Journal of Linguistic Geography (2024), 1–12
doi:10.1017/jlg.2024.6

CAMBRIDGE UNIVERSITY PRESS

Article

Corpus-based dialectometry with topic models

Olli Kuparinen and Yves Scherrer
University of Helsinki

Abstract

This paper presents a topic modeling approach to corpus-based dialectometry. Topic models are most often used in text mining to find latent structure in a collection of documents. They are based on the idea that frequently co-occurring words present the same underlying topic. In this study, topic models are used on interview transcriptions containing dialectal speech directly, without any annotations or preselected features. The transcriptions are modeled on complete words, on character n-grams, and after automatical segmentation. Data from three languages, Finnish, Norwegian, and Swiss German, are scrutinized. The proposed method is capable of discovering clear dialectal differences in all three datasets, while reflecting the differences between them. The method provides a significant simplification of the dialectometric workflow, simultaneously saving time and increasing objectivity. Using the method on non-normalized data could also benefit text mining, which is the traditional field of topic modeling.

Keywords: dialectometry; topic modeling; Finnish; Norwegian; Swiss German

1. Introduction

Corpus-based approaches have become increasingly common in quantitative dialectology in recent years. Dialect corpora, typically consisting of transcribed semi-directed interviews, have thus risen as an alternative to more traditional dialect atlases. Although, in comparison to atlases, they reveal more about the context and magnitude in which linguistic features are used, they come with their own issues. One problem is that the frequencies of the collected features typically need to be normalized to be comparable enough for dialectometrical analysis (Wolk & Szmrecsanyi, 2018). In the current work, we aim to surpass the issue by using transcribed interview data directly, without explicitly defining a list of features beforehand.

We approach the problem with the use of topic models. Their aim is to find latent structure in a collection of documents, inferring that frequently co-occurring items present the same underlying component. The method is most often used in text mining, in which the collection of documents is a set of texts, the items are words in these texts, and the components are the topics discussed in the texts (hence the name). In its standard usage, the topics emerge as representing semantic fields, and in that case, the words are typically standardized and lemmatized to reduce orthographic and morphological variation. In contrast, we propose to use topic models to find dialectological components (dialects) by *not* normalizing and lemmatizing the data, thus focusing on the differing forms.

Under the hood, many topic models rely on a dimensionality reduction algorithm. Such algorithms have been used in dialectometry before. Thus, one goal of this paper is to make the connection between dimensionality reduction and topic modeling explicit and to introduce standard practices of text mining into corpus-based dialectology.

We use phonetically transcribed interview data from three languages: Finnish, Norwegian, and Swiss German. All datasets have fairly good geographical coverage and can thus be compared with traditional dialect classifications based on atlas data. Besides the three languages, we also test the method with different segmentations of the transcriptions: complete words, character n-grams, and automatically segmented words. The results are visualized on maps and analyzed to give insightful examples of each language area. The focus of the paper lies less on a rigorous analysis of the dialects than on the assessment of the method's capabilities in dialectometrical tasks. We present the data in Section 2, the methods in Section 3, and the results in Section 4. The article ends on a concluding section.

2. Data

We use three different datasets in this study, representing the dialects of three distinct languages: Finnish, Norwegian, and Swiss German. All the datasets consist of transcribed recordings but differ in the time of recording, geographical coverage, and purpose of the recordings to begin with. Using three differently built corpora from two language families tests the methodology more exhaustively than what could be achieved with just a single corpus. The main characteristics of the three corpora are summarized in Table 1, and each dataset is described in turn in the three following subsections.

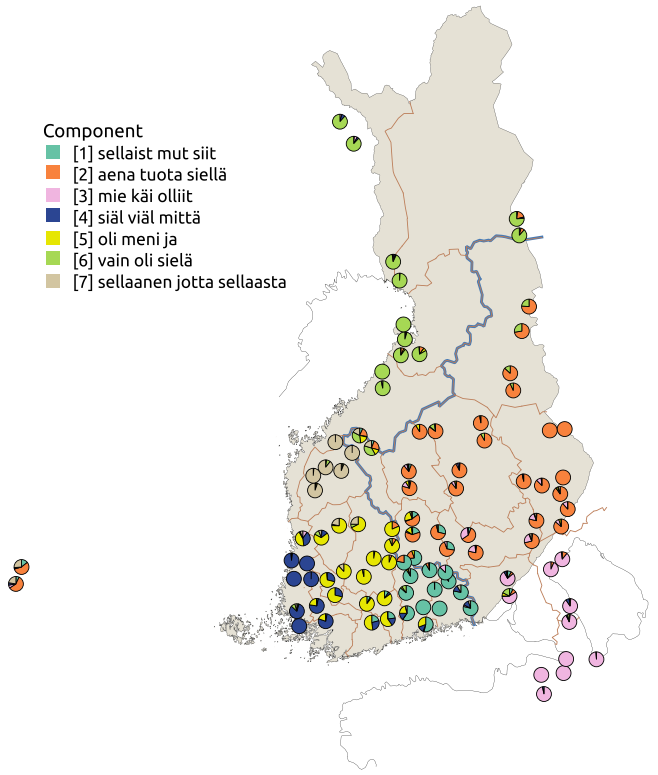
2.1. Samples of spoken Finnish

The Finnish dataset used in the study is *Samples of spoken Finnish* (Institute for the Languages in Finland, 2014). The corpus consists

Corresponding author: Olli Kuparinen; Email: olli.kuparinen@tuni.fi
Cite this article: Kuparinen O and Scherrer Y. Corpus-based dialectometry with topic models. *Journal of Linguistic Geography* https://doi.org/10.1017/jlg.2024.6

Table 1. The time and number of recordings in each corpus.

Corpus	Time of the recordings	Number of recordings
Samples of spoken Finnish	1960s–1970s	99
Norwegian Dialect Corpus	2006–2010	684 (438 informants)
Archimob Corpus (Swiss German)	1999–2001	43



Map 1. The component distribution of each speaker in a seven-component NMF model based on complete words in the Finnish dataset. The distributions are presented as pie charts, with each used component presented as a section with a specified color. The blue line denotes the border between Eastern and Western dialects, the red lines differentiate smaller dialect areas. The location in the far west is Värmland in central Sweden, where people from Savonia migrated to in the sixteenth century.

of dialect interviews recorded mostly in the 1960s. The interviews are from 50 Finnish-speaking municipalities, with two speakers born in the late nineteenth century (70 to 80 years of age) selected from each municipality. One location is limited to a single speaker, which means the corpus in total consists of 99 interviews that represent the dialect regions of Finnish comprehensively. The purpose of the corpus was to capture the rural Finnish dialects before urbanization and the standard language would influence them drastically. As a result, the speakers were hand-picked to represent their respective dialects. The interviews last for about an hour and have been phonetically transcribed and double-checked by language experts. Map 1 in Section 4.2 shows the geographic distribution of interviewed speakers across Finland.

2.2. Norwegian Dialect Corpus

The *Norwegian Dialect Corpus* consists of 684 recorded conversations in 111 locations in Norway. There are typically four informants per location: one male and female of over 50 years of age, and one male and female of under 30 years of age. The data includes recordings from both rural and urban locations. The corpus is part of the Nordic Dialect Corpus (Johannessen et al., 2009) that also includes data from other North Germanic languages. The conversations were recorded between 2006 and 2010 and have been transcribed both phonetically and orthographically to Norsk Bokmål, which is one of the standard languages of Norwegian. In this study, we only use the phonetic transcriptions.

Although the number of recordings is quite large, each informant typically appears in two recordings: once in an interview with a research assistant and once in an informal conversation with another informant. It is likely there is overlap between the interviews and conversations both content-wise and dialectologically. If we were to use them as separate documents, it would affect the modeling. We have thus combined the interview and conversation parts of a single informant into a single document, such that each document consists of the speech produced by exactly one informant. After this step, the dataset consists of 438 documents in total. The geographical coverage of the data is presented in Map 2 in Section 4.3.

