Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5343–5350
Marseille, 11–16 May 2020

Comparing methods for measuring dialect similarity in Norwegian

Andre Kåsen*, Janne Bondi Johannessen, Kristin Hagen**, Anders Nøklestad**, Joel Priestley****

* National Library of Norway, ** Department of Linguistics and Scandinavian Studies, University of Oslo
* Pb 2674 Solli 0203 Oslo, Norway ** Pb. 1102 Blindern, 0317 Oslo, Norway
andre.kasen@nb.no, {j.b.johannessen, kristin.hagen, anders.noklestad, joel.priestley}@iln.uio.no

Abstract

The present article presents four experiments with two different methods for measuring dialect similarity in Norwegian: the Levenshtein method and the neural long short term memory (LSTM) autoencoder network, a machine learning algorithm. The visual output in the form of dialect maps is then compared with canonical maps found in the dialect literature. All of this enables us to say that one does not need fine-grained transcriptions of speech to replicate classical classification patterns.

Keywords: dialectometry, Norwegian, transcription, geolinguistics, maps, the Levenshtein method, neural networks, long short term memory (LSTM) autoencoder network, machine learning algorithm

1. Introduction

This paper explores whether it is possible to use existing dialect transcriptions from a recently finished corpus project to automatically generate dialect areas. The phonetic transcriptions are quite coarse-grained, which indicates that they might be possible to generalise over by an automatic method. We have chosen to test two different methods: The Levenshtein method, which uses edit distance to compute distances between dialects, and the neural long short term memory (LSTM) autoencoder network, a machine learning algorithm. The resulting maps show that the transcriptions can indeed be used for the purpose of automatically generating dialect maps, but while the neural network method needs a large dataset, the Levenshtein method can get very good results with a small dataset, too.

The paper is structured a follows: Section 2 describes the LIA project from which the transcriptions have been taken, and also describes the actual transcriptions and traditional dialect maps, Section 3 describes the datasets we use, while Sections 4 and 5 present the results of using the two datasets on the two methods. Section 6 concludes the paper, while references are given in Sections 7 and 8.

2. LIA Norwegian - Corpus of Old Dialect Recordings, Transcriptions and Old Maps

2.1 The Project and the Recordings

LIA (Language Infrastructure made Accessible) is a national Norwegian project involving the four biggest universities in the country (the University of Oslo – UiO, the University of Bergen – UiB, UiT – The Arctic University of Norway and the Norwegian University for Science and Technology – NTNU), as well as the National Library. It is a five year project 2013–2019 funded by the Research Council of Norway with the aim of digitising dialect recordings at the four universities, and making them into a research infrastructure for research in language, language technology, history and other humanities sciences. The main resources developed in the project are three speech corpora: 1) LIA Norwegian - Corpus of older dialect recordings, 2) LIA Sápmi - Sámegiela hállangiellakorpus – Corpus of Sami speech, and 3) CANS – Corpus of American Nordic Speech. The present paper is about the first of these. The technical development and management of the corpus was based at the UiO, while the transcription tasks were divided among the four universities, securing local dialect expertise and also the possibility of creating a thriving dialect environment for interested students.

The recordings in LIA Norsk were transcribed (see Section 2.2), morphologically tagged (using the LIA tagger, a MarMoT tagger trained on the LIA material), and made available in a searchable corpus using the corpus search system Glossa (Nøklestad et al. 2017, Kosek et al. 2015), developed at the UiO. The corpus contains 3.5 Million tokens (interviews and conversations) spoken by 1374 different speakers from 222 different places. The corpus is searchable through a user-friendly interface, with easy click access to the relevant speech segment (audio and/or video) for each hit, a phonetic view with formant structure, pitch, waveform and spectrogram, and a translation option linked to Google Translate.

A side product is more than 200 “bilingual” dialect dictionaries with entries from the dialects and the corresponding words in the standard orthography.

The dialect recordings were from all four universities, and collected at very different times (1930s – 1990s), under very different circumstances and with diverse goals. Some of the recordings were clearly in aid of onomasiology, others of sociolinguistics or morphology (inflectional paradigms). This led to a disparate collection of recordings. The LIA project therefore had limited means of choosing, but at best could reject recordings that were irrelevant. Neither could new places be included, which meant that some places are vastly more represented in the corpus than others, in spite of efforts to represent all areas equally in the corpus. For example, the county of Telemark, which is often regarded as a heaven for dialectologists, only has 8 places, while the neighbouring Oppland county has 13.

