Finnish Dialect Identification: The Effect of Audio and Text

Mika Hämäläinen, Khalid Alnajjar, Niko Partanen and Jack Rueter
Department of Digital Humanities
University of Helsinki & Rootroo Ltd
firstname.lastname@helsinki.fi

Abstract

Finnish is a language with multiple dialects that not only differ from each other in terms of accent (pronunciation) but also in terms of morphological forms and lexical choice. We present the first approach to automatically detect the dialect of a speaker based on a dialect transcript and transcript with audio recording in a dataset consisting of 23 different dialects. Our results show that the best accuracy is received by combining both of the modalities, as text only reaches to an overall accuracy of 57%, where as text and audio reach to 85%. Our code, models and data have been released openly on Github and Zenodo.

1 Introduction

We present an approach for identifying the dialect of a speaker automatically solely based on text and on audio and text together. We compare the unimodal approach to the bimodal one. There are no previous dialect identification approaches for Finnish. There are several situations were a dialect identification method can be of use. For example, if we have ASR models fine tuned for specific dialects, the dialect identification from audio could be used as a preprocessing step. The model could also be used to label recorded materials automatically in order to create archival metadata. In order to make our contribution useful for others, we have released our code, models and processed data openly on GitHub[1] and Zenodo[2].

Finnish is a large Uralic language that is one of the official languages of Finland, and is used essentially at all levels of the modern society. There are approximately five million Finnish speakers. The language belongs to the Finnic branch of the Uralic language family, and is very closely related to Karelian, Meänkieli and Kveeni, and is also closely related to the Estonian language. It is more distantly related to numerous Uralic languages spoken in Russia.

The history of written Finnish starts in the 16th century. Current orthography is connected to this written tradition, which developed into the current form in the late 19th century with a conscious planning and systematic development of the lexicon. After this, the changes have been minor (Häkkinen, 1994, 16), and also impacted lexicon, especially what it comes to the development of the vocabulary of the modern society and traditional agrarian terminology becoming less known.

The Finnish spoken language, however, is still largely based on Finnish dialects. In the 20th century some of the strongest dialectal features have been disappearing, but there are still clearly distinguishable spoken vernacular varieties that are regionally marked. It has been shown that instead of clear disappearance of dialects there are various features that are spreading, but not at uniform rate, and reportedly younger speakers use the areally marked features less than the older speakers (Lappalainen, 2001, 92). Finnish vernaculars also represent historically rather different Finnic varieties, with major split between Eastern and Western dialects. There are, however, also dialect continuums and traditionally found gradual differentiation from region to region.

Many of the changes have been lexical due to technical innovations and modernization of the society: orthographic spelling conventions have largely remained the same. Spoken Finnish, on the other hand, traditionally represents an areally divided dialect continuum, with several sharp boundaries, and many regions of gradual differentiation from one municipality to another municipality.

As mentioned, in the later parts of the 20th century relatively strong dialect leveling has been taking place. Some of the Finnish dialects may already be concerned endangered, although the complex relationship between contemporary vernaculars and

[1] https://github.com/Rootroo-ltd/FinnishDialectIdentification
[2] https://zenodo.org/record/5330673

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8777–8783
November 7–11, 2021.

the most traditional dialectal forms makes this hard to ascertain. Dialect leveling in itself is a process known from many parts of Europe (Auer, 2018). However, in the case of Finnish the written standard has remained relatively far from the spoken Finnish, besides individual narrow domains such as news broadcasts were the written form is used also in speech.

Additionally there have been distinct text collections that include materials from this dialect archive. These include dialect books specific regions and municipalities, such as Oulun murrekirja [Dialect Book of Oulu] (Pääkkönen, 1994) or Savonlinnan seudun murrekirja [Dialect book of Savonlinna region] (Palander, 1986). There have also been more recent larger collections that contains excerpts from essentially all dialects (Lyytikäinen et al., 2013).