2.3. Archimob Corpus

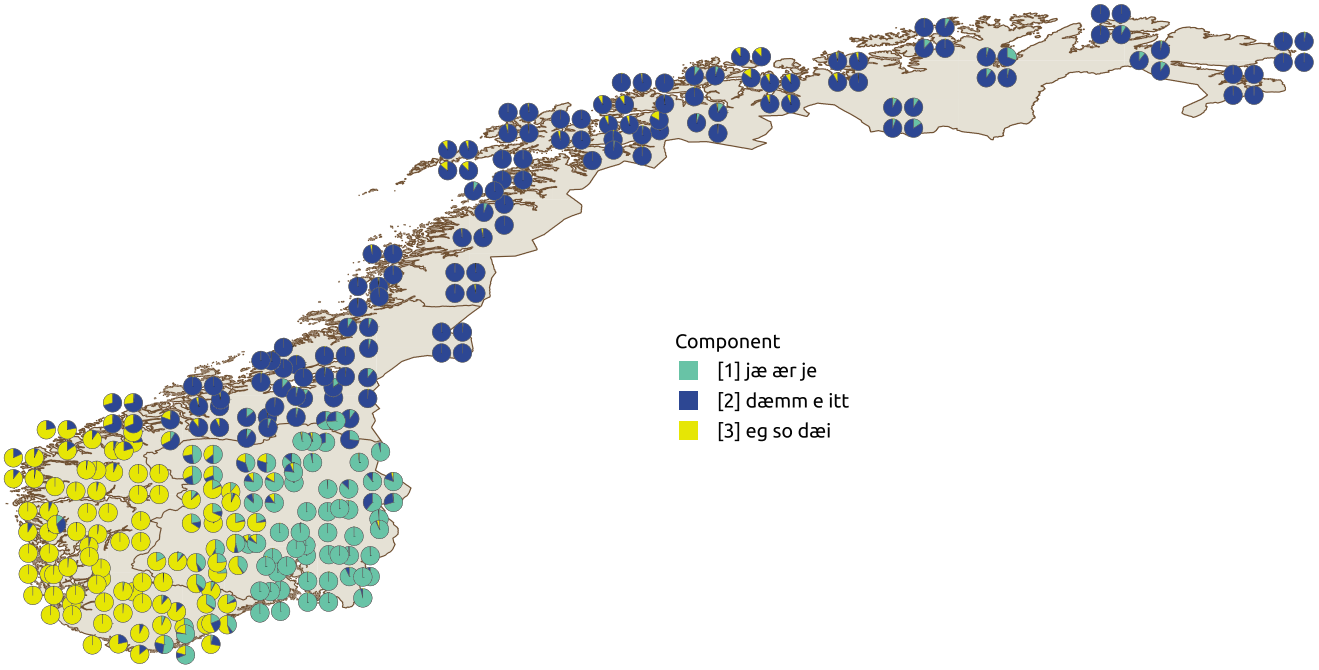
The Swiss German dataset used in the study is the Archimob Corpus. The corpus was originally collected between 1999 and 2001, when 555 interviews were recorded in the context of an oral history project focusing on the Second World War period in Switzerland. Of this large corpus, 43 interviews conducted in various Swiss German dialects have been transcribed for linguistic analysis (Scherrer, Samardžić, & Glaser, 2019). This dataset thus differs from the two other corpora in its original focus: the Finnish and Norwegian data were collected with linguistic inquiry in mind, whereas the Swiss German data was focused on historical events. The corpus is also smaller than the other two and is geographically concentrated on the cities in German-speaking Switzerland. Given that the purpose of the current work is to evaluate a new analysis method for dialect corpora, the differences between the datasets are in fact helpful: we can assess the robustness of the method when applying it to heterogeneous data. The locations of the informants in Switzerland are presented in Map 4 in Section 4.4.

3. Methods

3.1. Topic models

Topic models are statistical models that aim to discover underlying similarities in a collection of documents based on co-occurring items (Blei, 2012). Traditional topic models have been based on the concept of matrix factorization (Alghamdi & Alfalqi, 2015). A given collection of documents is represented as a data matrix $W_{D\times V}$ of D *documents* (one document per row) and V *items* (typically *words*, one per column). The values in W represent the (weighted) occurrence counts of particular words in particular documents. Topic models provide a way to decompose W into two matrices, $Z_{D\times K}$ and $H_{K\times V}$, where K is the number of *components* (or *topics*) and has to be chosen manually. Thus, Z contains the distribution of components (columns) over documents (rows), whereas H contains the distribution of items (columns) over

https://doi.org/10.1017/jlg.2024.6 Published online by Cambridge University Press



Map 2. The component distribution of each speaker in a three-component LDA model on complete words in the Norwegian dataset. The thin lines separate the four dialect areas of Eastern, Western, Trøndersk, and Northern dialects.

components (rows). Matrix factorization postulates that W can be decomposed into Z and H such that $W \cong ZH$. There are several topic modeling algorithms that differ in the exact way of building W (e.g., how to proceed with very frequent and very rare items) and in the matrix factorization methods. Besides Principal Component Analysis (PCA), Factor Analysis (FA), and Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization (NMF; Paatero & Tapper, 1994; Lee & Seung, 1999) has emerged as a popular factorization-based topic model.

More recently, Latent Dirichlet Allocation (LDA) has been proposed as an alternative for factorization-based models (Blei, Ng, & Jordan, 2003). LDA is a probabilistic approach that defines two probability distributions $p(z_k|d)$ and $p(w_v|z_k)$. The former corresponds to the probability of observing component k in document d, and the latter to the probability of observing item w_v given component k.

Topic models are traditionally used in text mining with the aim of identifying documents with similar semantic content. In this case, the co-occurring items are words in the documents, and the components are topics, hence the name of the model. For instance, the words *run*, *sneaker*, and *jog* are likely to occur together in several documents, whereas *vaccine*, *virus*, and *health* might constitute another topic in different documents. However, it is also possible that a document combines both: one could think of an information brochure describing the needed resting period after a vaccination before exercise. The main benefit of topic models, in comparison to so-called hard clustering, is that a document can combine multiple topics. For example, a document can be characterized as 70% *run–sneaker–jog*-related and 30% *vaccine–virus–health*-related.

In this study, we applied two of the most popular topic modeling approaches, namely LDA and NMF. They will be described in detail in Sections 3.3 and 3.4.

3.2. Topic models and dimensionality reduction in dialectometry

The core component of matrix-factorization-based topic models is a factorization algorithm that decomposes W into Z and H. This process is also known as dimensionality reduction since V (the vocabulary size, or the number of columns in W) is usually much higher than K (the number of components, or the number of columns in Z).

Dimensionality reduction has also been an integral part of the dialectometrical toolbox. In traditional atlas-based dialectometry, the data matrix W consists of one row per inquiry point, and one column per dialectal realization of a linguistic variable. Reducing W to Z is appealing because Z captures the most significant aspects of the overall variation in a small number of components, which can then be visualized and mapped, for example, with different colors. The most popular dimensionality reduction techniques in atlas-based dialectometric research have been factor analysis (FA) and principal component analysis (PCA) (cf., among others, Grieve, 2014; Nerbonne, 2006; Pickl, 2016; Shackleton, 2005). Leino & Hyvönen (2008) provide a comprehensive comparison of matrix factorization methods on dialect atlases.[1]

In this context, the values of Z are usually called *factor/component scores*, and H is referred to as *factor/component loadings*. It is generally observed that each component or factor corresponds to a dialect area, but some factors have also been found to indicate, for example, the distinction between urban and rural dialects (Pickl, 2016). While dimensionality reduction as such focuses on Z, the factor/component loading matrix H plays a crucial role in interpreting the results, as each component can be traced back to the dialectal variants that obtain the highest weights in it.

In atlas-based dialectometry, the data matrix is usually a binary presence–absence matrix, but some studies have derived count

https://doi.org/10.1017/jlg.2024.6 Published online by Cambridge University Press

data from atlases by aggregating linguistic variables (Nerbonne, 2015). In contrast, count data naturally occurs when using corpora.

Corpus-based dialectometry is interested in applying dialectometrical analysis techniques in general, and dimensionality reduction techniques in particular, to text corpora rather than dialect atlases. In practice, however, applications such as Szmrecsanyi & Wolk (2011) or Grieve, Speelman, & Geeraerts (2011) do not fall under the category of topic modeling because the set of linguistic features to study are established by a linguist beforehand, rather than automatically discovered. Hoppenbrouwers & Hoppenbrouwers (2001) proposed a phone frequency based method, which utilizes phone unigrams for the classification of Dutch dialects (see also Heeringa, 2004). The methodology resembles the one used here, but we were working on larger units of transcribed speech.