Some counties that were not considered interesting by the dialectologists are hardly represented at all, like Vestfold (1 place), and Akershus (3 places), while Hordaland has 18 places (and there were many more to choose from).

2.2 Transcription in the LIA Norwegian

The transcription standard was the same as that used in the previously developed Nordic Dialect Corpus (Johannessen et al. 2014), with both an orthographic transcription (using the standard Nynorsk Norwegian) and a coarse-grained phonetic transcription based on a common standard used for Norwegian dialects, described in Papazian and Helleland (2005). The phonetic transcription was done manually in Elan (Wittenburg et al. 2006), while the orthographic transcription was done semi-automatically using the Oslo Transliterator, developed at UiO, and used for many speech corpora.

The Oslo Transliterator takes as input a phonetically transcribed text and an optional dialect setting. Sets of text manually transliterated to orthography provide a good basis for training the program, enabling it to accurately guess the transliteration in subsequent bodies of text. The training process can be repeated, and the trained version can be used for similar dialects. Transcribing each recording twice, phonetically and orthographically, therefore does not take as much as twice the time of transcribing only once.

The motivation behind the decision to have two transcriptions was that they are good for different purposes. The orthographic transcription is useful for efficient corpus search and for automatic grammatical tagging and parsing. The phonetic transcription is useful since it makes it possible to read the actual pronunciations, and it also makes it possible to select or exclude certain pronunciations when searching the corpus. In addition, it makes it possible to see isoglosses when all the pronunciations of one word are projected to a map (an option which exists in the corpus).

In Map 1 we illustrate this. Here we have searched for the negation *ikkje* 'not', which varies very much across dialects. Choosing to highlight only one aspect of the transcribed words, gives a unique chance to see all occurrences of this particular feature. The corpus user can choose what to highlight. Here we have chosen to mark yellow all variants that are pronounced with a nasal followed by a laminal stop (for example *innte, ænnte, nte, ernte,* etc.).

Without the phonetic transcription, this map would not have been possible to make. However, it would also have been impossible to generate it without the orthographic transcription, since it was based on an orthographic search and a selection based on the phonetic forms of the hits.

Map 1: Yellow markers: All places where speakers have used a form of the negation *ikkje* 'not' containing a nasal and a palatal stop. Red markers: All places where somebody has used the negation in the recordings. (Screenshot from the LIA Norwegian Corpus.)

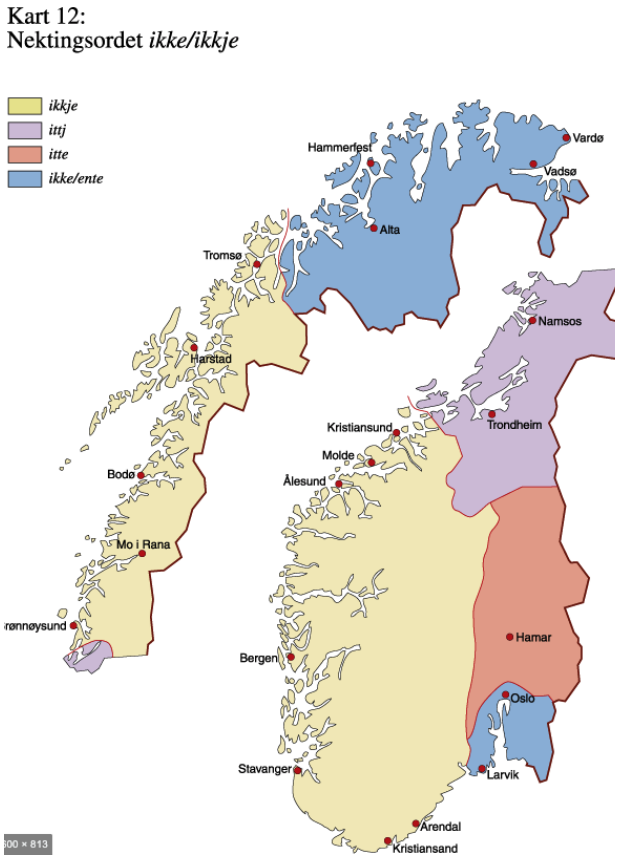
We provide an example of the two transcriptions, as they appear in the concordance hits in the corpus, Figure 1.

eg var	**ikkje**	kar om å # å styre han riktig veit du
je va	**ttje**	kær omm å # å styre n rikkti vett du

Figure 1: Orthographic (top line) and phonetic transcription (bottom line) of one of the more than 50 000 hits for the search for the negation *ikkje*. Translation: 'I wasn't man to steer it correctly, you know'. (Screenshot of one of the hits for speaker aamot_uio_0102 in the LIA Norwegian Corpus.)

