Interactive maps for corpus-based dialectology

Yves Scherrer
Dept. of Informatics, University of Oslo
Dept. of Digital Humanities, University of Helsinki
yves.scherrer@ifi.uio.no

Olli Kuparinen
Languages Unit, Tampere University
olli.kuparinen@tuni.fi

Abstract

Traditional data collection methods in dialectology rely on structured surveys, whose results can be easily presented on printed or digital maps. But in recent years, corpora of transcribed dialect speech have become a precious alternative source for data-driven linguistic analysis. For example, topic models can be advantageously used to discover both general dialectal variation patterns and specific linguistic features that are most characteristic for certain dialects. Multilingual (or rather, multilectal) language modeling tasks can also be used to learn speaker-specific embeddings. In connection with this paper, we introduce a website that presents the results of two recent studies in the form of interactive maps, allowing visitors to explore the effects of various parameter settings. The website covers two tasks (topic models and speaker embeddings) and three language areas (Finland, Norway, and German-speaking Switzerland). It is available at https://www.corcodial.net/.

1 Introduction

The traditional data collection method in dialectology has relied on structured surveys conducted in a particular language area. The results of such surveys can be presented in maps, typically one map per linguistic feature. These collections of maps, known as dialect atlases, are an important source of information about dialect divisions of different languages. For instance, the dialect atlas of Lauri Kettunen (Kettunen, 1940) still forms the basis of the division of Finnish dialects, even though it was collected almost 100 years ago.

As this example shows, dialect atlases were typically conceived in the first half of the 20th century and presented as paper maps. This poses problems of accessibility for modern dialectology, where computational models are often applied on dialect data, e.g. in the subfield of dialectometry (Goebl, 2011). Some atlases have already been digitized and can thus be used in computational analyses (Embleton and Wheeler, 1997; Scherrer and Stoeckle, 2016; Syrjänen et al., 2016). When digitized, the atlases are typically presented as two-dimensional data tables where the columns present linguistic features and the rows locations. Digitized atlases also make interactive visualizations possible (Scherrer, 2023).

In our recent research, we have experimented with topic modeling (Kuparinen and Scherrer, 2024) and representation learning (Kuparinen and Scherrer, 2023) to explore the dialectal divisions arising from corpora instead of atlases. Dialect corpora typically consist of spoken data (mostly interviews) which have been phonetically transcribed. Compared to the straightforward two-dimensional tabular data presented in dialect atlases, corpus data is more difficult to analyze computationally, because individual characteristics of speakers (addressed topics, length of interview, richness of vocabulary, etc.) are mixed with dialect features.

In the following sections, we briefly present the data and experiments, while focusing on the interactive website visualizing the results.

2 Data

We work with three datasets consisting of dialect interviews or conversations, which have been both phonetically transcribed and normalized to a standard variety. The datasets cover the Finnish, Norwegian and Swiss German language areas.

While the topic modeling experiments (Section 3.1) only make use of the phonetic transcriptions, the representation learning study (Section 3.2) is based on a dialect-to-standard normal-

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 634–638
March 3-4, 2025

ization task and uses both transcription layers.

2.1 Samples of Spoken Finnish

The Finnish data used in the experiments and visualized on the website comes from the Samples of Spoken Finnish corpus (fi. *Suomen kielen näytteitä*, SKN).[1] The corpus consists of interviews recorded in the 1960s and 1970s in 50 Finnish-speaking locations (Institute for the Languages of Finland, 2021). There are two speakers per location (with one exception) and approximately one hour of speech per person. The interviews were phonetically transcribed by professionals and normalized manually to standard Finnish. In total, the corpus contains 99 interviews and represents traditional Finnish dialects comprehensively.

