Achim Rabus. 2019. Vergangenheitsbildung in gesprochenen karpatorussinischen Varietäten: Quantitativ-statistische Perspektiven. *Die Welt der Slaven*, 64(1):15–33.

Achim Rabus and M. Zaidan Lahjouji-Seppälä. 2023. Stilometrie, Transkription und Fieldworker Isoglosses: Aspekte der quantitativen Analyse slavischer Minderheitensprachkorpora. In Jan-Patrick Zeller, Thomas Menzel, and Hauke Bartels, editors, *Einheit(en) in der Vielfalt von Slavistik und Osteuropakunde*, pages 357–372. Lang, Berlin.

Achim Rabus and Andrianna Šymon. 2015. Na novŷch putjach isslidovanja rusyns'kŷch dialektu: Korpus rozhovornoho rusyns'koho jazŷka. In Kvetoslava Koporová, editor, *Rusyn'skŷj literaturnŷj jazŷk na Slovakiji: Zbornyk referativ z IV. Midžinarodnoho kongresu rusyn'skoho jazŷka*, pages 40–54. Prjašiv.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

H. A. Skrypnyk, editor. 2013. *Ukrajinci-Rusyny: Etnolinhvistyčni ta etnokul'turni procesy v istoryčnomu rozvytku*. Instytut mystectvoznavstva, fol'klorystyky ta etnolohiji im. M.T. Ryl's'koho, Kyjiv.

Michal Vašíček. 2020. *Dynamika jihokarpatských nářečí*, volume 51 of *Práce Slovanského ústavu AV ČR. Nová řada*. Slovanský ústav AV ČR, v.v.i., Praha.

Ruprecht von Waldenfels. 2014. Explorations into variation across slavic: Taking a bottom-up approach. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology, Typology, and Register Analysis*, Linguistic Variation in Text and Speech, pages 290–323. De Gruyter, Berlin, Boston.

Christoph Wolk and Benedikt Szmrecsanyi. 2018. Probabilistic corpus-based dialectometry. *Journal of Linguistic Geography*, 6(1):56–75.

Curt Woolhiser. 2005. Political borders and dialect divergence/convergence in Europe. In Peter Auer, Frans Hinskens, and Paul Kerswill, editors, *Dialect change*, pages 236–262. Cambridge Univ. Press, Cambridge.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(18).