Emergent Dialectal Patterns : Analysis of regional variants in a vast corpus of Finnish spontaneous speech using a large-scale self-supervised model

Törö, Tuukka; Suni, Antti; Šimko, Juraj

2024

http://hdl.handle.net/10138/586701

Törö, T, Suni, A & Šimko, J 2024, Emergent Dialectal Patterns : Analysis of regional variants in a vast corpus of Finnish spontaneous speech using a large-scale self-supervised model. in Proceedings of Speech Prosody 2024. Speech Prosody, ISCA - International Speech Communication Association , Baixas, pp. 37-41, Speech Prosody, Leiden, Netherlands, 02/07/2024. https://doi.org/10.21437/SpeechProsody.2024-8

Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Emergent Dialectal Patterns: Analysis of regional variants in a vast corpus of Finnish spontaneous speech using a large-scale self-supervised model

Tuukka Törö, Antti Suni, Juraj Šimko

University of Helsinki, Finland
{tuukka.toro,antti.suni,juraj.simko}@helsinki.fi

Abstract

Traditional linguistic analyses, focused on morphological, syntactic and lexical features, as well as phoneme-level differences, divide Finnish into two major dialect groups, that subsequently further split into eight sub-groups. This paper presents a complementary dialectal analysis based solely on acoustic characteristics extracted from an extensive database of spontaneous speech from thousands of speakers from all Finnish dialectal areas. The distances among acoustic characteristics of speech from 17 administrative regions are approximated by prediction accuracies of binary classifiers. The classifiers are trained on principal components extracted from utterance embeddings obtained through a large-scale pretrained neural model. The clustering of regional varieties based on these distances yields geographically meaningful dialectal groupings, largely corresponding to the results of the traditional linguistic analyses. Our subsequent analysis indicates that the clustering makes use of prosodic characteristics of utterances.
Index Terms: sociophonetics, prosody, dialect identification, XLS-R, Finnish language

1. Introduction

Speakers index their social identities in multitude of ways, both by linguistic content as well as through acoustic cues such as prosody, voice quality and various phonetic features. The common traits in these characteristics of verbal communication among speakers from a certain (e.g., geographical) community, conditioned by variables such as age, gender and social class, form linguistic group identities, i.e., dialects [1, 2]. As such, dialectal distributions, based on similarities and differences among these traits, are best studied using speech material containing samples encompassing a wide range of socioeconomic, educational, regional and other aspects of social dynamics. Also, spontaneous recordings usually capture dialectal characteristics better than recordings of scripted controlled material.

In this paper we present a novel methodology for analysing a very large corpus of spontaneous, unscripted recordings in terms of geographical distribution of dialects. The analysed material, Donate Speech Corpus [3], is a vast dataset of spontaneous Finnish collected online. The corpus contains around 3,600 hours of speech (of which 1,600 hours is transcribed) by over 20,000 Finnish speakers varying in terms gender, age, region, and socio-economic class. We propose a measure of distance among regional varieties—giving rise to emergent dialectal patterns—based on accuracies of binary dialect classifiers trained on latent representations of speech material.

Outside linguistics, dialects have been a major interest in the field of speech technology, e.g., in order to improve performance of ASR systems on non-standard speech. The recent advent of large-scale models for speech representation learning capture subtle variations in speech along multiple dimensions including segmental and prosodic characteristics of audio-only material [4, 5, 6]. The latent spaces extracted from these neural network models' embeddings can be investigated to discover the acoustic underpinnings behind language variation. Foundation models such as XLS-R, pre-trained with a massive amount of audio data, can be fine-tuned with a smaller dataset for downstream tasks including language and dialect identification [7, 8].

Various self-supervised models for speech representation learning have recently been used to classify North Sami and Irish dialects [9, 10]. For Finnish, a combination of both audio and text features were used for dialect identification [11], and audio-only features for dialect levelling in the Satakunta dialect [12]. On the Donate Speech Corpus, topic identification and clustering has been investigated using audio-only material [13]. Relevant for the present work, Moisio *et al.* [3] attempted dialect classification both using transcripts and audio data as well as with audio-only data. They found that the accuracy of their classification models were relatively low in general due to the classifier never predicting some dialects but that the classification fared better with the audio-only data. They argued it suggests that the dialectal information is more salient in the acoustic signal than the linguistic content.