2.2 Norwegian Dialect Corpus

For Norwegian, we use a subset of the Nordic Dialect Corpus (Johannessen et al., 2009), which contains spoken language data from the North Germanic languages.[2] The Norwegian part (named Norwegian Dialect Corpus, NDC) is the largest and most thorough in transcription of the different subcorpora. There are 684 interviews (either with a single interviewee or with several) and 438 individual interviewees. For our experiments and visualizations, each data point represents the concatenation of all productions of one interviewee. The recordings were made between 2006 and 2010 and included speakers of different age groups. The recordings were phonetically transcribed and normalized to Bokmål.

2.3 ArchiMob Corpus (Swiss German)

The Swiss German data comes from the ArchiMob corpus (Samardžić et al., 2016; Scherrer et al., 2019), which consists of interviews conducted between 1999 and 2001.[3] It contains 43 phonetically transcribed interviews, which are used for the topic modeling experiments. We do not use this corpus for the representation learning experiments, since only six interviews were normalized manually (and the rest automatically).

[1] http://urn.fi/urn:nbn:fi:lb-2021112221, Licence: CC-BY.

[2] http://www.tekstlab.uio.no/scandiasyn/download.html, Licence: CC BY-NC-SA 4.0.

[3] https://www.spur.uzh.ch/en/departments/research/textgroup/ArchiMob.html, Licence: CC BY-NC-SA 4.0.

3 Experiments

3.1 Topic modeling

The topic modeling experiments are conducted on all three datasets presented in Section 2. We used two topic modeling techniques and five tokenization techniques to explore the dialect divisions of the three focus languages. The used models were non-negative matrix factorization (NMF; Paatero and Tapper 1994) and latent Dirichlet allocation (LDA; Blei et al. 2003), while the tokenizations were complete words, character n-grams from 2 to 4, and Morfessor-based subword tokenization (Smit et al., 2014). A more thorough explanation of the methodology and best results can be found in Kuparinen and Scherrer (2024).

3.2 Representation learning

In the second experiment, we trained a neural machine translation model to "translate" the phonetic transcriptions to standardized spelling. We used a relatively standard setup based on the Transformer architecture (Vaswani et al., 2017) and subword tokenization with BPE (Sennrich et al., 2016). Taking inspiration from multilingual translation modeling (e.g. Johnson et al., 2017), the speaker ID was added as the first token of each utterance on the source side. After training the model, we extracted the learned embeddings of these speaker IDs and used them as input data for three dimensionality reduction algorithms.

The dimensionality reduction algorithms were principal component analysis (PCA; Hotelling 1936), k-means clustering (MacQueen, 1967) and Ward agglomerative clustering (Ward Jr., 1963). The PCA is run with three principal components for visualization purposes (each component represented as a color in the RGB color scheme), while the clustering algorithms are run with the number of clusters ranging from 2 to 20. For further information on the experimental design and a quantitative evaluation of the clustering algorithms, see Kuparinen and Scherrer (2023).

4 Visualization

The website https://www.corcodial.net/ provides interactive visualizations of the two experiments described in the previous section. The maps are drawn with the Leaflet[4] mapping toolkit. The map backgrounds use the *Stamen*