To our best knowledge, the first paper to propose topic models for studying linguistic variation is Eisenstein et al. (2010). They extended an LDA-based model to incorporate two sources of lexical variation, namely semantic topic and geographical region. Their study resembled traditional text mining in the sense that it focused on lexical (rather than orthographical, phonetic, or morphological) differences. The latter would anyway be hard to detect in the examined dataset of US-American tweets. Kuparinen et al. (2021) used a Latent Dirichlet Allocation model (see Section 3.3) to discover lects in Helsinki Finnish. Although the aim of the study was similar to the current one, the data was differently pre-processed. The linguistic features were collected from the data by hand and the model was run on these collected features only.

3.3. Latent Dirichlet Allocation

Probably the most popular topic modeling algorithm is latent Dirichlet Allocation, or LDA (Blei et al., 2003). LDA is a probabilistic method, meaning that it produces probabilistic distributions of what best describes the data. Each component has a distribution over all the items, and each document has a distribution over all the components. This means that, for instance, the probability of each component in each document can be analyzed. Given the probabilities always sum to 1, the distributions are also easily comparable.

LDA uses raw term frequency as its input, which means that the most frequent items in the collection of documents usually obtain the highest weights. In text mining, this problem can be overcome by devising a list of so-called stop words, a collection of the most frequent words that do not impose much semantic meaning. To use English as an example, this normally includes such words as *and*, *the*, *if*, and so on. These stop words are excluded from the modeling. When working with dialect data, such a list is problematic for two reasons. Firstly, we do not want to exclude dialectal variants of these words since we are trying to find differences between the dialects. Secondly, such a list would grow huge in size, given the different realizations of the words. We thus used LDA without a list of stop words.

In order to avoid having the most frequent terms of the corpus as highest-weighted terms, we used a post-processing metric called *relevance* (Sievert & Shirley, 2014), which gives more weight to terms that appear in only a few components. The relevance metric does not affect the modeling itself but makes the output more interpretable. The metric can be controlled by a weight parameter λ (where $0 \leq \lambda \leq 1$). If $\lambda = 1$, the top terms are ranked by their probability in the component, resulting in the very frequent terms overpowering each component. If $\lambda = 0$, the top terms are ranked by the ratio of a term's within-component probability to its probability in the whole corpus. The approach is thus similar to tf–idf presented in the next section but is done on a component level and after the modeling. We used a λ value of 0.2 in all the LDA models, thus emphasizing the terms' exclusiveness to single components.

3.4. Non-negative matrix factorization

In contrast to LDA, non-negative matrix factorization (NMF) is a non-probabilistic method. The output is nonetheless similar, as the method produces two matrices: one of the components over items, and one of the documents over components (Paatero & Tapper, 1994).

It is common to transform the term counts in the data matrix to minimize the importance of very frequent items in the modeling. We used the term frequency–inverse document frequency (tf–idf) weighting scheme as input for the NMF model. Term frequency refers to the (relative) frequency in which a term appears in a document, whereas inverse document frequency is the inverse of the number of documents the term appears in. Tf–idf is simply the product of these two counts. This procedure gives more weight to terms that appear more rarely in the corpus but might be important for a given document. Its most significant effect is to downweigh frequent terms, such as the word *the* in English, and thus obviate the need for stop word lists or additional post-processing steps.

3.5. Applying topic models to phonetically transcribed dialect corpora

When topic models are used in text mining, the words are usually normalized and lemmatized. This means that the differing forms of the same word in the corpus are simplified to one standard form, so that the differences between documents are only semantic. When looking for dialectal differences, however, we wanted to include the differences in the forms as well. If we consider, for instance, the inessive case of the word *talo* "house" in Finnish, we meet at least three different realizations: *talossa*, *talosa*, *talos* "in a house." Standardizing and lemmatizing the words would lead to all being presented as *talo*, which would reveal nothing about the dialects. Thus, we kept the transcribed words intact in order not to lose any variation. There is a caveat to this strategy: more variation also means more noise. We will elaborate on this.

Firstly, the meaning of the words still affects the modeling. If there are a lot of placenames for instance, they could end up carrying a lot of weight in the components, even though they do not represent dialectal differences. Secondly, the dialectologically meaningful features are tied to the words, which means we are not actually calculating the frequency of the variants themselves (cf. Kuparinen et al., 2021), but the combinations of words and variants. If we modify the example from before, the occurrences *talosa* "in a house," *koulusa* "in a school," and *kirkosa* "in a church" would all end up as different tokens in the corpus, although they all have the same dialectal variant *-sa* of the inessive case. To tackle this issue, we tested the method on subword units as well. We used both character n-grams and automatic segmentation of the words.

Character n-grams are sequences of characters of set length n. Character bigrams are sequences of two characters, trigrams of three, and so on. Thus, the earlier word *talosa* would end up as the following character bigrams: *_t, ta, al, lo, os, sa, a_*, in which underscore is used to present the word boundary. We see that the inessive ending *-sa* is now presented as its own bigram. We have used bigrams, trigrams, and fourgrams in the current study.[2]

https://doi.org/10.1017/jlg.2024.6 Published online by Cambridge University Press

Table 2. An example sequence of the input types from the Swiss German dataset.

Words	mini eltere händ es zwäifamiliehuus ghaa
Bigrams	_m mi in ni i_ _e el lt te er re e_ _h hä än nd d_ _e es s_ _z zw wä äi if fa am mi il li ie eh hu uu us s_ _g gh ha aa a_
Trigrams	_mi min ini ni_ _el elt lte ter ere re_ _hä hän änd nd_ _es es_ _zw zwä wäi äif ifa fam ami mil ili lie ieh ehu huu uus us_ _gh gha haa aa_
Fourgrams	_min mini ini_ _elt elte lter tere ere_ _hän händ änd_ _es_ _zwä zwäi wäif äifa ifam fami amil mili ilie lieh iehu ehuu huus uus_ _gha ghaa haa_
Morfessor	mini eltere händ e s zwäi familie huus ghaa
Gloss	"my parents had a two-family house"

We hypothesized that the character n-grams would work well for finding phonetic features in the datasets, such as *l*-vocalization in Swiss German or diphthong opening in Finnish. Trigrams have been used similarly for the study of North Frisian dialects in Birkenes (2019). The pre-processing of data to character n-grams was done on Python 3 with the Natural Language Toolkit library (Bird, Loper & Klein, 2009).

For automatic segmentation of words we used Morfessor 2.0 (Virpioja et al., 2013). Morfessor was originally designed to perform automatic morphological segmentation for morphologically rich languages (Creutz & Lagus, 2005), but the current version can be used in any string segmentation task.[3] For dialects, we would want to produce important phonetic segments as well, for instance.

Morfessor is based on the minimum description length principle and works in an entirely unsupervised, data-driven way. It is based on the assumption that every word in a text corpus can be split into one or several so-called morphs according to a given segmentation strategy. The segmentation strategy is encoded in two ways, by the segmented corpus and the morph lexicon.

The more aggressive the splitting and the shorter the morphs, the higher the amount of morphs in the segmented corpus. Thus, the optimal way to encode the segmented corpus would be to leave each word intact and not split anything. On the other hand, the shorter the morphs, the fewer items in the morph lexicon. At the extreme, splitting all words into single characters would lead to the shortest morph lexicon. The goal of Morfessor is to jointly optimize the segmented corpus size and the morph lexicon size, leading to a segmentation strategy that lies somewhere between the two extremes.

When given a training corpus, Morfessor iteratively finds a segmentation strategy that most accurately describes the corpus. The segmentation model created from the training corpus can then be used to segment new documents. For our study, we have used the complete datasets as training corpora, and each document was thereafter segmented with the model built on the respective dataset.

To provide a clear picture of the different input segmentation strategies, an example sequence from the Swiss German dataset is provided in Table 2.

3.6. Experiments

We ran experiments with both topic modeling methods (LDA and NMF) on Python 3 using the library *scikit-learn* (Pedregosa et al., 2011), with a number of components ranging from 3 to 10. Items appearing in only one document were excluded from the modeling, but otherwise there were no limits on input.[4]

With five possible input types (words, bigrams, trigrams, fourgrams, and Morfessor-segmented items), two methods, and eight numbers of components, we ended up with 80 models for each of the three datasets. In the next section, we explain how we evaluate these models and present results of selected models.