It may be instructive to see that an existing map used for schools, Map 2, is massively inaccurate compared with Map 1.

The difference between Map 1 and Map 2 illustrates how much more comprehensive the map based on the speech corpus is compared with the one based on old maps. Traditional dialect maps are based on different materials, methods and sources, but are clearly not as comprehensive as a corpus based on actual recordings of people speaking naturalistically.



Map 2. A school map illustrating the negation word *ikkje* across Norway. It is the blue areas that are relevant and they represent not only pronunciations with nasal+stop, but also the form *ikke*. And even so, big areas along the coast are left out. (Screenshot from Skoleweb.)

2.3 Traditional Norwegian Dialectology

Traditional dialectology divides Norway into four or five areas. One common map looks like Map 3, where the country is divided into North Norwegian, Trønder Norwegian, East Norwegian and West Norwegian.

There are several isoglosses that such maps are based on, going back to the work by Hallfrid Christiansen (1946–48), and described by several authors afterwards. These are mainly: 1) The principle of equal syllable weight, causing infinitival forms to differ in those four areas, 2) tone realisation, separating the west and the north from the other two areas, 3) the retroflex flap, again separating the west and north from the rest (Venås and Skjekkeland 2019). No two isoglosses divide the country in the same way, and there are also others, like all those distinguishing the many ways of pronouncing the negation, palatalisation of dental consonants, the form of the personal pronouns for, for example, 1SG, 1PL, 3SGF.

Map 3: A traditional dialect map of Norway (Mæhlum & Røyneland (2012:178).

3. The Present Study

3.1 The Potential Value of the Present Transcriptions for Automatic Dialect Clustering

The transcription standard chosen for the LIA corpus was motivated in the needs of linguists and dialectologists, as well as in tradition and the available resources. The phonetic transcription is relatively coarse-grained; it can be written with letters of the Norwegian alphabet (the Latin alphabet plus the three extra vowel letters æ, ø and å). In fact, it is mainly written only by the small characters, with the exception of L, characterising the retroflex flap. Because of this coarse "semi-phonetic" transcription, there are more forms that are shared by speakers from several different places, than if a detailed IPA standard had been chosen instead. On the other hand, many details are also overlooked because of the choice of this coarse-grained transcription level.

Furthermore, the fact that we have two sets of transcriptions per dialect, both phonetic and orthographic, means that we have the same standard against which all the dialects can be measured.

The double transcriptions make it tempting to try to measure the Levenshtein distance on the dialects (see Sections 4.2 and 5.2). The fact that the phonetic transcriptions are coarse-grained, also suggest that an automatic clustering method, like neural networks (see Sections 4.3 and 5.3), may give good results. In both cases,

of course, the automatic methodology will not look for any predefined features, so it could be that other features might skew the results in unexpected ways.

The traditional descriptions rely especially on the three features mentioned in Section 2.3. Our transcriptions are not marked for tone, so that specific feature cannot be used by any of the two methods. We have used special marking of the retroflex flap, but this is not a particularly frequent phoneme, so it is not clear that it can be used as a defining feature. However, the feature of equal weight, which applies to all infinitives, should be possible to be made use of.

3.2 Two Datasets of Transcribed Words

We have chosen to use two datasets. Dataset 1 contains all those word pairs (i.e., the words with two transcriptions) that were shared by all the dialects.[1] These were only 23 pairs, and as can be expected, all are function words, i.e. no infinitives, and in fact no retroflex flaps. Three pairs did not vary at all across the 217 dialects (the preposition *på* 'on', the conjunction *og* /å/ 'and' and the infinitival marker *å* 'to'), and one pair only had one deviation from the form that all the others had, it pronounced as /e/ the preposition *i* 'in', pronounced /i/ elsewhere. The 19 pairs that were left were subjunctions, prepositions, pronouns, determiners and the negation. Some of these vary a lot in form across dialects, such as the preposition *av* 'of': *a, ao, av, ta, tao, tå*, the negation *ikkje* 'not': *ikke, ikkje, ikkji, innte, itte, ittj, ittje, kje, nte, tje*, etc., and the 1sg pronoun *eg* 'I': *e, i, je, jeg, jæ, æ, æg*, etc. Since there are no infinitives and no retroflex flaps in this dataset, and of course, no tone-marking, the prediction is that it is unlikely for any automatic method to use the dataset to build a map similar to the traditional map depicted in Map 3.