[4] https://leafletjs.com/

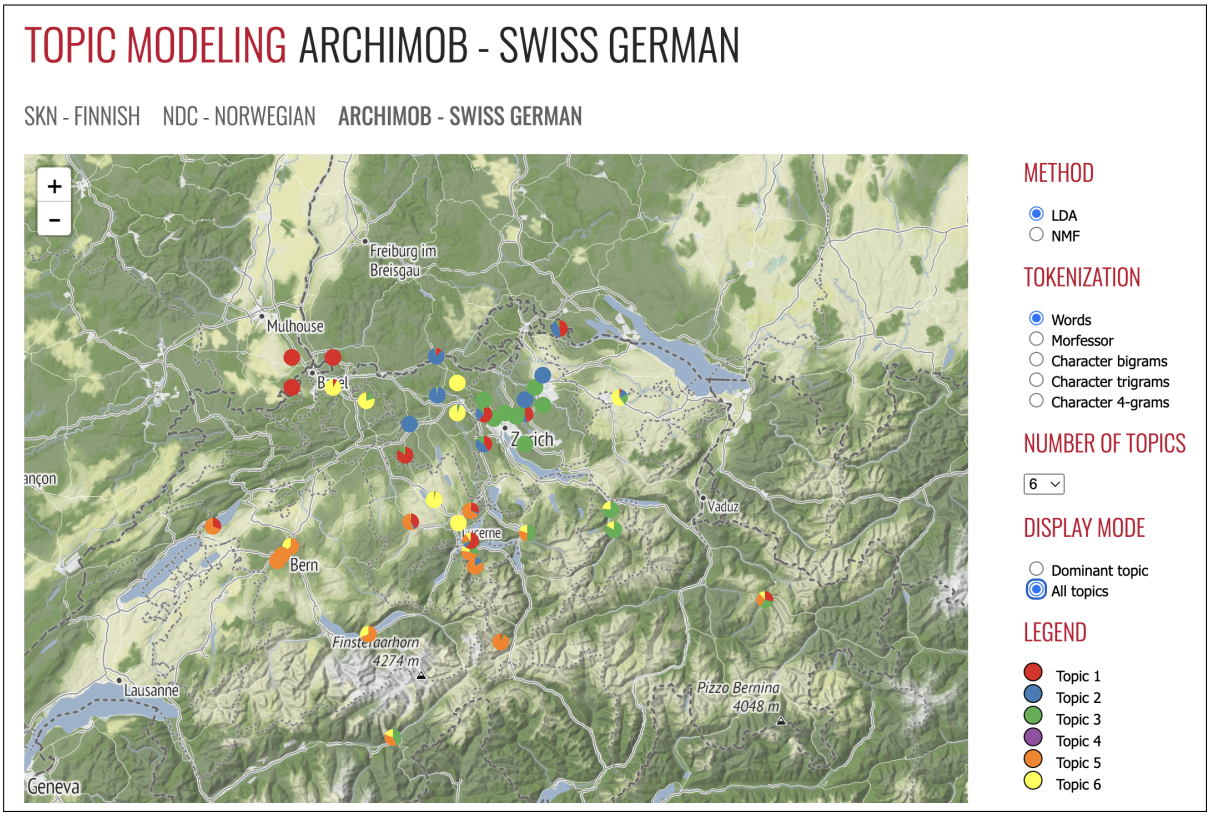

Figure 1: Interactive visualization of a topic modeling experiment for Swiss German. Each point represents one interview. The colored pie charts reflect the degree of membership in the different topics.

terrain style from Stadia Maps,[5] which are based on OpenStreetMap data.[6] The server-side backend is implemented in Flask.[7] All these libraries and sources are licensed under Creative Commons or other open source licenses. The current setup does not require any database, since all the data is available in precomputed CSV or JSON files.

Figure 1 shows a screenshot of a **topic modeling** experiment. The map itself takes up most of the screen, whereas the rightmost part is reserved for user interaction (e.g. to select a different parameter) and metadata display (e.g. the legend associating colors to topics). Each point on the map corresponds to one interview and each color corresponds to one inferred topic. The main benefit of topic models is that an interview can "belong" to several topics to varying degrees. The pie charts on the maps show the degree of membership to the different topics. A simpler visualization that only shows the dominant topic for each interview, is available by selecting *Dominant topic*. Further information about the composition of the topics (i.e., the tokens most strongly associated with each topic) can be shown in a popup window (not shown in Figure 1).

Technically, and quite similarly to geographic information systems in general, such a visualization relies on two data files: a corpus-specific GeoJSON file that describes the points (with their coordinates and IDs), and a task-specific JSON file that contains the distribution of topics for each point. Leaflet makes it easy to add the GeoJSON file as an additional layer on top of the map background, and to define the style (e.g. the colors) of each point based on the JSON file.

