Dialect Representation Learning with Neural Dialect-to-Standard Normalization

Olli Kuparinen and **Yves Scherrer**
Department of Digital Humanities
University of Helsinki
olli.kuparinen@helsinki.fi, yves.scherrer@helsinki.fi

Abstract

Language label tokens are often used in multilingual neural language modeling and sequence-to-sequence learning to enhance the performance of such models. An additional product of the technique is that the models learn representations of the language tokens, which in turn reflect the relationships between the languages. In this paper, we study the learned representations of dialects produced by neural dialect-to-standard normalization models. We use two large datasets of typologically different languages, namely Finnish and Norwegian, and evaluate the learned representations against traditional dialect divisions of both languages. We find that the inferred dialect embeddings correlate well with the traditional dialects. The methodology could be further used in noisier settings to find new insights into language variation.

1 Introduction

Starting with Johnson et al. (2017), multilingual neural models have become increasingly popular for both language modeling and sequence-to-sequence learning tasks. The most common type of multilingual model makes use of language labels that are prepended to the training and test instances to inform the model about the language being processed. The embeddings of the language models can then be analyzed to find emerging properties of the relationships between the languages (Östling and Tiedemann, 2017).

In this paper, we apply the same idea to a smaller granularity of linguistic variation, namely dialectal variation within a language area, and we use dialect-to-standard normalization as the modeling task. Focusing on two typologically different languages, we experiment with large datasets of Finnish and Norwegian dialects. We study the inferred dialect embeddings with different dimensionality reduction algorithms to see whether the neural normalization models learn dialectal differences. We find that the learned representations correlate well with the traditional dialect classifications.

2 Related Work

2.1 Representation Learning in Multilingual and Multidialectal Settings

Johnson et al. (2017) present a simple approach to multilingual machine translation that relies on additional input tokens signalling the model which target language it is supposed to generate. While they find interesting benefits of this approach (e.g., zero-shot translation), they do not specifically analyze the internal representations of the language labels. In contemporary work, Östling and Tiedemann (2017) analyze the structure of the language embedding space obtained from a multilingual language model. They find for example that the inferred clustering of Germanic languages corresponds closely to the established genetic relationships.

Abe et al. (2018) combine these two lines of research and apply them to dialectal data. Their training material includes texts from 48 Japanese dialects, each of which is aligned with the standard variety. They introduce a multi-dialectal neural machine translation model translating between the dialects and standard Japanese. Besides the practical benefits of dialect-to-standard and standard-to-dialect translation, the induced dialect label embeddings can be used for dialectometric analyses. For instance, they find that the clusters inferred from the dialect embeddings correspond to the major dialect areas of Japan. In this work, we apply a similar method to Finnish and Norwegian dialects.

Instead of training multi-dialectal translation or language models, Hovy and Purschke (2018) use a topic modelling approach to learn continuous document representations of cities in a large corpus of online posts from the German-speaking area. These city embeddings reflect the major German dialect areas according to earlier dialectological

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 200–212
May 5, 2023 ©2023 Association for Computational Linguistics

research.

2.2 Dialect-to-Standard Normalization

The dialect-to-standard translation task, often also referred to as dialect normalization, has been independently researched for a number of dialect areas, e.g., Swiss German (Scherrer and Ljubešić, 2016; Honnet et al., 2018), Finnish (Partanen et al., 2019) or Estonian (Hämäläinen et al., 2022). Most commonly, statistical or neural character-level machine translation models are used for this task.

Methodologically, dialect normalization is closely related to historical text normalization, and recent work in this field has notably investigated the optimal word segmentation strategies and hyperparameters (Bollmann, 2019; Wu et al., 2021; Bawden et al., 2022). We take these recent findings into account in our experiments.

2.3 Finnish and Norwegian Dialects

Both Finnish and Norwegian boast differing dialects which are used in everyday speech. There is also a long dialectological tradition for both languages, which is visible in the amount of available dialect corpora. In addition to the datasets used in this work (see Section 3), there are, for instance, the LiA corpus of historical dialect recordings in Norwegian (Hagen et al., 2021) and the Finnish Dialect Syntax archive (University of Turku and Institute for the Languages of Finland, 1985).

The dialects of Finnish are traditionally divided into Eastern and Western dialects (see Figure 1) and to eight more fine-grained dialect areas. The division is mostly based on Kettunen (1940) and explicitly defined in e.g., Itkonen (1989). We use this eight-dialect division for the evaluation of our representation learning.

The dialects of Norwegian are divided into four dialect areas: Western, Eastern, Central (or Trøndersk) and Northern dialects (Hanssen, 2010 - 2014), which in turn have several subgroups. We use the four-dialect division for evaluation. The dialect divisions for both languages are presented in Figure 1.

3 Data

3.1 Samples of Spoken Finnish

The Samples of Spoken Finnish corpus (fi. *Suomen kielen näytteitä*, SKN) is a collection of interviews conducted mostly in the 1960s (Institute for the Languages of Finland, 2021).[1] The corpus includes 99 interviews from 50 locations (2 for each location, with one exception) and presents the dialects of Finnish comprehensively. The key figures of the dataset are described in Table 1.

The interviews have been transcribed with the Uralic Phonetic Alphabet (UPA) on two levels of precision: a detailed transcription with diacritics and a simplified version which relies mostly on standard Finnish characters. We use the simplified transcriptions and only the utterances of the interviewees, not the interviewers. The transcriptions have been manually normalized to standard Finnish. The detailed transcriptions have been used for dialect-to-standard normalization in Partanen et al. (2019).