4. Results

We will begin our exploration of the results by describing our method of evaluating the models in Section 4.1. We will then present the best evaluated models for each dataset in Sections 4.2, 4.3, and 4.4. The results will be presented both as bar charts of the most important features of each component in the models, as well as geographical visualizations of the components in the interview locations. The bar charts were made in Python 3 with the matplotlib library (Hunter, 2007), and the maps were made with QGIS.

4.1. Model evaluation

In order to determine the best model parameters for a given dataset, we made use of two evaluation criteria. Ideally, a model should:

1. Assign the same dominant component, i.e., the component with the highest value in each document, to all texts that originate from the same dialectal area (as defined by earlier, independent dialectological studies).
2. Infer divergent but coherent components, i.e., the component's distributions should be different from each other without any outlier components.

We measured (1) using a completeness score (Rosenberg & Hirschberg, 2007). Completeness score measures how well a predefined class present in the data stays intact in the modeling. The completeness score is maximal (value of 1) when all members of each given class are assigned to the same inferred component. It is minimal (value of 0) when there are members of all classes in all components.

We wanted to maximize the completeness score on dialect areas presented by independent studies in the past. For Finnish, the dialect areas were based on the division presented in Itkonen (1989), with eight areas in total. The division was mostly based on phonological and morphological features. For Norwegian, the dialect areas were Western, Eastern, Northern, and Trøndersk dialects, as presented in Hanssen (2014). The division was based on phonological, morphological, and lexical differences. For the Swiss data, the dialectal areas were based on the ten-cluster solution of the atlas-based dialectometrical study by Scherrer & Stoeckle (2016).

LDA occasionally fails when facing large vocabularies (Dieng et al., 2020), which is the case in our study, especially when using complete words or Morfessor-segmented units as input. This results in either almost identical components or components containing very rare items. To ensure that the obtained components were both different from each other but also do not contain clear outliers, we used evaluation measure (2). We calculated pairwise cosine similarities between the terms' distributions over the components for each model. A difference between the maximum and minimum similarity was used as a filter

https://doi.org/10.1017/jlg.2024.6 Published online by Cambridge University Press

Table 3. The highest ranking models on dialectal completeness in the Finnish data.

Input type	Method	Number of components	Dialectal completeness	Cosine similarity difference
Words	NMF	7	0.84	0.31
Morfessor	NMF	7	0.84	0.27
Words	NMF	3	0.84	0.11
Morfessor	NMF	6	0.83	0.24
Words	NMF	6	0.83	0.30

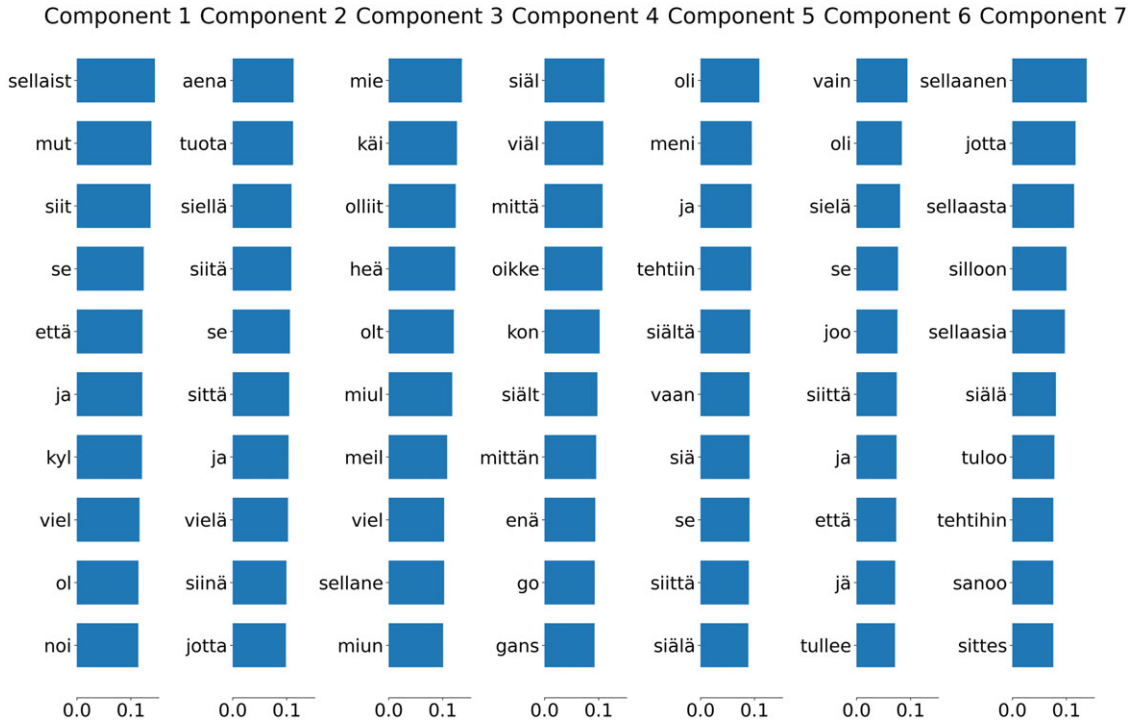


Figure 1. The highest-ranking terms in each component of a seven-component NMF model based on complete words in the Finnish dataset.

to exclude models that included either too similar components (producing maximum similarity close to 1) or clear outliers (minimum similarity close to 0).

The models that accumulated a top 5 score on the completeness metric and a difference of maximum and minimum cosine similarities of under 0.5 are presented for each dataset.

4.2. Samples of Spoken Finnish

The models that achieved the highest completeness scores for the Finnish data are presented in Table 3. The scores are high, but also indicate that the models diverge slightly from the traditional dialect areas. All top-performing models were based on the NMF method. Regarding input type, the longer sequences seem to work better for Finnish: there are no models based on character n-grams.

We further present the best model: the seven-component NMF model on complete words. The top ten terms for each component are presented in Figure 1, and each speaker's distribution of components is presented in Map 1. Features in components 3, 4, and 7 in Figure 1 correspond very clearly to traditional dialects (South-East, South-West, and Southern Osthrobothnia, respectively), and components 2 and 5 also exhibit some salient features of Savonian (C2) and Tavastian (C5) dialects. Components 1 and 6 are opaquer.

The first observation from Map 1 is that there were clear clusters of components in the traditional heartlands of the dialect areas, and more diversity in transitional areas. This is perhaps most evident when transitioning from Component 4 to 5 in the South-West. This area is traditionally known as the South-Western transitional dialects, that exhibit features from the South-West (Component 4) and Tavastia (Component 5). Moreover, the speakers from Lappajärvi near the western coast but on the eastern side of the blue line produced several different components. This is another traditional transition zone from the Savonian dialects to Ostrobothnian dialects. Thus, the modeling produced very expected results.

There were, however, some differences between the modeling and the traditional division into eight dialect areas. Component 6 would traditionally be divided into two dialects, with the northern locations separated from the others. Perhaps the most surprising difference between the modeled result and the division presented in Itkonen (1989) is the independence of Component 1. In the

https://doi.org/10.1017/jlg.2024.6 Published online by Cambridge University Press

Table 4. The highest ranking models on dialectal completeness in the Norwegian data.

Input type	Method	Number of components	Dialectal completeness	Cosine similarity difference
Words	LDA	3	0.79	0.07
Trigrams	NMF	3	0.75	0.03
Fourgrams	LDA	3	0.73	0.06
Bigrams	LDA	3	0.73	0.06
Fourgrams	NMF	3	0.72	0.09

traditional division this area would use Tavastian dialects (C5) but have some features from the southeast (C3). It seems the model is picking up on the unique combination of features, thus resulting in a completely separate component.

Based on the completeness scores, there seem to be quite distinct linguistic groups in the Finnish data. This was expected, as the dataset was collected to clearly portray the Finnish dialects. The most dialectal speakers were chosen from each municipality, and the transcriptions were very pedantic, which resulted in clear-cut differences.

4.3. Norwegian Dialect Corpus

As opposed to the Finnish dataset which focused on older speakers from the 1960s, the Norwegian dataset is more diverse. It includes speakers from urban and rural backgrounds as well as from different ages, and it was collected in the 2000s. We thus expected more variation in the results as well. The models that achieved the highest completeness scores for the Norwegian data are presented in Table 4.

In all models presented in Table 4, the number of components is three. This suggests that the division to three is prominent in the data, even though there is some diversity in terms of the method and input type. Four out of five models were based on character n-grams, whereas for Finnish there were no n-gram models. For Swiss German, the best models were also not based on character n-grams (see Section 4.4). We will therefore present two models for Norwegian to give the readers a possibility to see an n-gram based model as well. We will scrutinize the best model (LDA on complete words) and the second best model (NMF on trigrams).