Especially in the later parts of 21th century the spoken varieties have been leveling away from very specific local dialects, and although regional varieties still exist, most of the local varieties have certainly became endangered. Similar processes of dialect convergence have been reported from different regions in Europe, although with substantial variation (Auer, 2018). In the case of Finnish this has not, however, resulted in merging of the written and spoken standards, but the spoken Finnish has remained, to our day, very distinct from the written standard. In a late 1950s, a program was set up to document extant spoken dialects, with the goal of recording 30 hours of speech from each municipality. This work resulted in very large collections of dialectal recordings (Lyytikäinen, 1984, 448-449). Many of these have been published, and some portion has also been manually normalized. Dataset used is described in more detail in Section 3 Data.

In Finnish linguistics the dialect identification has primarily been studied in the context of folk linguistics. In this line of research the perceptions of native speakers are investigated (Niedzielski and Preston, 2000). This type of studies have been done for Finnish, for example, by Mielikäinen and Palander (2014), Räsänen and Palander (2015) and Palander (2011). It has been possible to suggest for individual dialects which features are the most stable and will remain as local regional markers, and which seem to be in retention (Räsänen and Palander, 2015, 25). In this study we conduct just individual experiments and report their results, but in the further research we hope the results could be analyzed in more detail in connection with the earlier dialect perception studies, as we believe the differences in perceived dialect differences could be compared to the difficulties and successes the model has to differentiate individual varieties.

Dialect	Short	Sentences
Etelä-Häme	EH	1860
Etelä-Karjala	EK	813
Etelä-Pohjanmaa	EP	2684
Etelä-Satakunta	ES	848
Etelä-Savo	ESa	1744
Eteläinen Keski-Suomi	EKS	2168
Inkerinsuomalaismurteet	IS	4035
Kaakkois-Häme	KH	8026
Kainuu	K	3995
Keski-Karjala	KK	1640
Keski-Pohjanmaa	KP	900
Länsi-Satakunta	LS	1288
Länsi-Uusimaa	LU	1171
Länsipohja	LP	1026
Läntinen Keski-Suomi	LKS	857
Peräpohjola	P	1913
Pohjoinen Keski-Suomi	PKS	733
Pohjoinen Varsinais-Suomi	PVS	3885
Pohjois-Häme	PH	859
Pohjois-Karjala	PK	4292
Pohjois-Pohjanmaa	PP	1801
Pohjois-Satakunta	PS	2371
Pohjois-Savo	PSa	2344

Table 1: Dialects and the number of sentences in each dialect in the corpus

2 Related work

The current approaches to Finnish dialect have focused on the textual modality only. Previously, bidirectional LSTM (long short-term memory) based models have been used to normalize Finnish dialects to standard Finnish (Partanen et al., 2019) and to adapt standard Finnish text into different dialectal forms (Hämäläinen et al., 2020). Similar approach has also been used to normalize historical Finnish (Hämäläinen et al., 2021; Partanen et al., 2021).

The closest research to our paper conducted for Finnish has been detection of foreign accents from audio. Behravan et al. (2013) have detected foreign accents from audio only by using i-vectors. However, foreign accent detection is a very different task to native speaker dialect detection. Many foreign accents have clear cues through phonemes that are not part of the Finnish phonotactic system, where as with dialects, all phonemes are part of Finnish.

There have been several recent approaches for Arabic to detect dialect from text (Balaji et al., 2020; Talafha et al., 2020; Alrifai et al., 2021). Textual dialect detection has been done also for German (Jauhiainen et al., 2018), Romanian (Zaharia et al., 2021) and Low Saxon (Siewert et al., 2020). The methods used range from traditional machine learning with features such as n-grams to neural models with pretrained embeddings, as it is typically the case in NLP research. None of these approaches use audio, as they rely on text only.

At the same time, North Sami dialects have been identified from audio by training several models, kNNs, SVMs, RFs, CRFs, and LSTM, based on extracted features (Kakouros et al., 2020). Kethireddy et al. (2020) use Mel-weighted SFF spectrogram to detect spoken Arabic dialects. Mel spectograms are also used by Draghici et al. (2020). All these approaches are mono-modal and use only audio.