3.2 Norwegian Dialect Corpus

The Norwegian Dialect Corpus (Johannessen et al., 2009) consists of interviews and informal conversations recorded in Norway between 2006 and 2010.[2] The corpus was collected as part of a larger study focusing on the dialectal variation of the North Germanic languages. The recordings come from 111 locations, with 438 speakers appearing in total. The same speakers appear in interviews and conversations with each other. We use the utterances of both contexts. The size of the dataset is described in Table 1.

The recordings have been phonetically transcribed and normalized to Bokmål (one of the standard languages for Norwegian). The normalization has been conducted semi-automatically: first with an automatic tool and thereafter manually checked.

The publicly available transcriptions and normalizations are not well aligned: the number of utterances is not identical, only one of the two layers contains quotation marks, and the orthographic transcriptions for some utterances are missing. We automatically re-align the transcriptions and normalizations before using them in our experiments.[3]

4 Experimental Setup

4.1 Preprocessing

We remove punctuation and pause markers from the transcriptions and normalizations, and exclude

[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] The re-aligned version of NDC is available at https://github.com/Helsinki-NLP/ndc-aligned.

		Speakers	Locations	Texts	Sentences	Words
SKN	(Samples of Spoken Finnish)	99	50	99	41,407	630,665
NDC	(Norwegian Dialect Corpus)	438	111	684	126,460	1,684,059

Table 1: The sizes of our two datasets.

Dialect	mie poikain kans olen kahen teäl
Standard	minä poikani kanssa olen kahden täällä
Dialect-BPE	<SKN34_Markkova> mi@@ e po@@ i@@ ka@@ in kan@@ s ol@@ en ka@@ h@@ en te@@ ä@@ l
Standard-BPE	minä po@@ i@@ ka@@ ni kan@@ ssa ol@@ en ka@@ hd@@ en tä@@ ä@@ llä
English gloss	‘Me and my son are alone here.’

Table 2: An example sentence from the Finnish dataset, with the source and target on top, preprocessed source and target (i.e. BPE-encoded and source label added) in the middle, and an English gloss below. The label in the beginning of the source identifies the speaker, and the embeddings learned on these label tokens are used for the analyses.

utterances that only include filler words (such as *mm, aha*, for instance). For NDC, we substitute all anonymized name tags with a capital X. The names in SKN are not anonymized, and we thus leave them as they are. Each speaker's utterances are split so that 80% of sentences are used for training, 10% of sentences are used for the development, and 10% of sentences are set aside for testing.

Following recent findings in historical text normalization (e.g., Tang et al., 2018; Bawden et al., 2022), we work on subword tokens instead of characters. We segment our data with the byte-pair encoding (BPE; Sennrich et al., 2016) algorithm. The number of merge operations is set to 200, following Gutierrez-Vasques et al. (2021). The vocabulary is shared between the source and the target. This results in a vocabulary of 336 tokens for SKN and 360 tokens for NDC. The vocabularies were evaluated qualitatively and they include meaningful units such as case markers and other morphological units for Finnish, as well as frequent words such as pronouns for both languages. Further tuning of the vocabulary size could anyhow enhance the results.

We add a speaker label at the beginning of each utterance. Note that labels generally indicate the target variety, whereas in our setup they represent the source variety. The target variety is fixed to be the standard. Therefore, the labels are not necessary for successful normalization, but we use them here to infer the speaker representations. An example of our preprocessing is shown in Table 2.

4.2 NMT Model Setup

Our NMT model is a classical Transformer with 6 encoder and decoder layers, vector size 512, and 8 attention heads each (Vaswani et al., 2017). We enabled position representation clipping because we found it to be beneficial in preliminary experiments. The models were trained for 100,000 steps with a batch size of 5000 tokens and gradient accumulation over 8 batches, and an initial learning rate of 4. The models were trained with the OpenNMT-py (Klein et al., 2017) toolkit with the default settings for all other parameters.[4]

4.3 Dimensionality Reduction

After training the NMT model, we obtain the embedding vectors for each of the speaker labels. This results in a matrix with 99 (SKN) or 438 (NDC) rows and 512 columns.

We run three dimensionality reduction methods on the matrices: a principal component analysis (PCA; Hotelling 1933), a k-means clustering (MacQueen, 1967), and hierarchical agglomerative clustering with Ward linkage (Ward, 1963). All methods are run on the scikit-learn toolkit (Pedregosa et al., 2011).

The PCA is used to visualize the dialect continuum (see 4.4). Because the visualization relies on three color channels (red, green, and blue), the PCA is run with three components, each being represented by one color. Both k-means and Ward clustering are run with the number of clusters ranging from 2 to 20, and the clusterings are evaluated

[4] We did initial testing with an RNN-based model as well, but the results were considerably better with the Transformer.

with the methodology described in Section 4.5. The number of clusters was defined by preliminary experiments, which showed that increasing the number above 20 did not enhance the results. K-means clustering is averaged over five runs, since it is known to fluctuate.