Finnish dialects have been extensively studied using traditional linguistic methods, and small, controlled speech datasets often limited to a specific region of the country [14]. The studies of dialectal groupings based on prosodic-acoustic characteristics are lacking. To our knowledge, this work presents the first analysis of this kind.

1.1. Finnish Dialects

Finnish language is traditionally divided into two major dialect groups: Eastern and Western dialects [15]. The division can be roughly traced back to the Nöteborg treaty of 1323 between Sweden and Novgorod which drew a border spanning in a northwest–southeast direction through what now is Finland (see Figure 3). The most distinctive phonetic feature of the division is considered to be the standard Finnish /d/ which in the eastern dialects is either not pronounced or is replaced with a semivowel, and in the western dialects is produced as /r/ or /l/ [16]. These dialect groups can be further divided into eight subgroups: Southwest, Transitional (between Southwest and Häme dialects), Häme, Etelä-Pohjanmaa (South Ostrobothnia), Keski- ja Pohjois-Pohjanmaa (Central and North Ostrobothnia), Far North, Savo (Savonia), and Southeast [17]. Studies about dialect attitudes have shown that respondents consider Finnish dialectal differences to include prosodic features such as rhythm and intonation [18].

10.21437/SpeechProsody.2024-8

2. Methods

2.1. Data Processing

We use the Donate Speech Corpus [3] of colloquial Finnish. The corpus contains around 1,600 hours of transcribed speech from over 20,000 Finnish speakers and includes metadata such as gender, age group, and the subjective judgement of the speaker's own dialect, based on one of 18 administrative regions of Finland[1].

The dialect classification presented here is based on *speaker-level representations* calculated using embeddings from the dialect identification model `voxlingua107-xls-r-300m-wav2vec` [8]. This model is trained on Voxlingua107 corpus using weights of the pre-trained XLS-R 300M model [19]. XLS-R itself is a large-scale self-supervised speech model based on wav2vec 2.0, trained with 436k hours of speech from 128 languages including 13,981 hours of Finnish. The Voxlingua107 corpus contains speech extracted from YouTube videos and includes 33 hours of Finnish. The voxlingua107 model is fine-tuned as a language identification model, and also outputs 2,048 dimensional latent embedding vectors of input utterances. These embedding vectors were used in this work as representations of utterances from the Donate Speech Corpus.

The speaker-level representations were then calculated as *means* of the `voxlingua107` embeddings of all utterance chunks (see below) by the given speaker. The original audio files from the Donate Speech Corpus are of varying length ranging from few seconds to several minutes long submissions. In order to have more uniform dataset in terms of utterance duration we split the recordings using provided alignments as follows: First, all recordings we split at silences longer than two seconds, leaving up to half a second silence at the edges. Then, the resulting chunks longer than 6 s were further iteratively split at silence intervals further than 6 s from the beginning. Resulting audio files longer than 10 seconds were discarded. Due to possible misalignments in Textgrid files, we took out audio files without silence at the beginning and end. Finally, the audio files were loudness normalized to -23 LUFS.

We balanced the dataset so that all administrative regions (to be subsequently clustered in terms of dialectal groupings) contain the same number of speakers, and the training material has a balanced gender distribution. Two, predominantly Swedish speaking, neighboring regions (Keski-Pohjanmaa and Pohjanmaa) both contained a small number of speakers and were combined into one overall Pohjanmaa unit. We were then left with 17 regions each consisting of 288 speakers in the training material, plus 33 speakers in the test set for a total of 5,457 speakers. The training material was ramdomly split into a training set (230 speakers for each region, approx. 80 %) and validation set (the remaining approx. 20 %).

2.2. Classification

In this article, we investigate how accuracies derived from classifiers can be used to cluster dialects based on the similarity of their acoustic characteristics, and what kind of acoustic cues the classifiers use for extracting distinctions between dialects.

Broadly following the approaches of [9, 3], we trained a single 17-way feed-forward DNN classifier with an input layer of 256 and three hidden layers (128, 64, 32) to classify the speaker embeddings from the regions [9]. Instead of the 2048-dimensional embeddings, their 256 first PCA components (PCs) were used as an input. Similarly, as reported by [3], the classifier never predicted some of the dialects, rendering this approach unusable for creating a distance metric.