Based on our literature review, the existing approaches use either text or audio for dialect detection. We, however, use both modalities and apply them on a language with no existing dialect detection models.

3 Data

The Finnish dialects are exceptionally well documented. In the 1950s the Finnish dialect archive was formed with the goal of recording 30 hours of speech from each Finnish municipality. This goal was reached fast, and exceeded, resulting in a very large collection of archived materials that is stored in the Institute for the Languages of Finland (Lyytikäinen, 1984, 448-449), and known as *Tape Archive of the Finnish Language*[3]. There have been numerous publications based on these materials, although it is hard to estimate into which extent this covers the entire body of recorded work, which totals 24,000 hours of audio.

The largest individual publication of these materials is beyond doubt the *Samples of Spoken Finnish* series that was published in 1978–2000 as 50 booklets.[4] Each book contained approximately two hours of transcriptions, from two different speakers, and represents a different municipality. Later these materials have been digitized and published as an openly licensed dialect corpus (Institute for the Languages of Finland, 2014). There are also other related corpora, most importantly *The Finnish Dialect Syntax Archive* that contains similar recordings annotated morphosyntactically (University of Turku and Institute for the Languages of Finland, 1985). Since 1980s follow-up research has been done in selected municipalities to track the changes in the dialects (Lyytikäinen and Yli-Paavola, 2010, 413), which is another significant line of research that complements these older dialect materials.

Later the work on these published materials has resulted in multiple electronic corpora that are currently available. Although they represent only a tiny fraction of the entire recorded material, they reach remarkable coverage of different dialects and varieties of spoken Finnish. Some of these corpora contain various levels of manual annotation, while others are mainly plain text with associated metadata. Materials of this type can be characterized by an explicit attempt to represent dialects in linguistically accurate manner, having been created primarily by linguists with formal training in the field. These transcriptions are usually written with a transcription systems specific for each research tradition. The result of this type of work is not simply a text containing some dialectal features, but a systematic and scientific transcription of the dialectal speech.

The corpus we have used in this study is the above-mentioned Samples of Spoken Finnish corpus (Institute for the Languages of Finland, 2014). The electronic version contains manually annotated normalization into standard Finnish. The corpus is almost 700,000 tokens large. The digital version, including audio, is published with CC-BY license, and is available in the Language Bank of Finland.[5] We have selected it into this study because of the open license and large dialectal scope. We have downloaded the corpus with the original audio files, and extracted from the audio all utterances that are shorter than 10 seconds in length. The dialect region classification is taken directly from the corpus metadata.

Despite the successful attempt of the authors of the corpus to include all dialects, the dialects are not entirely equally represented in the corpus. One reason for this is certainly the different sizes of the dialect areas, and the variation introduced by different speech rates of individual speakers. The difference in the number of sentences per dialect can be seen in Table 1. We do not consider this

[3] https://www.kotus.fi/en/corpora_and_other_material/spoken_language_corpora

[4] https://www.kotus.fi/aineistot/puhutun_kielen_aineistot/murreaanitteita/suomen_kielen_naytteita_-sarja

[5] http://urn.fi/urn:nbn:fi:lb-201407141

uneven distribution to be a problem, as it is mainly a feature of this dataset. The data has been tokenized and the dialectal transcriptions are aligned with audio on a sentence level. This makes our task with the dialect detection model easier as no alignment is required. We randomly sort the sentences in the data and split them into a training (70% of the sentences), validation (15% of the sentences) and test (15% of the sentences) sets. This means that the models are trained and tested on a sentence level rather than on smaller chunks.

4 Dialect detection

In this section, we describe the two different models we used to detect dialect automatically in the corpus. The first method is based on text only and the second method uses text and audio. Both of the methods used the same training, validation and test splits.

4.1 Text only model

We train a dialect classification model using a bi-directional long short-term memory (LSTM) based model (Hochreiter and Schmidhuber, 1997) by using OpenNMT-py (Klein et al., 2017) with the default settings except for the encoder where we use a BRNN (bi-directional recurrent neural network) (Schuster and Paliwal, 1997) instead of the default RNN (recurrent neural network), since BRNN based models have been shown to provide better results in a variety of tasks.