4.4 Visualization

The PCA weights are normalized to values between 0 and 1 and used to present the red, green and blue colors in a map visualization (Nerbonne et al., 1999). For example, having values such as 0.5 for PC1, 0.25 for PC2, and 0.75 for PC3 would translate to 128 on the red channel, 64 on the green channel, and 192 on the blue channel, since the maximum value per color is 256. Having a color channel for each of the three components therefore translates to a single color (purple in the example case). The method is used to create Figure 2. A similar approach has been presented in Hovy and Purschke (2018), and an often used technique in dialectometry called multidimensional scaling (MDS) functions on the same principle but with distance matrices (Nerbonne et al., 1999; Leinonen et al., 2016).

The best clustering results are also presented on maps. The map visualizations are created with QGIS (QGIS Development Team, 2023). For the Ward clustering results, we present the dendrograms (see Figure 6), which show the relations between clusters. The dendrograms are created with scipy (Virtanen et al., 2020) and matplotlib (Hunter, 2007) toolkits.

4.5 Evaluation

We evaluate the normalization performance on the development sets to ensure that our models are working as expected. We compare our results to Partanen et al. (2019), who produce a good baseline for the SKN dataset, even though they use the detailed transcriptions and different preprocessing[5] in their work. Since they evaluate their model performance on word error rate (WER), we use the same metric for the comparison.[6]

We evaluate the clusters produced by k-means and Ward primarily with V-measure (Rosenberg and Hirschberg, 2007). V-measure is the harmonic mean of homogeneity (how homogeneous the produced clusters are in terms of predefined classes) and completeness (how well the predefined classes stay complete in the clustering). Completeness is typically higher with fewer clusters (there are less clusters for the classes to spread out into) and homogeneity with a higher number of clusters (the clusters do not include as many classes). V-measure can thus be seen as an equivalent of F_1-score and homogeneity and completeness as precision and recall. The difference is that V-measure does not expect there to be an exact right number of clusters. The V-measure score is between 0 and 1, with 1 being a perfect match between the gold labels and the clustering solution.

As a more traditional metric, we also present the adjusted Rand index (Rand, 1971). As V-measure, the adjusted Rand index tries to compute the similarity between the gold labels and the predicted labels of a clustering algorithm. Mathematically, Rand index presents the probability that a randomly chosen pair of elements from the gold labels and the predicted labels will agree. The adjusted Rand index (ARI) is typically used instead of the plain version, as it is corrected for chance. The ARI score is between -1 and 1, with 0 being a random prediction and 1 being a perfect match. Scores below 0 are worse than the random baseline. Both V-measure and ARI are computed with the scikit-learn toolkit (Pedregosa et al., 2011).

We evaluate the clusterings against traditional dialect divisions. For Finnish, we use the eight-way classification presented in Itkonen (1989). For Norwegian, the ground truth is the four-way divide presented in Hanssen (2010 - 2014). The dialect divisions are presented in Figure 1. We compare our results to a geographically and administratively defined baseline, namely the regional units of Finland (NUTS3 in European Union Nomenclature), and the counties used in Norway from 1972 to 2018.[7]

5 Results

5.1 Normalization Performance

The word error rates for our models and for Partanen et al. (2019) are presented in Table 3. Our SKN model produces a similar, albeit slightly worse, score than in their work. As far as we are aware, there is no existing work on the normalization of

[5] On top of the different transcriptions, they use a character-level neural machine translation model with an RNN-architecture, and split the data to chunks of three words (non-overlapping trigrams).

[6] We use `https://github.com/nsmartinez/WERpp` for calculating the WER, as do Partanen et al. (2019).

[7] The number of counties was reduced from 19 to 11 in 2018.

	SKN	NDC
Partanen et al. (2019)	5.73	—
This work	6.11	4.89

Table 3: Word error rates (↓) for Partanen et al. (2019), our SKN model, and our NDC model.

the NDC dataset, and thus the score can not be compared. Achieving a similar score as Partanen et al. (2019) for Finnish, and a lower one for Norwegian, does not indicate issues with the model performance, and the learned representations of the speakers can therefore be used for further analysis.

5.2 Principal Component Analysis

Dialects create a continuum, with either subtle transitions from one area to another, or stronger borders between them. For instance in Finnish dialectology, a strong border is seen between the Western and Eastern dialects and smaller differences inside these large areas. To analyze whether the neural models have learned such differences, we run a three-component principal component analysis on the learned speaker embeddings.

Three components are chosen for visualization purposes, as each of the three components are presented with their own color on a map visualization. The speakers' locations are plotted on the map, and the degree of each component in each speakers' interview is presented as red, green, and blue colors, as explained in Section 4.4. Thus, similar hues indicate linguistic similarity of the speakers, and the degree of color change from one area to another indicates the degree of linguistic difference. The results of the principal component analysis are presented in Figure 2. The Finnish and Norwegian results are presented in the same figure for convenience, but the analysis is separate for both languages.

The explained variance of the principal component analysis model is low for both languages (14% for Finnish and 9% for Norwegian). We hypothesize this is due to the used data: we are working on the embedding space of the normalization model, which may include manifold variation, for example relating to the actual normalization task. The explained variance may thus not be as good a measure here as it is for multi-dimensional scaling, for instance, which works on distance matrices. Limiting the model to three components due to visualization might also affect the explained variance.