A particularity of corpus-based analyses is that there can be several interviewed persons from the same place, and the corresponding points on the map would be superimposed. The current implementation detects superimposed points and moves them away from their original locations to ensure their visibility. We plan to further improve this functionality.

The **representation learning** experiment is illustrated by Figures 2 and 3. Figure 2 shows a PCA of the speaker embeddings of the Norwegian NDC dataset. As is commonly the case

[5] https://stadiamaps.com/
[6] https://www.openstreetmap.org
[7] https://flask.palletsprojects.com/

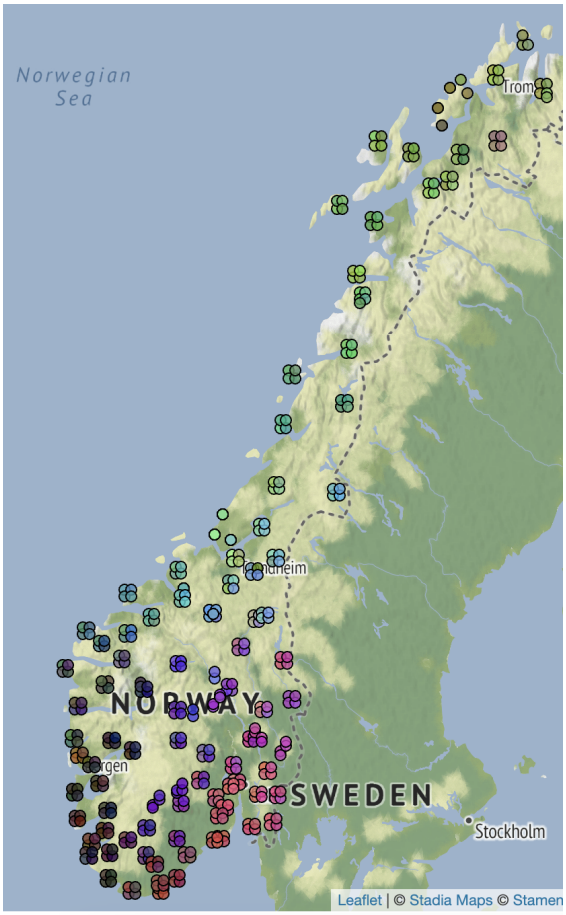

Figure 2: PCA map fragment of the learned speaker representations of the NDC dataset. The three PCA components correspond to the red–green–blue components of the colors.

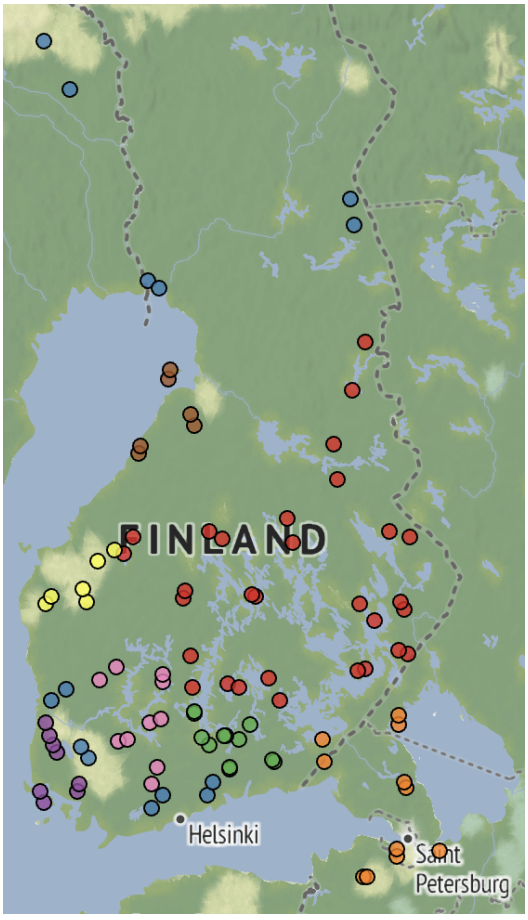

Figure 3: Cluster map of the learned speaker representations of the SKN dataset. The clustering is created with the Ward algorithm and displays 8 clusters.

with dimensionality reduction techniques, the map clearly shows the major dialect areas (Southwestern dialects in dark brown, Eastern dialects in red, central dialects in light blue and northern dialects in light green), without showing clear-cut borders between the areas.