We use the default of two layers for both the encoder and the decoder and the default attention model, which is the general global attention presented by Luong et al. (2015). The models are trained for the default of 100,000 steps. The model receives dialectal text[6] as input and predicts a dialect name as an output.

4.2 Text and audio model

Our multi-modal model makes use of the dialectal text and audio. The model combines BERT (Devlin et al., 2019) and XLSR-Wav2Vec2 (Baevski et al., 2020) neural models trained on Finnish data. We utilize the uncased Finnish BERT model[7] (Virtanen et al., 2019). The multilingual XLSR-Wav2Vec2 model released by Facebook does not support Finnish. Therefore we use a Finnish XLSR-Wav2Vec2 model[8] that is fine-tuned using readily available Finnish audio datasets: Finnish Common Voice (Ardila et al., 2020), CSS10 Finnish (Park and Mulc, 2019) and Finnish parliament session 2[9] for 30 epochs. All audio input is resampled to 16kHz.

Our multi-modal model follows a siamese neural network architecture, where one side of the network is dedicated to text and the other to audio. We ensure that both sides produce an equal size of features by 1) setting a fixed input length to BERT where padding and truncating is applied where necessary and 2) having two average pooling layers following the output of each side. For the textual output, a global average pooling is applied, whereas an adaptive average pooling is applied to the audio output. Afterwards, the pooled output is concatenated and followed by a dropout layer with a probability of 20%. Lastly, a fully connected dense layer is employed as the classification layer. In total, the network has 439 million trainable parameters and we fine-tuned it for 3 full epochs with a learning rate of 1e-4.

5 Results

The results of the two models can be seen in Table 2. These results were calculated using scikit-learn[10] (Pedregosa et al., 2011). It is clear from the results that the text only model performed worse for every single dialect than the audio and text model. In terms of overall accuracies, the text based model reached only an accuracy of 57%, where as the text and audio based model reach to an accuracy of 85%. This indicates that the audio has classificatory features that are not represented in the text version alone, although the text is in a transcription system that accurately captures various dialectal phenomena.

When comparing the per dialect performance of the better model with the amount of data available for each dialect, we can make an interesting observation that a high amount of data does not equal to a high F1-score. Out of the 10 dialects with the largest amount of samples in the data, only 3, *Kaakkois-Häme*, *Inkerinsuomalaismurteet* and *Kainuu*, reached to an F1-score of at least 0.90.

[6] We also experimented with a character-level model using the same neural network structure, but the accuracy remained low, only 37%

[7] https://huggingface.co/TurkuNLP/bert-base-finnish-uncased-v1

[8] https://huggingface.co/aapot/wav2vec2-large-xlsr-53-finnish

[9] http://urn.fi/urn:nbn:fi:lb-2017020201

[10] sklearn.metrics.classification_report

	bi-LSTM on text			Audio + BERT		
	precison	recall	f1	precision	recall	f1
EH	0.49	0.48	0.49	0.97	0.89	0.93
EK	0.51	0.44	0.47	0.86	0.57	0.69
EP	0.72	0.67	0.69	0.68	0.93	0.79
ES	0.5	0.53	0.51	0.79	0.82	0.8
Esa	0.38	0.37	0.38	0.6	0.97	0.74
EKS	0.44	0.38	0.41	0.9	0.85	0.87
IS	0.74	0.75	0.75	0.96	0.86	0.91
KH	0.67	0.74	0.7	0.86	0.97	0.92
K	0.53	0.49	0.51	0.97	0.83	0.9
KK	0.57	0.54	0.56	0.92	0.95	0.93
KP	0.46	0.45	0.46	0.81	0.87	0.84
LS	0.47	0.38	0.42	0.98	0.74	0.84
LU	0.56	0.52	0.54	0.97	0.98	0.98
LP	0.34	0.32	0.33	0.94	0.92	0.93
LKS	0.34	0.46	0.39	0.72	0.99	0.83
P	0.55	0.58	0.56	0.71	0.93	0.81
PKS	0.4	0.38	0.39	0.93	0.62	0.75
PVS	0.75	0.72	0.73	0.91	0.74	0.82
PH	0.32	0.31	0.31	0.83	0.63	0.72
PKS	0.6	0.58	0.59	0.92	0.8	0.86
PP	0.4	0.45	0.42	0.74	0.38	0.5
PS	0.5	0.53	0.51	0.9	0.89	0.89
PSa	0.43	0.47	0.45	0.68	0.87	0.76