We commence with an analysis of the Finnish speakers in Figure 2. There are clearly differing areas in the South-West (bright green), South-East (light green), South-East Häme (blue), and Savo (red). The South-West, South-East, and Savo are traditional dialect areas, but South-East Häme has been traditionally seen as a part of a larger Häme area (dark blue in Figure 2). The shade of the blue thus indicates that South-East Häme, although related to the rest of Häme, is somewhat different from it. Regarding the transitions from one area to another, there is a clear difference between East and West in the South and center of the country (from blue to red), but not as big a difference in the North. This reflects the understanding that the Northern dialects are a combination of Western and Eastern influence (e.g., Leino et al. 2006).

For Norwegian, the color changes in Figure 2 are more subtle than for Finnish, indicating transitional areas between the dialects. There is a clearly red area (PC1) around Oslo, a purple cluster (PC1 and PC3) in the center of the country and dark hues in the West. The Trøndersk area in the middle has a cyan quality (PC2 and PC3), which turns green (PC2) in the North and yellow in the far North (Finnmark). Regarding the four-way division of Eastern, Western, Trøndersk, and Northern dialects, the map shows that there is internal variation in the areas.

5.3 Clustering Evaluation

We run k-means clustering and agglomerative clustering with Ward linkage on the learned speaker representations to examine whether the methodology captures similar divisions as in traditional dialectology. We evaluate each clustering with the number of clusters ranging from 2 to 20, and compare them to dialect divisions presented in the past, as explained in Section 4.5. We use V-measure and adjusted Rand index as metrics, and a geographically and administratively defined baseline against which to compare the clustering performance. The results for both methods and datasets are presented in Figure 3 and Figure 4. In case of ambiguity between the V-measure and adjusted Rand index, we prefer the V-measure.

Figure 3 and Figure 4 show that agglomerative clustering with Ward linkage outperforms k-means on both datasets, with the difference being clearer for Finnish. Similar findings have been reported before (Heeringa, 2004; Prokić and Nerbonne, 2008;

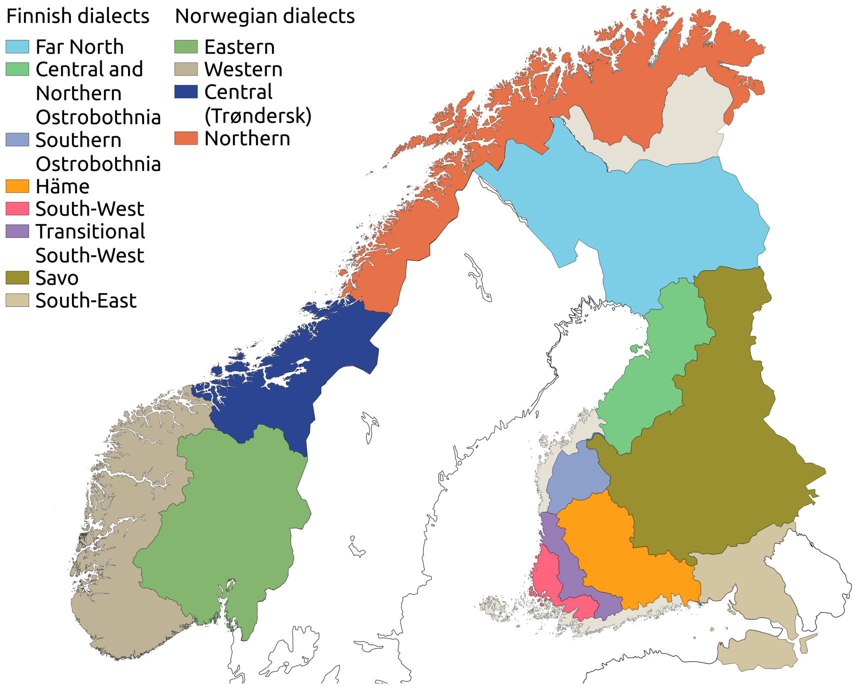

Figure 1: The dialect areas used as gold labels. The Norwegian division is based on Hanssen (2010 - 2014) and the Finnish one on Itkonen (1989).