Dataset 2 consists of the 2000 most frequent words in the entire corpus, but with no requirement that they should also be found in all the dialects. Of course, we know already that only 23 pairs can be found in all, but many words will be found in 90 % of the dialects, for example, both the 3SGF and the 3SGM pronouns are very frequent, but there are some places in which the speakers have not talked about something that made these pronouns necessary. For example, the council Sauherad has got no occurrence of the pronoun *ho* 'she.3SGF', while Sørfrøya has no occurrence of the pronoun *han* 'he.3SGM'.

The most frequent words are of course function words, but occasionally even they are distributed unevenly. For example, the preterite *vart* 'became.PRET' is not used in all parts of the country, as there exists a popular competitor, *ble* 'became.PRET'. Some lexical words that are used in many of the councils are *år* 'year.SG+PL' (used in recordings from 207 out of 217 places), *si* 'say.PRES', *går* 'walks.PRES', *kommer* 'comes.PRES', and *øver* 'practices.PRES'. The rarest ones are *idrett* 'sports', *pene* 'beautiful.POS.PL', *hotell* 'hotel.SG+PL', *morosamt* 'fun.N.SG', and *segle* 'sail.INF'. The latter, and least frequent, word in Dataset 2 is only used in 17 places out of 217 places.

4. The Levenshtein and Neural Network Methods for Dialect Analysis on the Small Dataset 1

Table 1 shows what the datasets look like as input to the algorithms below. The leftmost column contains place names and the next two columns give dialect pronunciations of the words in the heading. The headings (orthographic forms of the words) and place names are fed into the algorithms, where they are used as labels.

	Noko 'something'	**Ikkje 'not'**
Ål	noko	ikkji
Åmli	nåke	kje
Åmot	no	itte
Åsane	noe	kje
Åseral	nåkå	kje

Table 1: Excerpt of dialect matrix used as input for both methods.

4.1 Dialectometry and Dataset 1: Introduction

Quantitative, data-driven dialectology, or dialectometry, often uses edit distance algorithms like the Levenshtein method to compute and measure similarities between related dialects (for example Gooskens and Heeringa 2005, Heeringa, Johnson and Gooskens 2009). Recently a neural network approach has also been proposed (see Rama & Çöltekin 2016). In the following we will present a number of similarity experiments applied to the LIA Norwegian Corpus; we both extract parallel (token) data from the corpus, as well as test the proposed alignment agnostic features of the neural long short term memory (LSTM) autoencoder network.

The experiments show that the semi-phonetic transcription standard used in the LIA project to a large extent replicates the dialectal maps that have figured in the literature at least since Christiansen (1946/1969). In order to produce the present dialect maps we use the open analysis tool Gabmap, with which we concentrate on the fuzzy clustering functionality. Gabmap is both documented (Nerbonne et al. 2011; Leinonen et al. 2016) and reviewed (Snoek 2014).

4.2 Levenshtein and Dataset 1

The Levenshtein method computes the edit distance between two transcribed words (phonetic realizations of the same lexeme in two dialects) at a time (see Figure 2). This is done on a segmental basis, where total similarity gives the value zero, and then there can be graded numerical values according to the phonetic distance between each comparison pair of sounds. The lower the value, the higher the similarity is between the two transcribed words compared.

[1] In the transcriptions it is often the case that one orthographic word form has several different phonetic realisations. We have simplified the data and chosen only the most frequent one for each word.

Philadelphia — Jefferson

d	ʒ	ɔ	ɚ	d	ʒ	ə	
d	ʒ	ɔˆ	r̩	d	ʒ	ə	
		0.5	1				1.5

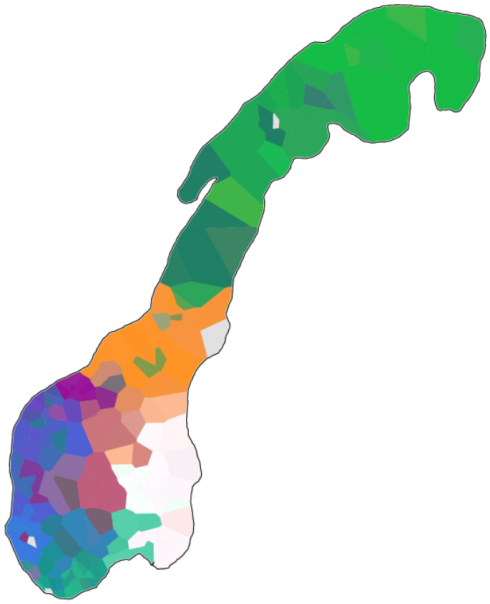
Philadelphia — Lancaster

d	ʒ	ɔˆˇ	ɚ	d	ʒ	ə	
t̲	ʃ	ɔ	r	t̲	ʃ	ə	
1	1	0.5	1	1	1		5.5

Figure 2: Example of computation of Levenshtein distance between two strings taken from Nerbonne et al. (2011).