Therefore, we subsequently trained binary classifiers (of the same architecture as above), one for each region pair. The input data for these classifiers were the first PCs calculated for the training data for the given region pair. The networks were trained for 50 epochs with learning rate of 1e-4 using the Adam optimizer, a cross-entropy loss function, and a dropout after the first hidden layer with probability 0.1. The most accurate model of the 50 epochs on the validation set was used to predict dialects on the test set.

We thus trained 272 DNN models, one for every possible pair of regions for binary classification, and used the obtained prediction accuracies on the test set as an acoustic based distance metric (see Figure 1). The lower classification accuracy is assumed to indicate a greater similarity between the acoustic representations used for training; the accuracy in the region of 50 % (the baseline for binary classifiers) means that the varieties cannot be distinguished using the present method.

2.3. Acoustic analysis

To analyse how the classification results depend on the acoustic signal-based characteristics, we extracted several simple utterance-level acoustic measures from the audio files using the Parselmouth Python library [20], and calculated the means for every speaker. The following measures were used in the analysis:

- f_0-MEAN and f_0-STD (standard deviation)
- Spectral TILT
- Speaking RATE
- Formants F1, F2 and F3

Intensity related measures were omitted due to the loudness normalization of the speech material.

The f_0 features were measured within range of 75 to 450 Hz, in semitones with pitch floor set to 75 Hz. Speaking RATE was approximated by dividing the length of transcription in characters by the duration of the utterance, and Spectral TILT by computing the slope of the Long-Term Average Spectrum (between 1000 and 4000 Hz) of the utterance. Formant frequencies were estimated with the default settings of Praat.

While f_0, spectral tilt and speaking rate measures depict standard global (utterance level) prosodic features of the speech material (intonation, tempo and voice quality), formants are primarily related to segmental articulatory characteristics (such as average vowel frontness and openness).

3. Results

3.1. Dialect clustering

Figure 1 depicts the pairwise classifier accuracies for region pairs. The heatmap indicates possible clusters with relatively low pairwise accuracies (i.e., greater similarity) among the data from some regions (see the light areas in top-left and bottow right portions of the figure).

Subsequently, we used the accuracy-based distance measure for hierarchical agglomerative clustering (HAC; using [21]) of the regions. The dendrogram in Figure 2 shows the results of the HAC clustering. Given the hierarchical nature

[1] Data from Swedish speaking Åland as well as all non-native speakers of Finnish from other areas were excluded.

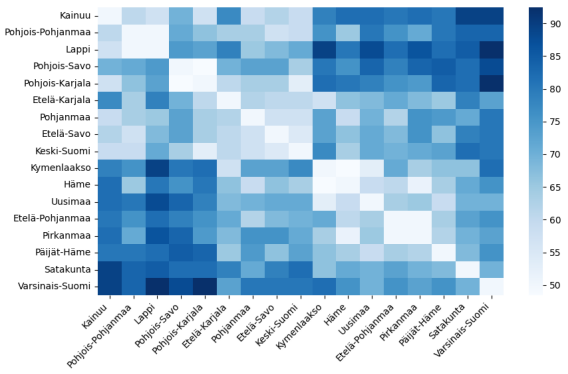

Figure 1: *Classifier accuracies for every region pair.*

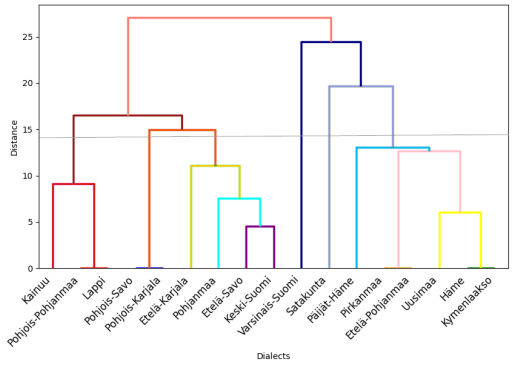

Figure 2: *A dendrogram of agglomerative clustering with average linkage based on classifier accuracy. The grey line denotes 6 clusters.*

of the clustering algorithm, multiple levels of clusters can be formed.

The primary division splits the regions in Northeast–Southwest dimension. Geographically, this division is shown in Figure 3; the regions from the left branch of the dendrogram are depicted in warm colours and the ones from the right branch in cold hues. Subsequent splits follow the same Northeast–Southwest pattern. The coloring in Figure 3 is based on a 6-way split indicated by the horizontal line in Figure 2. Subsequent splitting would in fact still make geographical sense but will start producing clusters containing a single region.