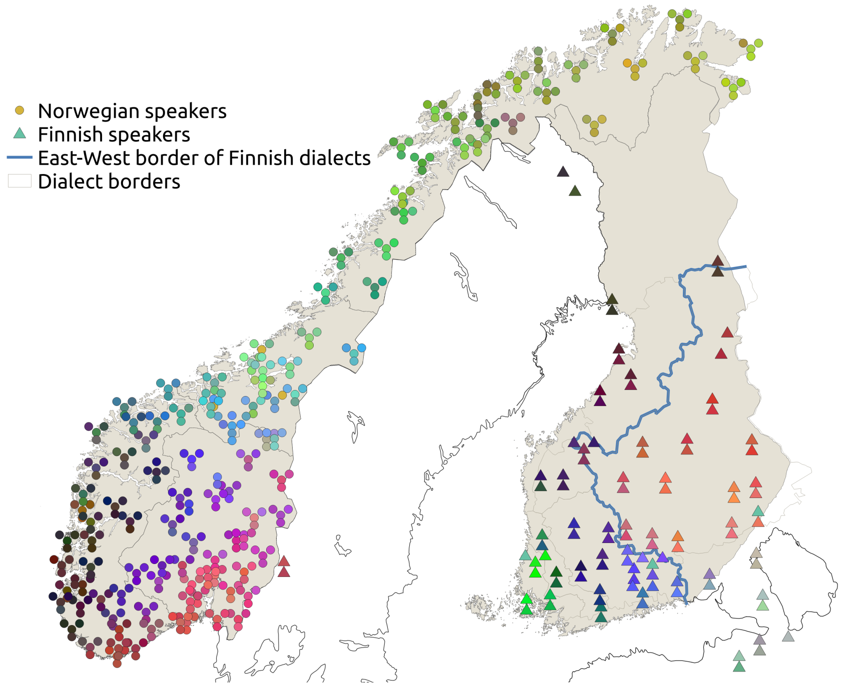

Figure 2: Visualization of a three-component principal component analysis. The Norwegian speakers are presented with circles and Finnish speakers with triangles. The dialect areas that are used as ground truth are presented with thin grey lines. The first principal component is presented as red, second as green, and third as blue. The color shade of each speaker is thus a combination of these three colors. Note that the PCA is different for both languages, and they are presented side by side because of geographical proximity. Also note that there are two locations of Finnish in Sweden (in the North and in the far West, close to the Norwegian border.)

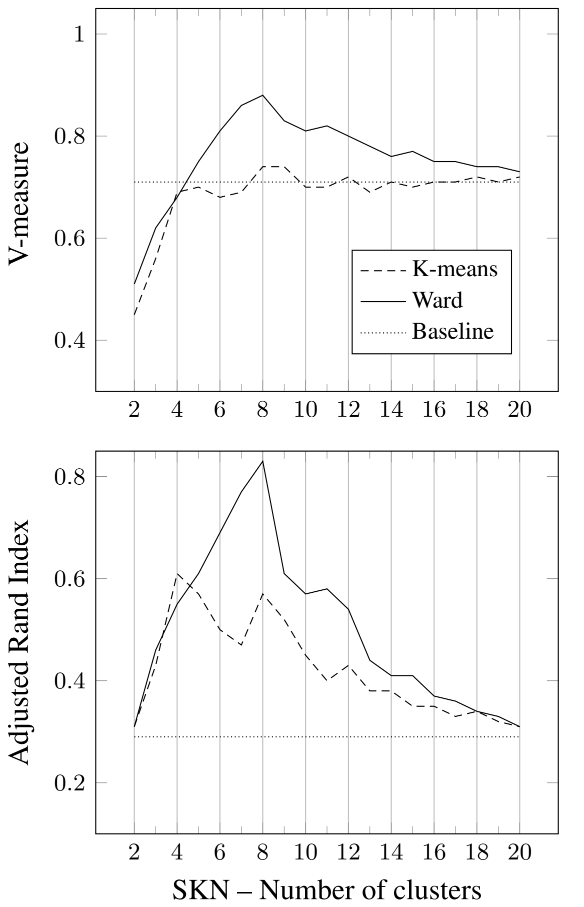

Figure 3: Evaluation of the clustering methods on the SKN dataset. K-means averaged over five runs. Baseline is presented as a horizontal line.

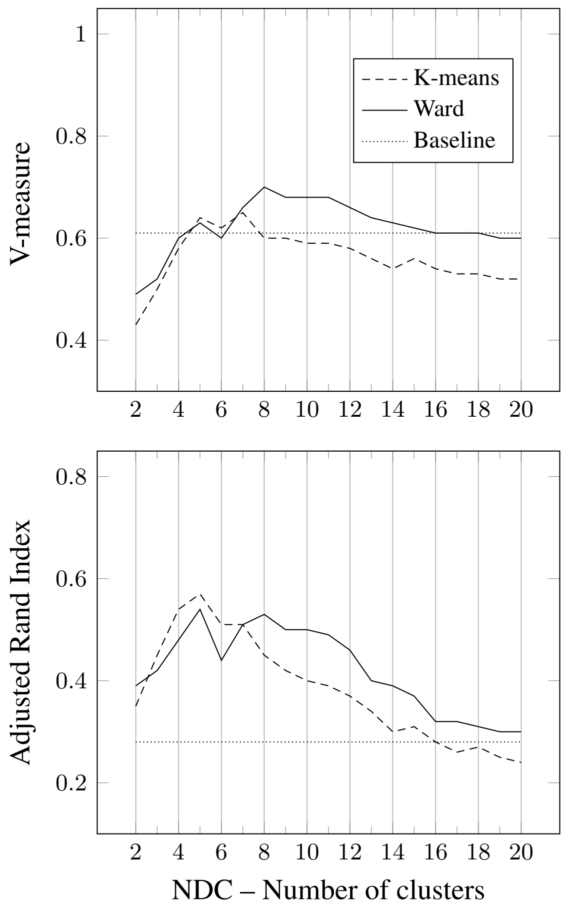

Figure 4: Evaluation of the clustering methods on the NDC dataset. K-means averaged over five runs. Baseline is presented as a horizontal line.

Hovy and Purschke, 2018). The scores are also generally worse for Norwegian, with the models barely outperforming the V-measure baseline. The best V-measure scores are achieved with Ward having 8 clusters for both languages. For Finnish, the 8-cluster solution also achieves the clearly best Rand index score. For Norwegian, the 8-cluster solution is on par with a 5-cluster solution on the adjusted Rand index. The 8-cluster solutions with Ward for both languages are presented in Figure 5.

For k-means, the scores differ between the two metrics: best scores are achieved with 5 (Rand) or 7 (V-measure) clusters for Norwegian, and with 4 (Rand) or 8 (V-measure) clusters for Finnish. Since the k-means scores are generally worse, they are presented in Appendix A in Figure 7.