To compute a Levenshtein distance between dialects one needs parallel data. Since the data in LIA Norwegian are from naturalistic conversations, and not collected using standardized questionnaires, the amount of parallel data risks being scarce. Yet there are certain highly frequent words as described above that occur across all the dialect in LIA Norwegian. These words were extracted for all dialects and put in a spreadsheet where each cell is a dialect-word pair also called a site-item pair. This matrix was then fed to Gabmap. As described in Leinonen et al. (2016), Gabmap uses the string edit distance or Levenshtein distance to measure the (linguistic) distance between the dialects in the spreadsheet; other types of input data are also possible to input. The Levenshtein method has been used before for Norwegian and Nordic languages (Gooskens and Heeringa 2005, Heeringa, Johnson and Gooskens 2009), but with other data and other goals than those of the present one, so they are not directly comparable.

Using Dataset 1, the resulting map is presented in Map 4a.



Map 4a: Noisy cluster map for the Levenshtein method for Dataset 1.

Given that Map 4a is in practice only based on 19 word-pairs, the map shows remarkably clear areas, and the four areas of Map 3 are easily recognised in Map 4a.[2] The orthographic transcription was not used in the calculations, but it helped the program to understand which phonetic strings to compare with which.

4.3 Neural Networks and Dataset 1

Unlike the Levenshtein method, the neural networks method does not compare word pairs. Instead, it generalizes over all the phonetic forms in a single dialect and then these generalisations are compared in order to obtain a measure of the similarity between the dialects. The method uses an autoencoder consisting of a pair of LSTM networks, which is an example of a recurrent neural network with a gating mechanism. The autoencoder consists of an encoder and a decoder where the encoder encodes a phonetic form into a hidden state and the decoder maps the hidden state back to the same phonetic representation, see Figure 3, and Rama & Çöltekin (2016: 26-7) for a detailed description. Rama & Çöltekin argue that their method have a number of advantages over the classical string edit distance: 1) it does not need explicitly aligned data, since the hidden state of the autoencoder groups similarly transcribed words together and 2) it can discover long-distance word internal dependencies, such as vowel harmony.

When we train our models we use the same network configuration as Rama & Çöltekin (2016). The output of the neural network is converted to a site-site matrix and input to Gabmap as difference data rather than as discrete string representations.

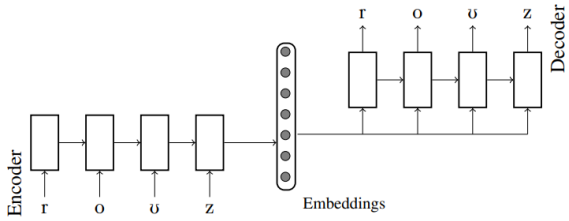
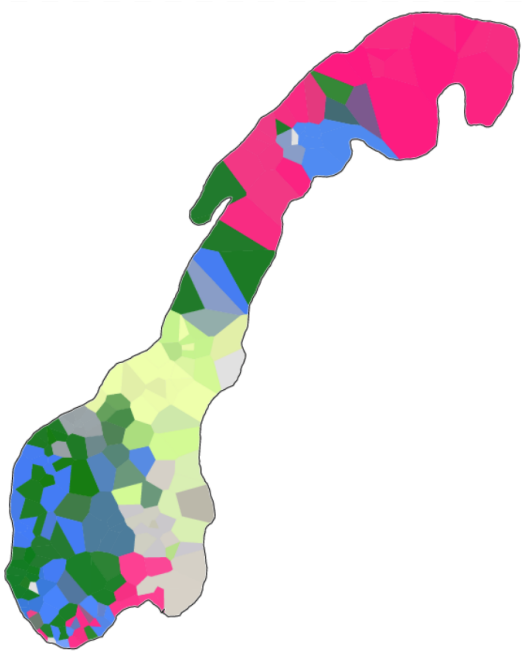


Figure 3: An illustration of the autoencoder with a pair of LSTM networks functioning as encoder and decoder, respectively, with a hidden state of seven dimensions (Rama & Çöltekin 2016).