We will commence with the three-component LDA model on complete words. The top terms of each component are presented in Figure 2 and the map representation in Map 2. The top terms included very frequent variant words such as *jæ, e, eg* "I" and *itte, itt, tje, ittje* "not."[5]

There were very clear clusters of used components in Map 2, with some overlap when transitioning from the east to the west. The Western (C3) and Eastern (C1) dialects followed the assumed dialect areas (divided by the thin lines) quite well, but the Northern and Trøndersk dialects were merged in C2. This combination is quite surprising, given that the Northern dialects are traditionally combined with the Western dialects and the Trøndersk dialects with the Eastern dialects, when the dialects are divided in two.

Another interesting finding is that the speakers of Kristiansand (the sixth largest city in Norway, located on the south coast) diverge from their surroundings: all speakers use dominantly C1 instead of C3. The city is the de facto center of the area of Agder and might thus differ from the areas surrounding it. It is encouraging that the method is capable of discovering the differences between urban and rural areas.

The three-component NMF model on trigrams is presented in Figure 3 and on Map 3. In Figure 3, we saw some complete words such as *eg* "I," *då* "then," and *ja* "yes" along with phonetic features such as retroflex flap [ɽ] denoted by a capital L in the corpus. The inclusion of complete words was due to word size: there were short, very frequent words in the Norwegian dialects.

Map 3 included much more diversity than Map 2, especially in the center of the country (Trøndersk). This is mostly due to the retroflex flap, denoted by a capital L in C1, which is used both in the Eastern and Trøndersk dialects. In Map 3, the traditional two-way division also becomes more apparent. There are traces of Component 2 (Western dialects) in the north and vice versa. Moreover, there is Eastern influence (Component 1) in the Central Trøndersk area and vice versa.

As opposed to the very clear-cut differences in the Finnish dataset, there was more diversity in the Norwegian data. Given that the dataset is much more modern and includes also younger and more urban speakers, this is no surprise. However, there were still very clear differences between the dialects in the most frequent words and some salient phonetic features. The method was also able to discover the diverging speech of the city of Kristiansand.

4.4. Archimob Corpus

Our third and final dataset consisted of 43 interviews in Swiss German. It was not originally collected with linguistic inquiry in mind and was thus different from the Finnish and Norwegian datasets (see Section 2.3). The top models for the dataset are presented in Table 5. Morfessor-based models worked best for the Swiss German dataset, with four out of the five best models being based on Morfessor-segmented units. All the models presented in Table 5 were NMF-based. We will analyze the best Morfessor-based model (seven components). The highest-ranking terms are presented in Figure 4 and the component distribution of each speaker in Map 4.

The highest-ranking terms in Figure 4 demonstrated how the Morfessor segmentation works: most of the top terms were complete words, but there were also smaller segments included such as *e*, *i*, *äu*, and *öü*. Regarding the components themselves, there were interesting differences: C1 used the diphthong *äi*, C2 diphthong *ai*, and C3 diphthong *ei*. This was most evident in the word "said," which takes the forms *gsäit*, *gsait*, and *gseit*. Other phonetic features included l-vocalization in C3 and C6 (*viu*, *mau*, *vöu*, *mou*, and *wöu*, Standard German *viel*, *mal*, *weil*, respectively), the retainment of word initial *kh* (instead of *ch*) in C5, and the palatalization and unrounding of vowels in C7 (*chenne* vs. *chönne*, *miesse* vs. *müesse*, *hit* vs. *hüt*, *üf* vs. *uf*).

In Map 4 we saw much more variation than in the Finnish and Norwegian datasets, but also some well-defined clusters in C3 around Bern, C5 around Basel (the name of the city also appears as a top term in Figure 4), and C7 in the Alpine region. The area in and around Zurich seems to have had the most variation, with components 1 and 2 appearing the most. These two components diverged in the use of the diphthong ending in -i, with C1 using *äi* and C2 *ai*.

Upon thorough examination of these two components, it can be seen that they formed a complementary pattern: if one is used, the other is not. The distribution of both along with the transcriber information is shown in Map 5. From the map, it became apparent that this division was indeed based on transcriber differences: the

https://doi.org/10.1017/jlg.2024.6 Published online by Cambridge University Press

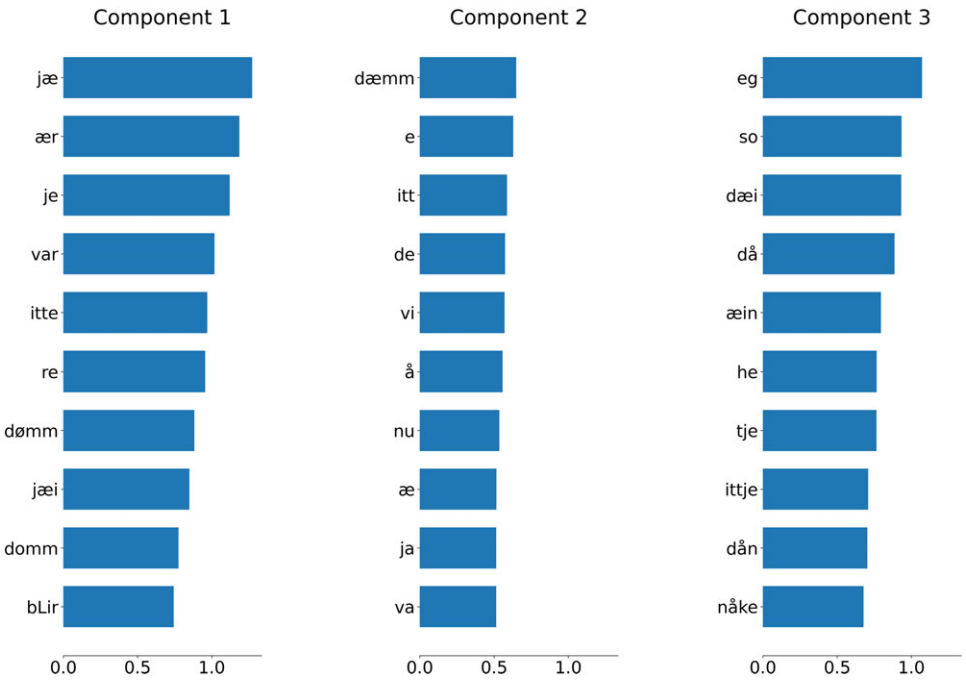


Figure 2. The highest-ranking terms in each component of a three-component LDA model on complete words in the Norwegian dataset.

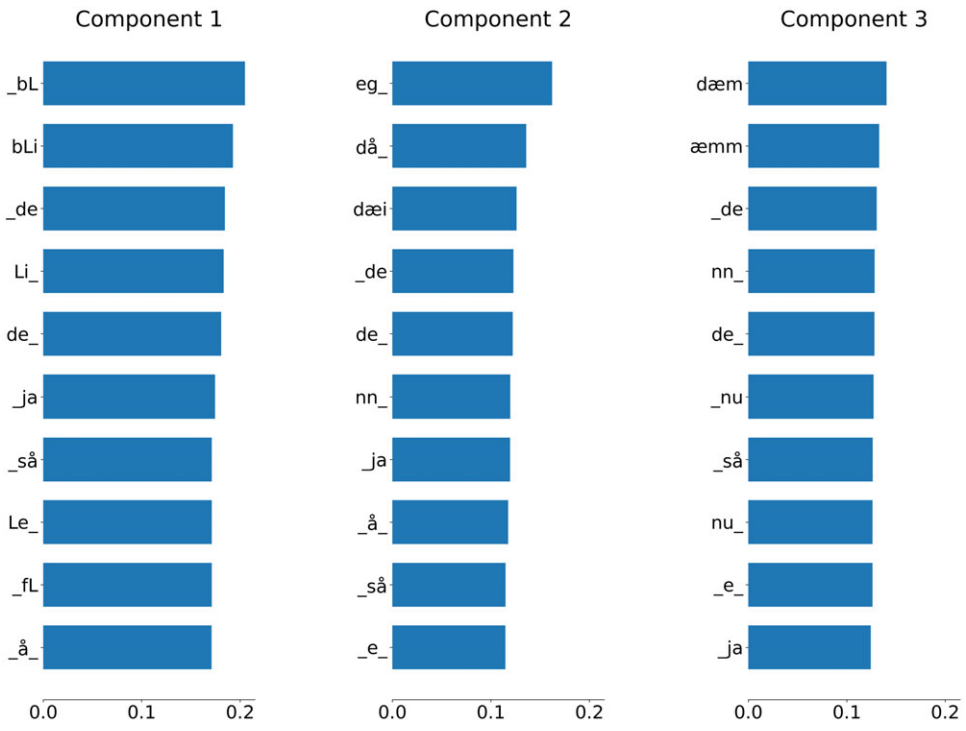


Figure 3. The highest-ranking terms in each component of a three-component NMF model on trigrams in the Norwegian dataset.