5.4 Ward Clustering

The 8-cluster solutions for agglomerative clustering with Ward linkage are presented on a map in Figure 5 and as dendrograms in Figure 6. The colors and cluster labels are shared between the figures.

The 8-cluster solution for Finnish presented in Figure 5 manages to capture five of the eight traditional dialect areas completely. The South-Western dialects are presented in cluster number 3 (hereafter C3; presented in purple), Southern Ostrobothnia in C5 (brown), Central and Northern Ostrobothnia in C6 (pink), Savo in C0 (orange[8]) and South-East in C4 (green). The Far North is also homogeneously presented in C1 (yellow), but some speakers from the South are in the same cluster. Häme is divided, with the South-East Häme generating its own cluster (C2 / red; rest of Häme in C7 / grey). The division of Häme seemed apparent also in Figure 2 and has been reported in dialectometry before (Leino and Hyvönen, 2008). Overall, the learned representations correspond to the traditional dialects of Finnish very well, which was evident in the V-measure and ARI scores in Figure 3.

The dendrogram in Figure 6 further presents the

[8] Värmland in Western Sweden was inhabited by immigrants from Savo.

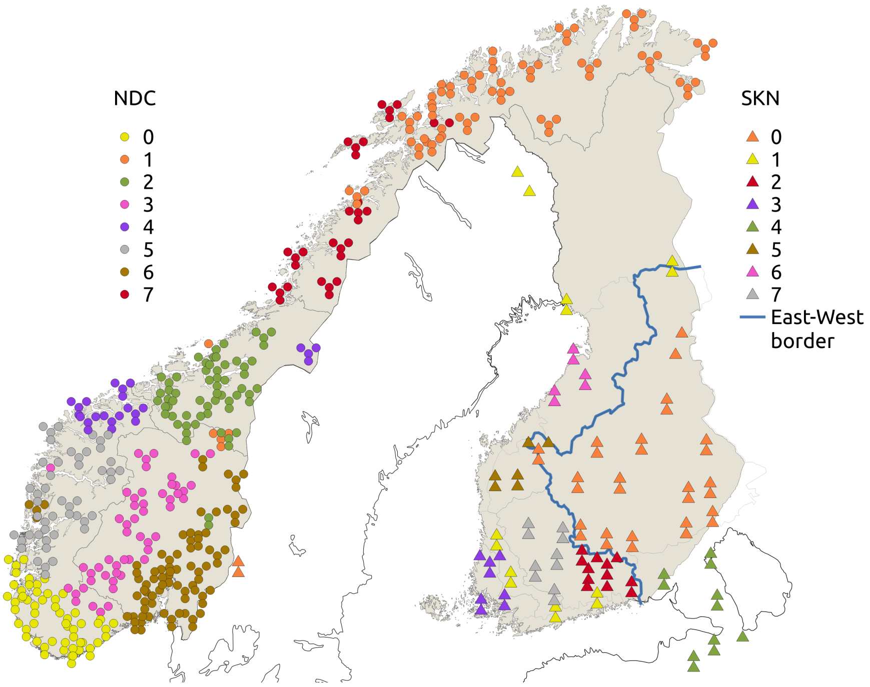

Figure 5: Agglomerative clustering (Ward linkage) based on highest V-measure. Eight clusters for both languages. Norwegian speakers are presented with circles and Finnish speakers with triangles.

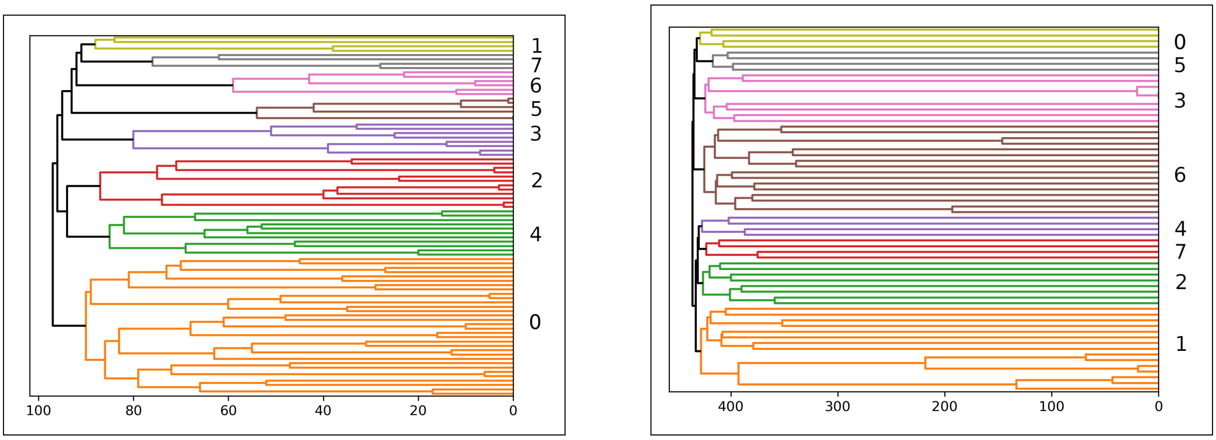

Figure 6: Dendrograms for the agglomerative clustering (Ward linkage). SKN on the left and NDC on the right. The dendrogram for NDC has been truncated for clarity. Cluster labels and colors match those of Figure 5.

relations between the clusters. The first division happens between Savo (C0 / orange) and all other dialects. Further divisions are between South-East (C4 / green and C2 / red) and the Western dialects, which in turn split up one dialect area at a time (in order: South-West (C3 / purple), Southern Ostrobothnia (C5 / brown), Central and Northern Ostrobothnia (C6 / pink) and finally Häme (C7 / grey) and the Far North (C1 / yellow)).