The result of the neural network model applied to Dataset 1 is given in Map 4b.

[2] Here and in the rest of the paper, similarity between maps is based on visual inspection. Our next step is to use an image similarity metric, such as Sum-Square-Difference as implemented in OpenCV Template Matching.



Map 4b: Noisy cluster map for LSTM autoencoder for Dataset 1.

Although it is possible to detect some areas in this map, too, the areas are much less clearly marked than on Map 4a.

4.4 Dataset 1: Summary

The results of the two methods on the tiny Dataset 1, in practice only 19 word pairs, are clearly seen in Maps 4a and 4b. The Levenshtein method makes good use of the transcriptions, and since the same word pairs are present in all the places, there is no noise: Every phonetic transcription is compared with the other transcriptions of the same word, so all the results are maximally comparable. Map 4a has clearly captured the main dialect areas, as can be seen when it is compared with the traditional Map 3.

The neural network method, on the other hand, obviously finds the dataset too small, and the differences between the transcriptions too large. Map 4b, resulting from this method, is not as clearly representing the same dialect areas of Map 3, and is inferior to the Levenshtein method with such a small set of data as Dataset 1. Since machine learning methods generalize over data, they require a sufficient amount, and clearly a set of 19 words is too small.

The prediction in Section 3.2, that Dataset 1 would be of little use for marking dialect areas, given that it contains none of the three major dialect-characterising features, no retroflex flaps, no infinitives and no tone description, turns out to be false. Clearly, Dataset 1, with its many function words, has captured core differences and similarities between the dialects.

5. The Levenshtein and the Neural Network Methods for Dialect Analysis on the Large Dataset 2

5.1 Dialectometry and Dataset 2: Introduction

The results from Dataset 1 showed very clear differences for the two methods applied to the two datasets. However, Dataset 2 is quite large, with 2000 words, and contains many verbs in the infinitival form, many words containing the retroflex flap, though still no tone making. It will also contain many other dialect differences, like pronouns and the negation. It is to be expected that Dataset 2 will fare quite well with the neural network method due the increased data size, and quite badly with the Levenshtein method, since many phonetic strings will have no counterpart in a number of the other dialects, and thus generate faulty alignments.

5.2 The Levenshtein Method and Dataset 2

The result from applying the Levenshtein method on Dataset 2 is given in Map 5a. Contrary to the prediction, the Levenshtein method has resulted in a map with well-defined dialect areas, in spite of the fact that there are lots of faulty alignments.

Map 5a: Noisy cluster map for the Levenshtein method for Dataset 2.

5.3 Neural Networks and Dataset 2

The neural network method applied on Dataset 2 is depicted in Map 5b.