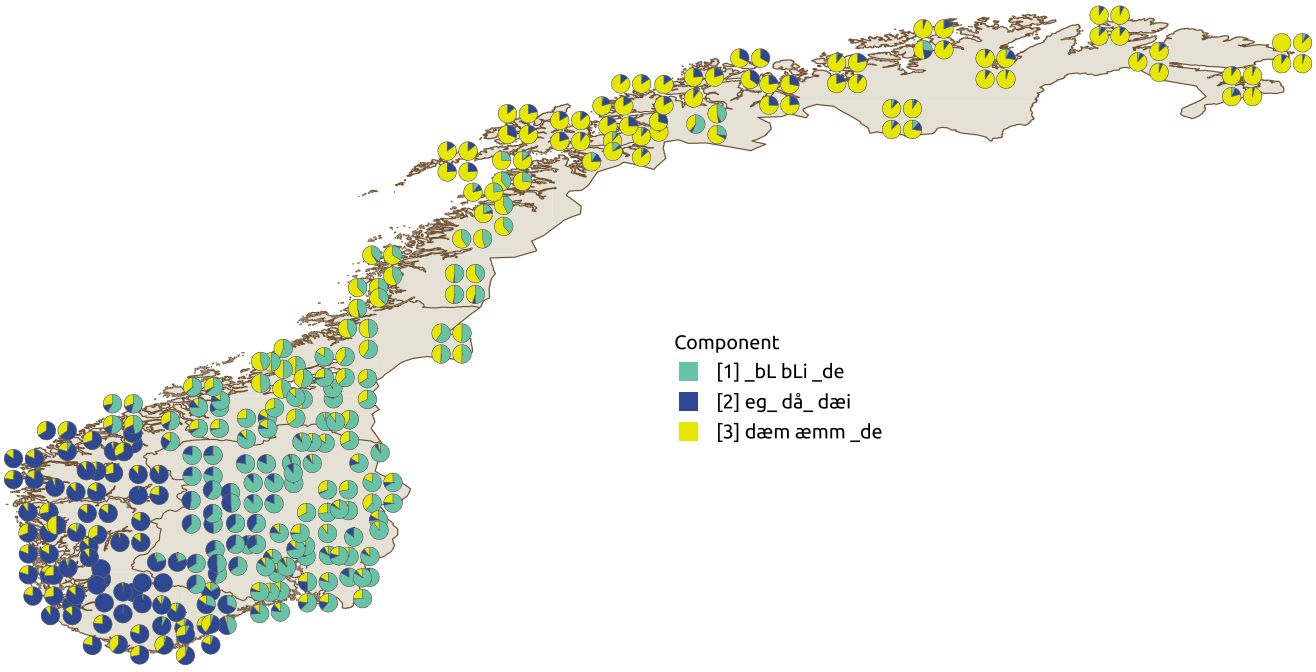
interviews transcribed by B and C have high values of C2, whereas the interviews transcribed by A (and to some extent E) have high values of C1.[6] Although this sort of pattern was not desired in this study, it showed the possibilities of the method in the study of orthographical variation. The method could be used in studies of historical texts or social media, for instance.

The Swiss German dataset proved to be more diverse than the Norwegian and Finnish datasets, but clear clusters still appeared around Bern and Basel, and in the Alpine region. The variation pattern around Zurich turned out to indicate transcriber differences. The Swiss dataset was thus able to simultaneously show that one is very dependent on transcription quality when

https://doi.org/10.1017/jlg.2024.6 Published online by Cambridge University Press

Table 5. The highest ranking models on dialectal completeness in the Swiss German data.

Input type	Method	Number of components	Dialectal completeness	Cosine similarity difference
Morfessor	NMF	7	0.64	0.24
Morfessor	NMF	4	0.62	0.16
Morfessor	NMF	8	0.62	0.27
Morfessor	NMF	6	0.61	0.21
Fourgrams	NMF	7	0.61	0.26



Map 3. The component distribution of each speaker in a three-component NMF model on trigrams in the Norwegian dataset.

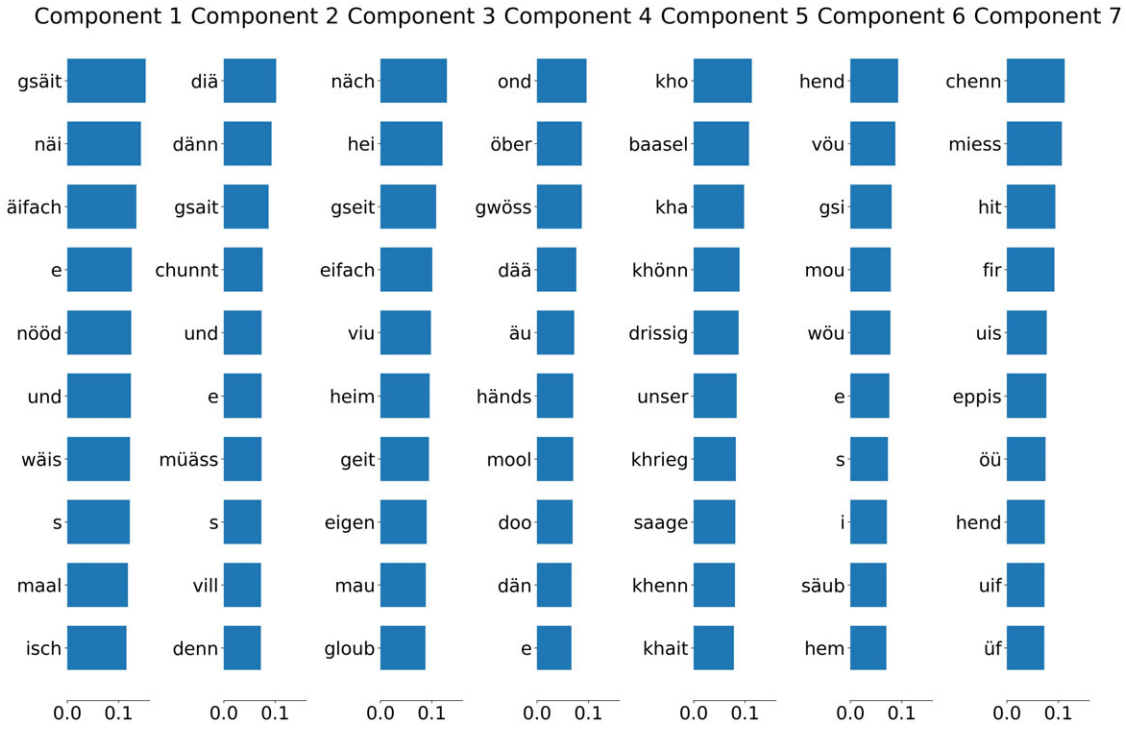
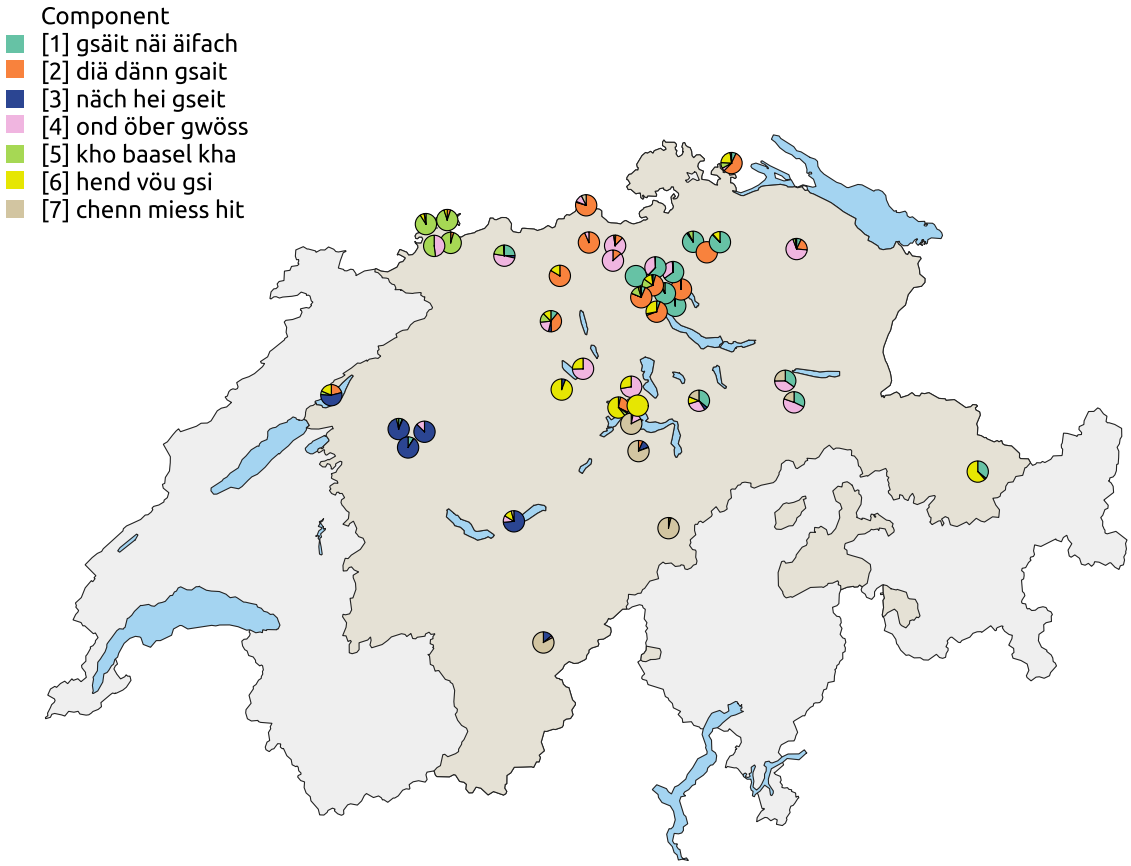