Table 2: Results for the two models

The F1-score of the dialect with the second highest number of samples, *Pohjoinen Keski-Suomi*, was only 0.86. Other dialects that had an F1-score of at least 0.9 were the 11th most resourced *Etelä-Häme*, the 14th most resourced *Keski-Karjala* and the 16th and 17th most resourced *Länsi-Uusimaa* and *Länsipohja*.

The lowest F1-score was 0.5 for *Pohjois-Pohjanmaa*. This is interesting as the dialect is the 12th most resourced one. Even the two least resourced dialects in our dataset, *Etelä-Karjala* and *Pohjoinen Keski-Suomi* got higher F1-scores, 0.69 and 0.75 respectively. These results are an indication that some of the dialects are more clearly marked making them easier to detect even with less data, while some other dialects may have undergone a process of dialect leveling (see Hinskens 1998) making them less distinct from other dialectal forms of Finnish. It is also possible that some dialects are already significantly close to one another, and thereby the model simply cannot distinguish them accurately. Further error analysis could reveal important details of this type.

6 Conclusions

We have presented the first model for Finnish dialect classification for a relatively large number of different dialects, 23 in total. Based on our experiments, a text only model is not as effective in dialect classification as a model with text and audio. It is clear that the amount of data alone is not the only variable that constitutes a high performance of the model for a given dialect, but also how distinctive a given dialect is from other dialects. Since the speakers in the test set were not present in the training, we are confident that the dialect is the feature that the model has learned to predict.

Using the audio materials offers in itself new interesting possibilities for dialect clustering and comparison. Traditional dialect atlases have also been used in automatic comparison and grouping of different Finnish dialects (Syrjänen et al., 2016). In further research we believe also this kind of information could be connected to the analysis to show how the dialect identification exactly interacts with the dialectal variation and differences at close municipality level. At the same time the identifiability of a dialect must be connected to the degree of dialect leveling, linguistic distances and differences between them, so applying the model into newer recordings could also yield information about these processes.

We have made all the data, code and models openly available on Github[11] and Zenodo[12]. We believe that this is the only way to ensure this line of research continues for the Finnish language in the future as well.

References

Khaled Alrifai, Ghaida Rebdawi, and Nada Ghneim. 2021. Arabic tweeps dialect prediction based on machine learning approach. *International Journal of Electrical & Computer Engineering (2088-8708)*, 11(2).

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Peter Auer. 2018. Dialect change in europe–leveling and convergence. *The Handbook of Dialectology*, pages 159–76.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33.

[11]https://github.com/Rootroo-ltd/FinnishDialectIdentification

[12]https://zenodo.org/record/5330673

Appiah Balaji, Nitin Nikamanth, and Bharathi B. 2020. Semi-supervised fine-grained approach for Arabic dialect detection task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 257–261, Barcelona, Spain (Online). Association for Computational Linguistics.

Hamid Behravan, Ville Hautamäki, and Tomi Kinnunen. 2013. Foreign accent detection from spoken Finnish using i-vectors. In *INTERSPEECH*, volume 2013, page 14th.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Alexandra Draghici, Jakob Abeßer, and Hanna Lukashevich. 2020. A study on spoken language identification using deep neural networks. In *Proceedings of the 15th International Conference on Audio Mostly*, AM '20, page 253–256, New York, NY, USA. Association for Computing Machinery.