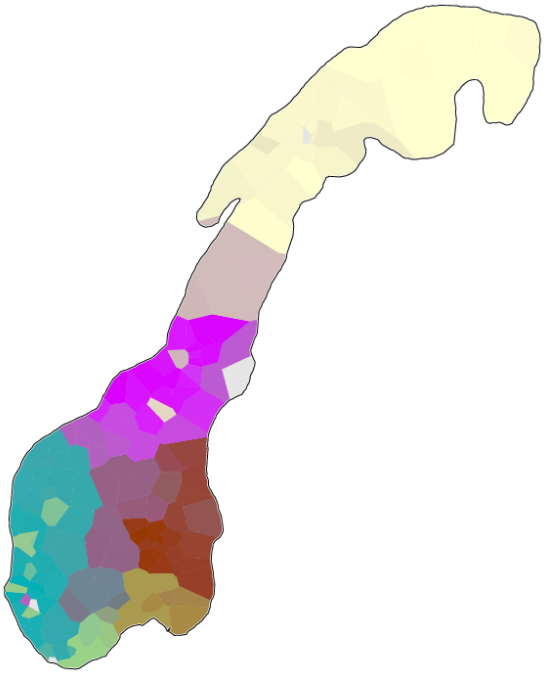


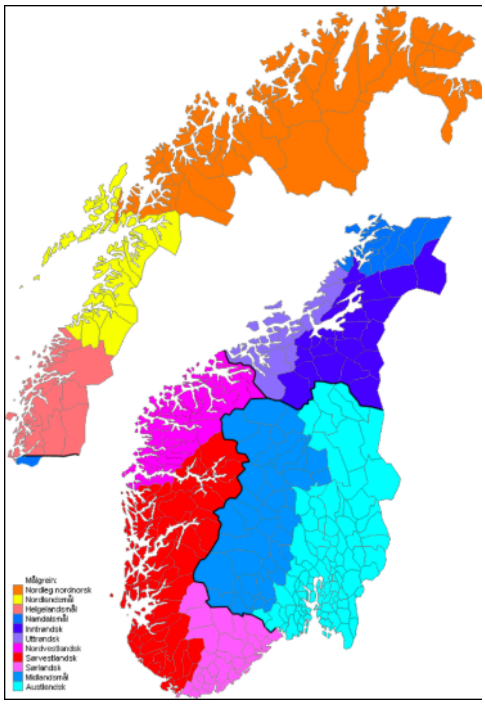
Figure 5b: Autoencoder noisy cluster map for top 2000 word forms.

Map 5b is also clearly depicting the major four dialect areas illustrated in Map 3. This is not so surprising, but still a welcome result.

5.4 Dataset 2: Summary

The two methods applied to Dataset 2 have both captured the four main dialect areas in Map 3. This means that they have been able to generalise over the forms presented by the transcriptions. There are many such differences, not just the shape of the infinitives and the retroflex flaps, but also the forms of function words like the negation and pronouns, preterite suffixes, prepositions, palatalisation of consonants, patterns of plosive voicing, the form of question words, feminine noun suffixes, masculine noun plural suffixes, strong verb present tense suffixes, etc. Each of these dialect features has their own isogloss dividing the map in different ways (see, for example, Skjekkeland 1997).

The two dialectometry methods have in fact captured more than the four major dialect areas from Map 3. We see that there is a separate colour down this middle of South Norway (purple-coloured in both Maps 5a & 5b), and also a smaller separate colour in the very south (dark purple in Map 5a and light green in Map 5b). These represent further dialect distinctions, and if we look at a different, more fine-grained map based on the traditional dialect features, such as Map 6, we recognise those extra areas that we see in Maps 5a & 5b. They are also suggested in Maps 4a & 4b, but less clearly.



Map 6: More fine-grained dialect areas, shown in a traditional type of map. (Wikiwand.com: Målmerke)

Maps 5a & 5b with their similarity to the traditional Map 6, illustrate that both methods make use of all the transcriptions and cluster in ways that represent the generalised traditional maps 3 & 6. This is remarkable, since we know that each dialect feature has its own isogloss, as depicted in Map 1 (with the negations containing a nasal), and that some of these may not even be known, as suggested by Map 2. However, since a language is not just a collection of isoglosses, but can be described by dialect areas, there must be some clustering of the isoglosses reflecting real dialect areas. This has been depicted in the traditional dialect maps, and is confirmed by the two automatic dialectometrical methods.

6. Conclusion

We started out with a transcribed corpus of dialects from across all of Norway. The phonetic transcription standard was quite coarse-grained, in order for the transcription to be quickly carried out by the manual transcribers using the alphabetic keys of an ordinary keyboard. This coincided with a standard way of transcribing dialects among dialectologists. Since the transcription standard used was coarse-grained, the present authors assumed that this would aid transcriptions to be categorised automatically to generate maps.

This paper has shown that a coarse-grained transcription of speech is sufficient to replicate known dialectal boundaries, and perhaps even discover more dialect areas. We have explored different methods of visualizing the dialect data. For the smaller Dataset 1, with only words represented in every dialect (23 words overall, but only 19 that actually showed some degree of variation), we found that the Levenshtein method gave a much more reasonable map then the neural network model, confirming that neural networks need more data in order to generalise. Next we

followed Rama & Çöltekin (2016: 31) and provided the network with "a few thousands of words ...", the transcriptions of the 2000 most frequent words of the corpus as a whole, Dataset 2. The neural network method worked much better on this dataset, but surprisingly, did not set it apart from the Levenshtein method. Rather, the neural network gives smoother transitions between the clusters, as is expected.

What we draw from the experiments is that the Levenshtein method is very good for small datasets, especially when there is an explicit (orthographic) norm against which each transcribed word can be compared. Both methods are good when there is much data, even if the Levenshtein method then has to compare some of the transcribed words with a null representation in some other dialects.

Our work differs from that of Rama & Çöltekin (2016) in that we have chosen Norwegian dialects rather than Dutch and German, that the transcriptions are new, that our transcriptions are more coarse-grained than the Dutch and German ones, that our word selection comes from a real dialogue corpus (thus containing many function words), and that we have compared the generated maps with traditional maps drawn by dialectologists.