Figure 4 visualises the distances among regions in a two dimensional space using the metric multidimensional scaling (MDS) method. MDS is used to represent distances between points in high-dimensional data in a lower dimension, reflecting the original distances as much as possible. While the MDS does not indicate clear clusters, the relationships between dialects reflect the dendrogram and follow the regions' geographical positions in a remarkable fashion (cf. Figure 3).

A direct comparison between the clusters and traditional studies' dialect groups is not possible, as the borders of the 17 administrative regions differ from historical dialect regions. For example, a part of Satakunta is in the Transitional dialects and another in the Southwest dialects. The lower we look in the dendrogram hierarchy, the more differences there are with the traditional studies, suggesting that there are acoustic distinc-

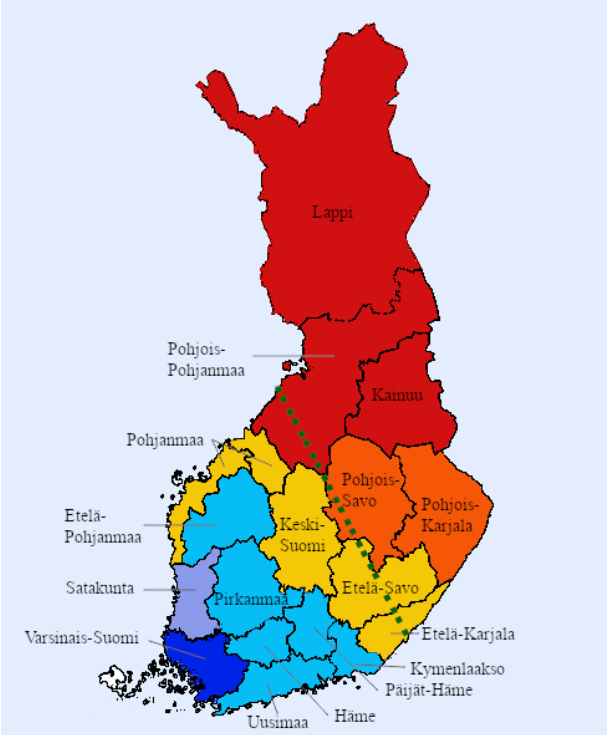

Figure 3: *Finnish dialect regions clustered. Red (warm) versus blue (cold) hues represent the 2-way clustering, different shades represent the 6-way clustering. The dotted green line denotes an approximation of the Nöteborg treaty border.*

tions that are not explained with linguistic and phoneme level variation. Still, the clusters are geographically meaningful and largely follow the overarching divisions of conventional wisdom about dialects. Figure 3 shows a map [22] of Finland with the 6 clusters in different hues and the east-west division in blue versus red.

3.2. Linear regression

In order to evaluate the relevance of the speaker-level prosodic-acoustic characteristics on the classification (see Section 2.3), we fitted linear models with the prosodic-acoustic features as dependent variables and dimensions of latent acoustic representations as predictors. The better the fit (i.e., the higher the adjusted R-squared value), the more faithfully the feature is represented in the given latent space. The separate models were fitted for each binary classifier (2 x 272 models in total).

Two types of latent representations were used in the analysis: the 256 PCs of the `voxlingua107` embeddings that served as inputs to the binary classifiers, and the 32-dimensional last hidden layer outputs of the classifiers. Table 1 summarises the adjusted R-squared values of the fits.

The 256 `voxlingua107` PCs strongly predict most of the analysed features; this means that the features are, unsurprisingly, represented in the embeddings and provide a significant source of variation among the speakers. In case of the classifiers' hidden layer, the fits are considerably worse, as indicated by lower means of the R-square values. This suggests that the classifiers use the features (encoded in the input) to a lesser degree, and they do not overwhelmingly rely on uncontrolled biases in the analysed material. The relatively high standard de-

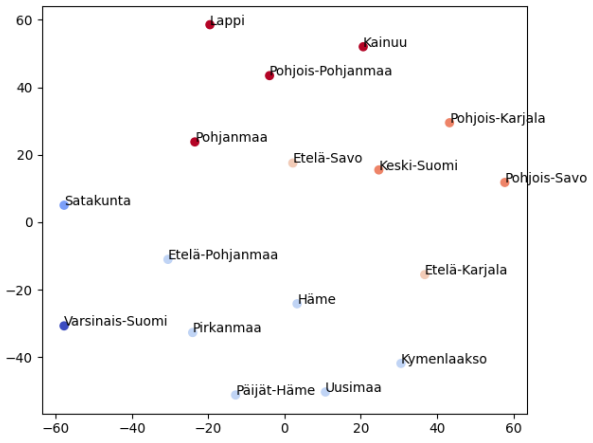

Figure 4: *Multidimensional scaling to two dimensions using a distance matrix based on the classification accuracies.*

viations for the hidden layer compared to the input layer in Table 1 indicate that the features might be more relevant for some region pairs than for others.