The clusters for Norwegian in Figure 5 are also quite distinct. The central Trøndersk area is mostly presented in cluster number 2 (hereafter C2; presented in green color), but the three other dialect areas are divided, with two clusters in Eastern (C3 / pink and C6 / brown) and Northern (C1 / orange and C7 / red) dialects, and three clusters in the Western (C0 / yellow, C5 / grey and C4 / purple) dialects. The clusters tend to stay inside the traditional dialect areas, apart from some Western speakers belonging to cluster number 3 (pink) and the municipality of Lierne (in the central Trøndersk area, near the Swedish border) belonging completely to cluster number 4 (purple).

The Norwegian Eastern dialects are moreover divided into mountain communities (*fjellbygdmål*) and lower elevation communities (*flatbygdmål*) (Hanssen, 2010 - 2014), and our clusters number 3 (pink) and 6 (brown) follow this division quite well. Likewise, the Northern dialects have a subdivision into Nordland and Troms-Finnmark, which is also reflected in clusters number 7 (red) and 1 (orange). The Western dialects have three subgroups, as do our clusters, but the areas are not as clear. The clustering is thus quite faithful to the subdivisions of the major dialect areas.

The dendrogram for NDC in Figure 6 presents the relations between the clusters. The first division is between North and South, as C2 (green), C4 (purple), C7 (red), and C1 (orange), presenting the Central and Northern dialects, are divided from the Western and Eastern dialects, presented in C5 (grey), C0 (yellow), C3 (pink) and C6 (brown). This is somewhat unexpected, as a two-way division is typically seen to be between East and West.

In the North, C1 (orange; the area of Finnmark) is divided from the three others, and C2 (green; Trøndersk) is further divided from C4 (purple) and C7 (red). In the South, C6 (brown) around Oslo (*flatbygdmål*) is first divided from the others, followed by C3 (pink; *fjellbygdmål*). This is to be expected, as both C3 and C6 clusters belong to the Eastern dialects.

All in all, it is apparent that the learned representations of the neural normalization models reflect dialect divisions. For Finnish, the clustering produced by Ward in Figure 5 is very close to the gold labels. For Norwegian, it is likely that using a more fine-grained division as gold standard could produce even higher V-measure scores, since in our clustering the four major dialect areas are divided in a way that reflects traditional understanding of dialectal subgroups.

6 Conclusions

In this paper, we apply neural dialect-to-standard normalization models to two typologically different languages and use the learned speaker representations to study the dialect continuum and division of the languages. We use large datasets of Norwegian and Finnish dialects, which have been manually transcribed and manually or semi-automatically normalized to a standard form. We add speaker labels to each dialect utterance (source) and normalize to the standard language, using byte-pair encoded data.

The model learns representations of the speakers based on the speaker labels added to the dialect utterances. The learned representations are further studied with principal component analysis, agglomerative clustering with Ward linkage, and k-means clustering. The results are evaluated against gold standard divisions of the dialects using V-measure and adjusted Rand index as metrics. Agglomerative clustering with Ward linkage outperforms k-means clustering for both languages on V-measure.

We find that the learned representations of the speakers correspond well to traditional dialect divisions. We also show that some dialect areas, such as the Häme dialect in Finnish are not as homogenic as could be assumed by the traditional division. The methodology could be further used with noisier data from social media for instance, which could reveal new insights into areal variation.

Limitations

We use clean, systematically transcribed and normalized datasets. Further evaluation of the methodology on noisier data is left for future work. We focus on two typologically different languages, but our work is still tied to the linguistic and dialectal practices of Northern Europe.

The used neural normalization model has not gone through extensive hyperparameter tuning, since the aim of the paper is not in the best possible normalization quality. It is however possible that the learned representations would perform even better if such tuning was to be executed. This also applies to the chosen dimensionality reduction methodology: using different methods might offer better results.

There are multiple ways to divide the Finnish and Norwegian dialects. We have chosen one such division for both languages, and used them as the gold standards. Using different divisions could result in different models achieving the highest scores. One could also try to avoid using gold labels altogether to find new insights into areal variation. It is anyhow apparent that the models learn dialectal differences between speakers, and that the selection of the gold standard only affects which models are deemed to perform best.

Acknowledgements

This work has been supported by the Academy of Finland through project No. 342859 "CorCoDial – Corpus-based computational dialectology".

References

Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. 2018. Multi-dialect neural machine translation and dialectometry. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. 2022. Automatic normalisation of Early Modern French. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3354–3366, Marseille, France. European Language Resources Association.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of BPE merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Online. Association for Computational Linguistics.

Kristin Hagen, Gjert Kristoffersen, Øystein A. Vangsnes, and Tor A. Åfarli, editors. 2021. *Språk i arkiva: Ny forsking om eldre talemål frå LIA-prosjektet*. Novus forlag.

Mika Hämäläinen, Khalid Alnajjar, and Tuuli Tuisk. 2022. Help from the neighbors: Estonian dialect normalization using a Finnish dialect generator. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 61–66, Hybrid. Association for Computational Linguistics.