Kaisa Häkkinen. 1994. *Agricolasta nykykieleen: suomen kirjakielen historia*. Söderström.

Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. 2021. Lemmatization of historical old literary Finnish texts in modern orthography. In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 189–198, Lille, France. ATALA.

Mika Hämäläinen, Niko Partanen, Khalid Alnajjar, Jack Rueter, and Thierry Poibeau. 2020. Automatic dialect adaptation in Finnish and its effect on perceived creativity. In *11th International Conference on Computational Creativity (ICCC'20)*. Association for Computational Creativity.

FLMP Hinskens. 1998. Dialect levelling: a two-dimensional process. *Folia Linguistica Historica*, 32:35–51.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Institute for the Languages of Finland. 2014. Suomen kielen näytteitä - Samples of Spoken Finnish [online-corpus], version 1.0. http://urn.fi/urn:nbn:fi:lb-201407141.

Tommi Sakari Jauhiainen, Heidi Annika Jauhiainen, Bo Krister Johan Linden, et al. 2018. Heli-based experiments in swiss german dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. The Association for Computational Linguistics.

Sofoklis Kakouros, Katri Hiovain, Martti Vainio, and Juraj Šimko. 2020. Dialect identification of spoken north s\'ami language varieties using prosodic features. *arXiv preprint arXiv:2003.10183*.

Rashmi Kethireddy, Sudarsana Reddy Kadiri, Paavo Alku, and Suryakanth V. Gangashetty. 2020. Mel-weighted single frequency filtering spectrogram for dialect identification. *IEEE Access*, 8:174871–174879.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proc. ACL*.

Hanna Lappalainen. 2001. Sosiolingvistinen katsaus suomalaisnuorten nykypuhekieleen ja sen tutkimukseen. *Virittäjä*, 105(1):74–74.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Erkki Lyytikäinen. 1984. Suomen kielen nauhoitearkiston neljännesvuosisata. *Virittäjä*, 88(4):448–448.

Erkki Lyytikäinen, Jorma Rekunen, and Jaakko Yli-Paavola. 2013. *Suomen murrekirja*. Gaudeamus.

Erkki Lyytikäinen and Jaakko Yli-Paavola. 2010. Suomen kielen nauhoitearkisto 50-vuotias. *Virittäjä*, 114(3).

Aila Mielikäinen and Marjatta Palander. 2014. *Miten suomalaiset puhuvat murteista? — kansanlingvistinen tutkimus metakielestä*. Suomalaisen Kirjallisuuden Seura.

Nancy A Niedzielski and Dennis R Preston. 2000. *Folk linguistics*. De Gruyter Mouton.

Matti Pääkkönen. 1994. *Oulun seudun murrekirja*. Suomalaisen Kirjallisuuden Seura.

Marjatta Palander. 1986. *Savonlinnan seudun murrekirja*. Suomalaisen Kirjallisuuden Seura.

Marjatta Palander. 2011. *Itä- ja eteläsuomalaisten murrekäsitykset*. Suomalaisen Kirjallisuuden Seura.

Kyubyong Park and Thomas Mulc. 2019. Css10: A collection of single speaker speech datasets for 10 languages. *Proc. Interspeech 2019*, pages 1566–1570.

Niko Partanen, Khalid Alnajjar, Mika Hämäläinen, and Jack Rueter. 2021. Linguistic change and historical periodization of old literary Finnish. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 21–27, Online. 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.

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.

Maaret Räsänen and Marjatta Palander. 2015. Kansandialektologinen testi murrepiirteiden keskinäisestä murteellisuusjärjestyksestä [a perceptual test on the hierarchy of Eastern-Finnish dialect features]. *Virittäjä*, 119(1).

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Janine Siewert, Yves Scherrer, Martijn Wieling, and Jörg Tiedemann. 2020. LSDC - a comprehensive dataset for low Saxon dialect classification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 25–35, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

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

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification. *arXiv preprint arXiv:2007.05612*.

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

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. 2021. Dialect identification through adversarial learning and knowledge distillation on romanian bert. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 113–119.