Figure 4. The highest-ranking terms in each component of a seven-component NMF model on Morfessor-segmented units in the Swiss German dataset.

https://doi.org/10.1017/jlg.2024.6 Published online by Cambridge University Press



Map 4. The component distribution of each speaker in a seven-component NMF model on Morfessor-segmented data in the Swiss German dataset. The German-speaking area is presented in beige and the big lakes in blue.

using the method on dialects and that the method could be very helpful in discovering orthographic variation in written language.

5. Conclusions

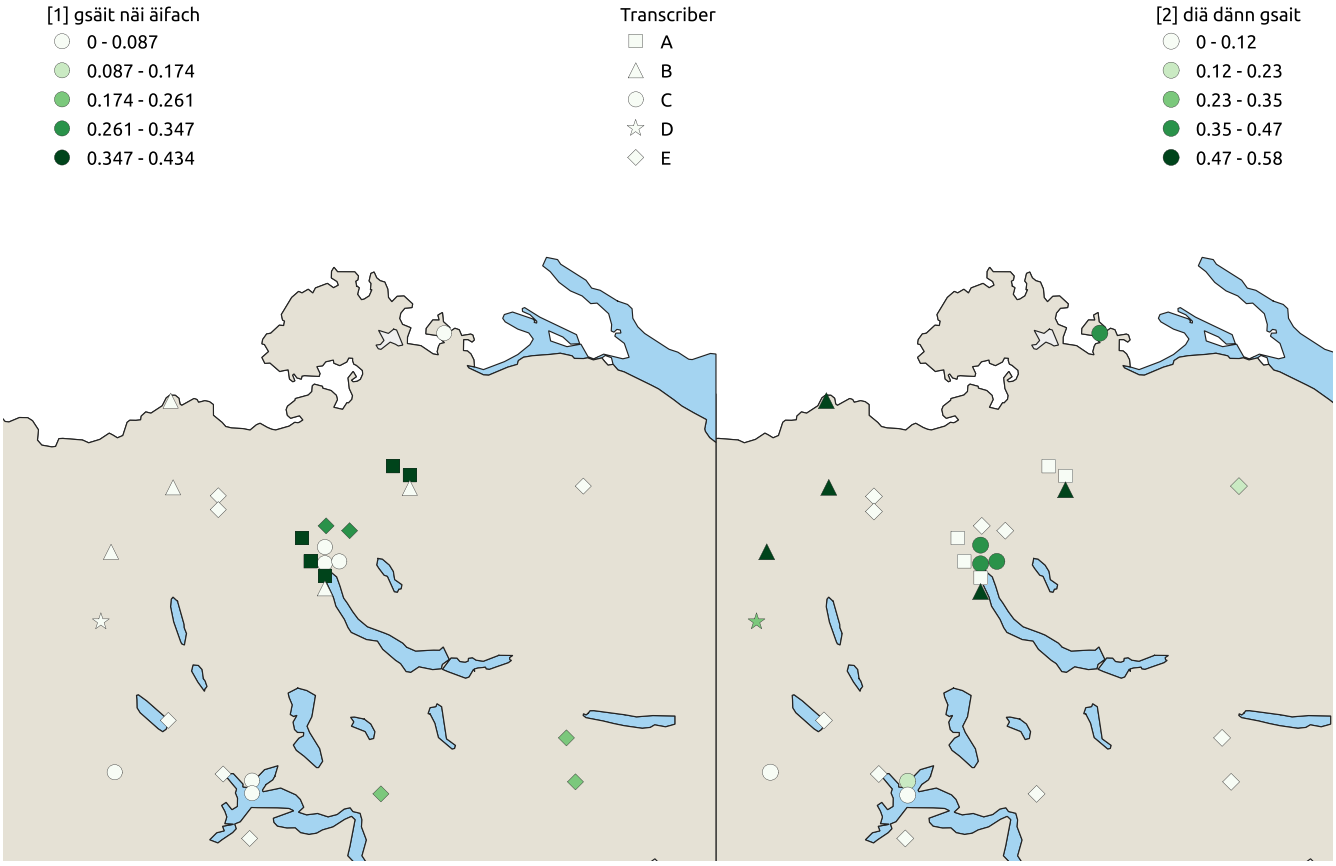
In this study we apply two topic modeling methods to three dialectal datasets, with the number of components ranging from three to ten and with five different input types. In conclusion, the method presented works admirably: regardless of input type, the best models result in meaningful clusters of component usage. Regarding the two possible methods, NMF appears more stable and is also substantially faster to run than LDA. In the dialectological setting, LDA also needs post-processing of the top terms, as a list of stop words is not desirable. As such, NMF would be the easier solution of the two. However, it is also likely that the results provided here could just as well be achieved by other dimensionality reduction algorithms, such as factor analysis or principal component analysis (cf. Leino & Hyvönen, 2008). The benefit of topic models over these methods is the easy interpretation of the results, as the models work only on non-negative numbers and offer the highest weighted words or subwords as the model output. These top terms essentially represent the most important features in each dialect.

The method makes evident the differences between the datasets. The Finnish dataset was collected in the 1960s, and the speakers were handpicked to present the ideal dialects. This is mirrored in our results, as there is very little variation in the modeling result. The Norwegian dataset, collected in the 2000s, is more diverse, with a bigger city (Kristiansand) diverging from the surrounding areas. All in all, there is still remarkably little overlap between the components, and the correspondence with the dialect areas presented in Hanssen (2014), for instance, is relatively high. Finally, the Swiss German dataset is the most diverse. This is mostly due to two factors: the dataset was not collected for linguistic purposes and there seems to be variation in transcription styles. The finding suggests extra care must be taken when modeling dialects based on transcriptions.

We experimented with five different input types: complete words, bigrams, trigrams, fourgrams, and Morfessor-segmented data. No type turns out to be better than the rest for all datasets: words and Morfessor-segmented units worked best for Finnish, words and character n-grams for Norwegian, and Morfessor-units for Swiss German. The most significant takeaway of the study is that the method can discover dialect areas and the features of these dialects based on the transcriptions alone, without any annotations or preselected features. This provides a significant simplification of the corpus-based dialectometric workflow.