The explorative nature of the present study does warrant a close-reading of the different data matrices, but should, however, invite both computer scientists and linguists to explore new data and new methods. In the future we would like to extend our analyses to different levels of grammar, for example syntax, and possibly aggregate the analyses based on metadata like year of recording to see if it is possible to observe diachronic phenomena in the data. There are also other spoken language corpora that have never been subject to such an analysis, for example CANS and LIA Sápmi. A final point would be to optimize the hyperparameters of the LSTM network.

7. Bibliographical References

Christiansen, Hallfrid. 1946-48: *Norske dialekter*.

Gooskens, Charlotte and Heeringa, Wilbert. (2005). De moderne norske dialekters stilling indenfor den nordiske sprogfamilie. In Svein Lie, Gudlaug Nedrelid and Helge Omdal (eds.): *Utvalde artiklar frå det tiande Møte om norsk språk i Kristiansand* (MONS10). Kristiansand: Høyskoleforlaget

Heeringa, Wilbert, Keith Johnson and Charlotte Gooskens. (2009). Measuring Norwegian Dialect Distances using Acoustic Features. *Speech Communication* 51, 167–183.

Johannessen, Janne Bondi; Vangsnes, Øystein Alexander; Priestley, Joel; Hagen, Kristin. (2014). A multilingual speech corpus of North-Germanic languages. In Raso, Tommaso; Mello, Heliana (eds.): *Spoken Corpora and Linguistic Studies.* John Benjamins Publishing Company, p. 69-83.

Kosek, Michal, Anders Nøklestad, Joel Priestley, Kristin Hagen, and Janne Bondi Johannessen. (2015). In Gintarė Grigonytė, Simon Clematide, Andrius Utka and Martin Volk (eds.): Visualisation in speech corpora: maps and waves in the Glossa system, *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania, NEALT Proceedings Series 25*, 23–31.

Leinonen, Therese, Çagri Çöltekin, and John Nerbonne. (2016). Using Gabmap. *Lingua*, 178:71–83.

Mæhlum, Brit og Røyneland, Unn. (2012). *Det norske dialektlandskapet*, Oslo: Cappelen Damm Akademisk.

Nerbonne, John, Peter Kleiweg, Wilbert Heeringa & Franz Manni. (2008). "Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering", in Christine Preisach, Lars Schmidthieme, Hans B urkharty & Reinhold Decker (eds.), Data Analysis, Machine Learning, and Applications. *Proc. of the 31st Annual Meeting of the German Classification Society*, Berlin: Springer, 647-654.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. (2011). Gabmap – a webapplication for dialectology. *Dialectologia*, Special Issue II:65–89.

Nøklestad, Anders, Hagen, Kristin, Johannessen, Janne Bondi, Kosek, Michal and Joel Priestley. (2017). A modernised version of the Glossa corpus search system. In Jörg Tiedemann (ed.): *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*. 2017, 251-254.

OpenCV. https://docs.opencv.org/2.4.13.7/index.html

Papazian, Eric & Helleland, Botolv. (2005). *Norsk talemål*. Kristiansand: Høyskoleforlaget.

Rama, Taraka, and Çağrı Çöltekin. (2016). "LSTM autoencoders for dialect analysis." *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*.

Skjekkeland, Martin. 1997. *Dei norske dialektane. Tradisjonelle særdrag i jamføring med skriftmåla.* Kristiansand: Høyskoleforlaget.

Venås, Kjell & Martin Skjekkeland. (2019). Dialects i Norway. *Store Norske Leksikon:* https://snl.no/dialekter_i_Norge

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In: *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.

8. Language Resource References

Elan software https://tla.mpi.nl/tools/tla-tools/elan/ Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands

LIA (Language Infrastructure made Accessible) http://tekstlab.uio.no/LIA/

LIA Norwegian - Corpus of older dialect recordings https://tekstlab.uio.no/glossa2/lia_norsk

LIA Sápmi - Sámegiela hállangiellakorpus – Corpus of Sami speech https://tekstlab.uio.no/glossa2/saami

CANS – Corpus of American Nordic Speech https://tekstlab.uio.no/glossa2/cans2

The Oslo Transliterator https://www.hf.uio.no/iln/english/about/organization/text-laboratory/services/oslo-transliterator/

Skoleweb https://www.skoleweb.net/fagsider-norsk-spraak-dialekter-kart