Figure 3, on the other hand, visualizes the speaker embeddings of the Finnish SKN dataset. In this case, a hierarchical clustering algorithm has been selected. The result shows clearly identifiable dialect areas corresponding relatively well with atlas-based divisions.[8] An exception is the cluster represented in blue on the map, which includes points in the Greater Helsinki area, in a transition area in the Southwest, as well as in Northern Finland. At the moment, the visualization website supports two clustering algorithms (Ward and K-means) and any number of clusters between 2 and 10.

5 Conclusion

Following up on our recent research where we propose to use topic modeling and representation learning to explore the dialectal divisions arising from corpora of transcribed dialect speech, we present an interactive website where it is possible to view the experimental results in the form of maps. Different parameter settings and modes of visualization can be easily chosen.

At the moment, the website covers two tasks (topic modeling and representation learning) and three linguistic areas (Finland with the SKN corpus, Norway with the NDC corpus, and German-speaking Switzerland with the ArchiMob corpus). The design of the website is modular and permits the easy inclusion of additional tasks, language areas and corpora.

[8] The dendrogram of the hierarchical clustering can be displayed on demand (not shown here).

Acknowledgements

This work is supported by the Research Council of Finland through project No. 342859 "CorCoDial – Corpus-based computational dialectology" and project No. 360356 "Speech as speech – acoustic modeling in variational linguistics".

References

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

Sheila Embleton and Eric S. Wheeler. 1997. Finnish dialect atlas for quantitative studies. *Journal of Quantitative Linguistics*, 4(1-3):99–102.

Hans Goebl. 2011. Dialectometry and quantitative mapping. In Alfred Lameli, Roland Kehrein, and Stefan Rabanus, editors, *An International Handbook of Linguistic Variation*, volume 2, pages 433–464. De Gruyter Mouton, Berlin, New York.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Institute for the Languages of Finland. 2021. Samples of Spoken Finnish, VRT Version.

Janne Bondi Johannessen, Joel James Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus – an advanced research tool. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 73–80, Odense, Denmark. Northern European Association for Language Technology (NEALT).

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Lauri Kettunen. 1940. *Suomen murteet. 3, A, Murrekartasto*. Suomalaisen Kirjallisuuden Seuran toimituksia ; 188. Suomalaisen kirjallisuuden seura, Helsinki.

Olli Kuparinen and Yves Scherrer. 2023. Dialect representation learning with neural dialect-to-standard normalization. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 200–212, Dubrovnik, Croatia. Association for Computational Linguistics.

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

J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967).

Pentti Paatero and Unto Tapper. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob - a corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066, Portorož, Slovenia. European Language Resources Association (ELRA).

Yves Scherrer. 2023. dialektkarten.ch – Interactive dialect maps for German-speaking Switzerland and other European dialect areas. In Thomas Krefeld, Stephan Lücke, and Christina Mutter, editors, *Berichte aus der digitalen Geolinguistik (II)*, volume 9 of *Korpus im Text*. Ludwig-Maximilians-Universität München, Germany.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. Digitising Swiss German: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53(4):735–769.

Yves Scherrer and Philipp Stoeckle. 2016. A quantitative approach to Swiss German – Dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, 24(1):92–125.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden. Association for Computational Linguistics.

Kaj Syrjänen, Terhi Honkola, Jyri Lehtinen, Antti Leino, and Outi Vesakoski. 2016. Applying population genetic approaches within languages: Finnish dialects as linguistic populations. *Language Dynamics and Change*, 6(2):235 – 283.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Joe H. Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.