Table 1: *Mean and standard deviation of adjusted R-squared values of linear regression fits over all dialect pairs on the training set where the acoustic features are explained.*

Acoustic measure	**256-dim input** mean (*std*)	**32-dim hidden** mean (*std*)
Speaking rate	0.90 (*0.01*)	0.23 (*0.06*)
f0 mean	0.89 (*0.02*)	0.29 (*0.07*)
f0 SD	0.73 (*0.03*)	0.13 (*0.06*)
Spectral tilt	0.89 (*0.01*)	0.13 (*0.06*)
F1	0.81 (*0.01*)	0.11 (*0.05*)
F2	0.84 (*0.01*)	0.14 (*0.05*)
F3	0.81 (*0.02*)	0.13 (*0.05*)

4. Discussion

The presented methodology using pairwise classification approach yields meaningful dialectal clustering of Finnish language. The differences in the standard deviations of the linear fits indicate that prosodic features are present in dialectal variation. The results of our analysis are broadly in line with standard linguistic analyses of Finnish dialects. The Northeast–Southwest dialectal split is robustly manifested in our results. The more fine-grained clustering reflects geographical proximity of the regions. Some of the clusters do not perfectly reflect the traditional dialect studies. For example, Pohjois-Savo, Pohjois-Karjala, Kainuu, Keski-Suomi and parts of Pohjois-Pohjanmaa and Päijät-Häme are traditionally considered Savonian dialectal areas, while according to our analysis these regions cluster in four dialectal sub groups (see Figure 3).

This discrepancy can be partly attributed to a mismatch between the dialectal regions and administrative regions used as basic units of analysis in our work. The geographically meaningful and gradual dialectal distribution presented in Figure 3 also suggests that our analysis of a large dataset of "wild" spoken material reflects somewhat different features of speech compared to traditional analyses. Our approach likely captures more holistic characteristics (including prosody in spontaneous, uncontrolled utterances) than previous phonetic studies primarily focused on segmental quality.

In addition to dialects, our analysis might be influenced by other sources of variance captured in the signal. While the input data for the classifiers were balanced in terms of gender, in this study we do not control for other aspects like age, income or education level. It is possible that at least some of the emergent clustering reflects the differences in demographics among the regions of Finland. This highlights the need to look at social variables, not as cleanly separable categories, but as intersecting underlying structures that condition the way we communicate.

These differences can potentially also correlate with signal characteristics of the recordings; e.g., participants from more affluent areas may have better recording equipment. Signal quality has been shown to affect dialect identification in previous studies [23]. While we have concentrated on dialectal analysis in this work, the size and the nature of the corpus will allow us to investigate these potential socioeconomic and demographic influences in the future.

While the recordings include background noise, such as music or family members speaking in the background, this comes with its advantages: the recording settings are very much natural, eliciting real spontaneous speech, and the variation in background noise and recording quality should smooth out on the dialectal level compared to datasets that are, for example, recorded from regional TV broadcasts with overarching noise and channel characteristics.

In order to identify the sources of dialectal clustering, we will need to develop a way to associate the embeddings with the acoustic and prosodic features. This is challenging due to the black box nature of the speech model, i.e., the way the acoustic-prosodic characteristcs are encoded in the embeddings. Also, acoustic-prosodic measurements extractable from a large and varied dataset (such as used here) are by necessity quite simple and may perform poorly on unscripted spontaneous speech. One way to tackle this, is to simplify the data, either by selecting more controlled data for the test set or by creating synthesized stimuli. Neural text-to-speech synthesizers can be trained with the speech model's embeddings along with the corresponding audio signal. It is possible to control the synthesizer output with dialectally representative embeddings from our dataset. Synthesizing identical sentences for every dialect, we could control variation stemming from the text, and create stimuli more suitable for the acoustic measurements and perception experiments.