Eskil Hanssen. 2010 - 2014. *Dialekter i Norge*, 3. opplag. edition. LNUs skriftserie ; nr. 184. Fagbokforlaget, Bergen.

Wilbert Jan Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph.D. thesis, University of Groningen.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520.

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.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

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

Terho Itkonen. 1989. *Nurmijärven murrekirja*. Kotiseudun murrekirjoja ; 10. Suomalaisen kirjallisuuden seura, Helsinki.

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. Osa. Suomalaisen kirjallisuuden seura, Helsinki.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Antti Leino and Saara Hyvönen. 2008. Comparison of component models in analysing the distribution of dialectal features. *International Journal of Humanities and Arts Computing*, 2(1-2):173–187.

Antti Leino, Saara Hyvönen, and Marko Salmenkivi. 2006. Mitä murteita suomessa onkaan? murresanaston levikin kvantitatiivista analyysiä. *Virittäjä*, 110(1):26.

Therese Leinonen, Çağrı Çöltekin, and John Nerbonne. 2016. Using gabmap. *Lingua*, 178:71–83. Linguistic Research in the CLARIN Infrastructure.

J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press.

John Nerbonne, Wilbert Heeringa, and Peter Kleiweg. 1999. *Edit Distance and Dialect Proximity*, pages 433–464. CSLI Press, Stanford.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Niko Partanen, Mika Hämäläinen, and Khalid Alnajjar. 2019. Dialect text normalization to normative standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 141–146, Hong Kong, China. Association for Computational Linguistics.

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.

Jelena Prokić and John Nerbonne. 2008. Recognizing groups among dialects. *International Journal for Humanities and Arts Computing*, 2(Special Issue on Language Variation):153–172.

QGIS Development Team. 2023. *QGIS Geographic Information System*. QGIS Association.

William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Yves Scherrer and Nikola Ljubešić. 2016. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, Bochumer Linguistische Arbeitsberichte 16, Bochum.

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.

Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre. 2018. An evaluation of neural machine translation models on historical spelling normalization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1320–1331, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

University of Turku and Institute for the Languages of Finland. 1985. The Finnish Dialect Corpus of the Syntax Archive, Downloadable Version.

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

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

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

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

A K-means Clustering

Figure 7 presents the best k-means clustering results (evaluated by the V-measure). This results in 7 clusters for Norwegian and 8 or 9 clusters for Finnish. Note that while we averaged over 5 runs when evaluating, we only present the single best run with the said number of clusters. Therefore we present the 8-cluster solution for Finnish, since it achieved a higher single run score than a 9-cluster solution.

The Finnish division achieves to capture the South-Eastern (C4 / green), Southern Ostrobothnian (C5 / brown), Northern Ostrobothnian (C7 / grey), Häme (C1 / yellow), and South-Western (C2 / red) dialects for the most part. The traditional dialect areas of South-West transitional, Far North, and Savo are however divided into several clusters. This results in a lower V-measure score than for the Ward clustering in Figure 5.

The Norwegian clusters produced by the k-means are reminiscent of the Ward clustering, presented in Figure 5. The central dialects (Trøndersk) are mostly presented in C3 (pink). The Eastern dialects are divided into mountain community (C1 / orange) and lower elevation (C5 / grey), Western dialects are divided into three groups (C2, C6, C4), and the Northern dialects into two groups (C4 / purple and C0 / yellow). There are however considerably more outliers, with some speakers belonging to different clusters than their surrounding speakers. This results in low V-measure when evaluated against the dialect areas.

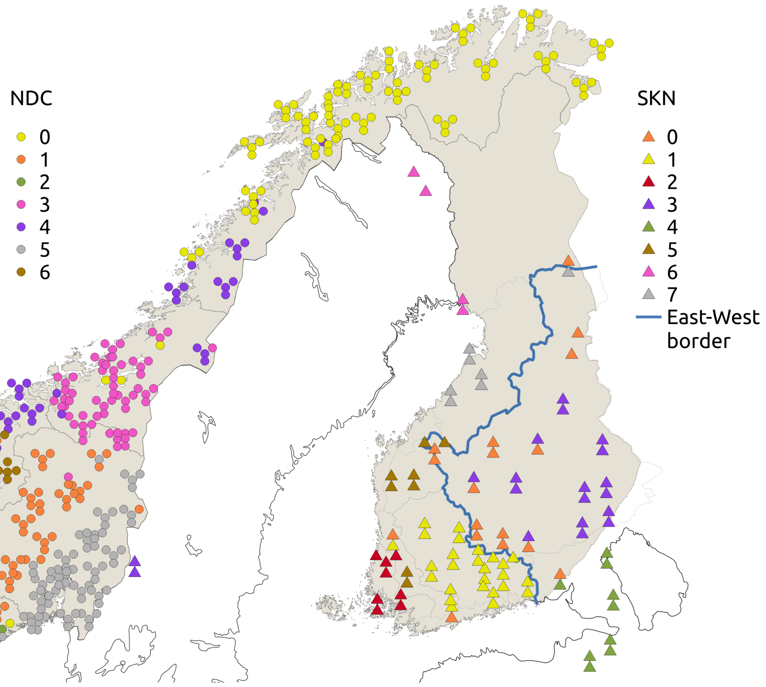

Figure 7: K-means clustering based on highest V-measure. Seven clusters for Norwegian, and eight clusters for Finnish. Norwegian speakers are presented with circles and Finnish speakers with triangles.