The current study has focused on transcriptions of spoken dialect corpora. The logical next step would be to extend the method to variation in written language. Either historical texts with significant orthographic variation or modern social media data could serve as possible corpora. There has been plenty of research on social media with topic modeling (e.g., Eisenstein et al., 2010; Steinskog, Therkelsen, & Gambäck, 2017; Dahal, Kumar, & Li, 2019), but the focus has so far been on semantic meaning. Scrutinizing structural variation could thus result in new and

https://doi.org/10.1017/jlg.2024.6 Published online by Cambridge University Press



Map 5. The distribution of components 1 and 2 around Zurich compared with the transcribers of each interview.

insightful findings. Historical texts on the other hand could provide a long timescale and thus a possibility to observe changes in written language with a topic modeling approach.

All in all, the method can discover interesting variation on several levels: morphology, phonology, and orthography. Extending the input to word n-grams could possibly even leverage syntactic studies.

Topic modeling in text mining has often been used on normalized and lemmatized data. The current study shows that this is not always necessary or even desirable and that this traditional domain of topic modeling could benefit from different inputs as well. Especially for lower-resourced languages for which normalization or lemmatization models do not exist, the method itself could still provide useful insights.

Acknowledgments. This work has been supported by the Academy of Finland through project No. 342859 "CorCoDial—Corpus-based computational dialectology." Both authors worked for the University of Helsinki during the study, but have since changed affiliations. Olli Kuparinen is now affiliated with Tampere University and Yves Scherrer with the University of Oslo.

Competing interests. The authors declare none.

Notes

1. A related approach is known under the name of Multidimensional Scaling (MDS). In this case, the data matrix is converted to a square distance matrix containing real values between 0 and 1, and a dimensionality reduction algorithm is applied on the distance matrix. MDS is mathematically closely related to PCA, with the main difference being that MDS is applied to distance matrices and PCA to data matrices.

2. We also experimented with a combination of bi- and trigrams, but the results did not significantly differ from simple bi- or trigrams.

3. Another popular subword segmentation method is the non-probabilistic byte pair encoding (BPE; Sennrich, Haddow & Birch, 2016), which in our topic modeling experiments performed slightly worse than Morfessor.

4. The scripts for the pre-processing and topic modeling are available on Github: https://github.com/Helsinki-NLP/dialect-topic-model.

5. Some models also capture variants with -k (*ikke/ikkje*) but they did not appear in the top models.

6. Note that the Zurich area has been traditionally found to be a transition zone between *äi* and *ai* diphthongs, e.g. in the *Linguistic atlas of German-speaking Switzerland* (Hotzenköcherle et al., 1962-1997): http://dialektkarten.ch/mapviewer/swg/index.de.html#point:1109_Geiss. It is therefore not surprising that transcriber variation is observed for this particular phenomenon in this particular geographic area.

References

Alghamdi, Rubayyi & Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications* 6(1). doi: 10.14569/IJACSA.2015.060121.

Bird, Steven, Edward Loper, & Ewan Klein. 2009. *Natural language processing with Python*. O'Reilly Media Inc.

Birkenes, Magnus Breder. 2019. North Frisian dialects: A quantitative investigation using a parallel corpus of translations. *US WURK* 68(3–4). 119–168. doi: 10.21827/5c98880d173a4.

Blei, David. 2012. Probabilistic topic models. *Communications of the ACM* 55(4). 77–84.

Blei, David, Andrew Ng, & Michael Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*. 993–1022.

Creutz, Mathias & Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology.

Dahal, Biraj, Sathish Alampalayam Kumar, & Zhenlong Li. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining 9*. 1–20.

Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics 8*. 439–453.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, & Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277–1287. Cambridge, MA: Association for Computational Linguistics.

Grieve, Jack. 2014. A comparison of statistical methods for the aggregation of regional linguistic variation. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech* (Lingua & Litterae 28), 53–88. Berlin & New York: Walter de Gruyter.

Grieve, Jack, Dirk Speelman, & Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change 23*(2). 193–221.

Hanssen, Eskil. 2014. *Dialekter i Norge*, 3rd edn. Bergen: Fagbokforlaget.

Heeringa, Wilbert Jan. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. dissertation. Groningen: University of Groningen.

Hoppenbrouwers, Cor & Geer Hoppenbrouwers. 2001. *De indeling van de Nederlandse streektalen. Dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Assen: Koninklijke Van Gorcum B.V.

Hotzenköcherle, Rudolf, Robert Schläpfer, Rudolf Trüb, & Paul Zinsli (eds.). 1962–1997. *Sprachatlas der deutschen Schweiz*. Bern: Francke.

Hunter, John. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering 9*(3). 90–95. doi: 10.1109/MCSE.2007.55.

Institute for the Languages in Finland. 2014. *Samples of spoken Finnish, downloadable corpus*. The Language Bank of Finland. Retrieved from http://urn.fi/urn:nbn:fi:lb-2020112937.

Itkonen, Terho. 1989. *Nurmijärven murrekirja* (Kotiseudun murrekirjoja 10). Helsinki: Suomalaisen Kirjallisuuden Seura.

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, & Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus—an advanced research tool. In Kristiina Jokinen and Eckhard Bick (eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA*, 73–80. Odense: Northern European Association for Language Technology (NEALT). Corpus available at: http://tekstlab.uio.no/nota/scandiasyn/.

Kuparinen, Olli, Jaakko Peltonen, Liisa Mustanoja, Unni Leino, & Jenni Santaharju. 2021. Lects in Helsinki Finnish—a probabilistic component modeling approach. *Language Variation and Change 33*(1). 1–26. doi: 10.1017/S0954394521000041.

Lee, David & H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature 401*. 788–791. doi: 10.1038/44565.

Leino, Antti & 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. doi: 10.3366/E1753854809000378.

Nerbonne, John. 2006. Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing 21*(4). 463–475.

Nerbonne, John. 2015. Various variation aggregates in the LAMSAS South. In Catherine Evans Davies & Michael D. Picone (eds.) *New Perspectives on Language Variety in the South: Historical and Contemporary Approaches*, 369–382. Tuscaloosa, AL: The University of Alabama Press.

Paatero, Pentti & Unto Tapper. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics 5*. 111–126. doi: 10.1002/env.3170050203.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, & Bertrand Thirion. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*. 2825–2830.

Pickl, Simon. 2016. Fuzzy dialect areas and prototype theory: Discovering latent patterns in geolinguistic variation. In Marie-Hélène Côté, Remco Knooihuizen, & John Nerbonne (eds.), *The future of dialects*, 75–98. Berlin: Language Science Press. doi: 10.17169/langsci.b81.84.

Rosenberg, Andrew & 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*, 410–420. Prague: Association for Computational Linguistics.

Scherrer, Yves, Tanja Samardžić, & Elvira Glaser. 2019. Digitising Swiss German: How to process and study a polycentric spoken language. *Language Resources & Evaluation 53*. 735–769. doi: 10.1007/s10579-019-09457-5.

Scherrer, Yves & Philipp Stoeckle. 2016. A quantitative approach to Swiss German—dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica 24*(1), 92–125. doi: 10.1515/dialect-2016-0006.

Sennrich, Rico, Barry Haddow, & 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*, 1715–1725. Berlin: Association for Computational Linguistics.

Shackleton, Robert. 2005. English–American speech relationships. *Journal of English Linguistics 33*(2). 99–160.

Sievert, Carson and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. Baltimore, MD: Association for Computational Linguistics.

Steinskog, Asbjørn, Jonas Therkelsen, & Björn Gambäck. 2017. Twitter topic modeling by Tweet aggregation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, 77–86. Gothenburg: Association for Computational Linguistics.

Szmrecsanyi, Benedikt & Christoph Wolk. 2011. Holistic corpus-based dialectology. *Revista Brasileira de Linguística Aplicada 11*(2). 561–592.

Virpioja, Sami, Peter Smit, Stig-Arne Grönroos, & Mikko Kurimo. 2013. *Morfessor 2.0: Python implementation and extensions for Morfessor baseline* (Aalto University publication series Science + Technology 25). Retrieved from http://urn.fi/URN:ISBN:978-952-60-5501-5.

Wolk, C. & B. Szmrecsanyi. 2018. Probabilistic corpus-based dialectometry. *Journal of Linguistic Geography 6*(1). 56–75. doi: 10.1017/jlg.2018.6.

https://doi.org/10.1017/jlg.2024.6 Published online by Cambridge University Press