If we manage to close the gap between the embeddings and their acoustic correlates, these methodologies could prove useful for studying complex and subtle changes in speech.

5. Acknowledgements

This study was funded by the Academy of Finland project: *Predictive Processing Approach to Modelling Prosodic Hierarchy for Speech Synthesis*.

6. References

[1] W. Wolfram, "Dialect in society," *The handbook of sociolinguistics*, pp. 107–126, 2017.

[2] A. Etman and A. L. Beex, "Language and dialect identification: A survey," in *2015 SAI intelligent systems conference (IntelliSys)*. IEEE, 2015, pp. 220–231.

[3] A. Moisio, D. Porjazovski, A. Rouhe, Y. Getman, A. Virkkunen, R. AlGhezi, M. Lennes, T. Grósz, K. Lindén, and M. Kurimo, "Lahjoita puhetta: a large-scale corpus of spoken finnish with

some benchmarks," *Language Resources and Evaluation*, vol. 57, no. 3, pp. 1295–1327, 2023.

[4] M. Bartelds, W. de Vries, F. Sanal, C. Richter, M. Liberman, and M. Wieling, "Neural representations for modeling variation in speech," *Journal of Phonetics*, vol. 92, p. 101137, 2022.

[5] P. Sullivan, A. Elmadany, and M. Abdul-Mageed, "On the robustness of arabic speech dialect identification," *arXiv preprint arXiv:2306.03789*, 2023.

[6] M. M. Shaik, D. Klakow, and B. M. Abdullah, "Self-supervised adaptive pre-training of multilingual speech models for language and dialect identification," *arXiv preprint arXiv:2312.07338*, 2023.

[7] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[8] J. Valk and T. Alumäe, "VoxLingua107: a dataset for spoken language recognition," in *Proc. IEEE SLT Workshop*, 2021.

[9] S. Kakouros and K. Hiovain-Asikainen, "North s\'{a} mi dialect identification with self-supervised speech models," *arXiv preprint arXiv:2305.11864*, 2023.

[10] L. Lonergan, M. Qian, N. N. Chiaráin, C. Gobl, and A. N. Chasaide, "Towards spoken dialect identification of irish," *arXiv preprint arXiv:2307.07436*, 2023.

[11] M. Hämäläinen, K. Alnajjar, N. Partanen, and J. Rueter, "Finnish dialect identification: The effect of audio and text," *arXiv preprint arXiv:2111.03800*, 2021.

[12] H. Behravan, V. Hautamaki, S. M. Siniscalchi, Siniscalchi, E. Khoury, T. Kurki, T. Kinnunen, and C.-H. Lee, "Dialect levelling in finnish: a universal speech attribute approach," in *INTERSPEECH*, 2014, pp. 2165–2169.

[13] T. Grosz, Y. Getman, R. Al-Ghezi, A. Rouhe, and M. Kurimo, "Investigating wav2vec2 context representations and the effects of fine-tuning, a case-study of a finnish model," in *Interspeech*, 2023.

[14] T. Kurki, "Kielikäsitysten mosaiikki: havaintoja puhekielen näytteistä," *Sananjalka*, vol. 60, no. 60., pp. 71–95, 2018.

[15] M. Rapola, *Johdatus suomen murteisiin*. Finnish Literature Society (SKS), 1969.

[16] Institute of the Languages of Finland (Kotus). (n.d.) Suomen murteet. Accessed on [Dec 14, 2023]. [Online]. Available: https://www.kotus.fi/kielitieto/murteet/suomen_murteet

[17] E. Lyytikäinen, J. Rekunen, and J. Yli-Paavola, *Suomen murrekirja*. Helsinki: Gaudeamus, 2013.

[18] A. Mielikäinen and M. Palander, "Suomalaisten murreasenteista," *Sananjalka*, vol. 44, no. 1, pp. 86–109, 2002.

[19] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[20] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[21] 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, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] S. Finland. (n.d.) Statistics finland. Retrieved December 5, 2023. [Online]. Available: https://www.stat.fi/org/avoindata/paikkatietoaineistot/

[23] H. Boril, A. Sangwan, and J. H. Hansen, "Arabic dialect identification-'is the secret in the silence?'and other observations." in *INTERSPEECH*, 2012, pp. 30–33.