



# Automatic Dialectal Transcription: An Evaluation on Finnish and Norwegian

Olli Kuparinen

Faculty of Information Technology and Communication Sciences, Tampere University, Finland

olli.kuparinen@tuni.fi

## Abstract

The fields of dialectology and sociolinguistics are highly reliant on phonetically transcribed spoken language data, which are often written in language-specific styles and alphabets. Transcribing dialectal speech phonetically is time-consuming and thus a major bottleneck for data collection in variational linguistics. Meanwhile, the field of automatic speech recognition (ASR) has taken leaps forward in recent years.

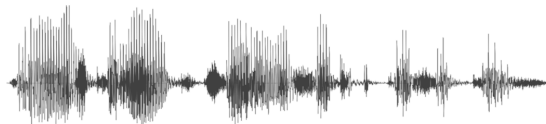
In this work, we introduce automatic dialectal transcription as a distinct ASR task and investigate solutions using data from two unrelated languages (Finnish and Norwegian) with two levels of transcription precision. We find a large performance gap between the languages, as Finnish models are much more efficient than those for Norwegian. We further evaluate the character-level errors of both languages and dialectal difficulties of the best Finnish model. We find that the most geographically central dialects tend to be easier to transcribe than the more distant ones.

**Index Terms:** speech recognition, dialect, transcription

## 1. Introduction

Dialectal transcription is an important part of studies in the fields of dialectology and sociolinguistics. The spoken audio recordings, typically collected in interviews or informal conversations, are transcribed to closely represent the recorded speech. This entails transcribing features such as false starts, hesitations, filler words, and pauses, as well as phonological and morphological features typically not written in the standard language. While the International Phonetic Alphabet (IPA) is sometimes used in dialectal transcription, it is also normal to employ transcription styles and alphabets that are specific to different languages or language families. Moreover, word boundaries are typically marked based on the language being transcribed. The task thus includes elements from standard language transcription, phoneme recognition [1], and rich (or verbatim) transcribing [2]. Figure 1 highlights the differences between phoneme recognition, dialectal transcription, and standard language text.

Since dialectal transcriptions are often created by professional linguists and their production is time-consuming, automating the process would ease and accelerate data collection massively. Readily available automatic speech recognition (ASR) tools are typically trained to output standard language (or at least the standard alphabet), which often makes them unsuitable for the task off-the-shelf. In this work, we introduce automatic dialectal transcription as a distinct ASR task. Even though earlier studies have worked on dialectal data in ASR [4, 5], the focus has been in standard language output nonetheless [6].



Phon. di mai j t e r e a k ʈ s i o : n e s e ĝ d p o s i t i v  
Transcr. di maischte reakzioone seged positiv  
Stand. Die meisten Reaktionen seien positiv

Figure 1: An example from a Swiss German corpus [3] with the audio waveform on top, followed by the phoneme presentation in IPA, dialectal transcript and a standard language version. English gloss: ‘Most reactions seem to be positive.’

As a result of the costly creation of transcriptions, the available labeled data are typically scarce. This makes automatic dialectal transcription also a low resource task, even if dealing with large languages. The inherent variation in speech, non-standard writing and special characters used in transcription also mean that additional language models are not necessarily helpful.

We evaluate the difficulties and possibilities of automatic dialectal transcription using data from two unrelated languages (Finnish and Norwegian). Both languages are known for their dialectal variation, and dialects are also often used in everyday life. Moreover, there are large collections of dialectal data available for the two languages. We use two levels of transcription precision to evaluate the difficulty of the task.

We use existing pre-trained multilingual and monolingual models for the task, and experiment with them off-the-shelf and with fine-tuning on transcribed data. We do further error analysis on the Finnish experiments to explore problematic linguistic features and dialects. The main contributions of the paper are thus to introduce the task of automatic dialectal transcription, and to evaluate self-supervised models based on transcription quality of Finnish and Norwegian dialects.<sup>1</sup>

## 2. Data

We use spoken dialectal data from Finnish and Norwegian. Both datasets are transcribed in language-specific styles and include interviews from different dialects. We segmented the original recordings based on annotated utterances and excluded the interviewers’ turns. The data were divided based on utterances: 80% of utterances in an interview were used for training, 10% for validation and 10% for testing.

<sup>1</sup>The best models and data are published at the HuggingFace Hub: <https://huggingface.co/collections/okuparinen/dialectal-transcription-fi-no-68398ac5ae9224cb1f8dd05b>.

Table 1: Examples from the two datasets with detailed transcription on top and simple transcription below. A standard language alternative and English gloss are presented in the bottom.

Transcription	SKN (fi)	LIA (no)
Detailed	mnää osasi oððal lyäð , veri nokast truiskat	ja # då va æ dær ute
Simple	mnää osasi ottal lyäd veri nokast truiskat	ja då va æ dær ute
Standard	minä osasin otsalla lyödä, veri nokasta ruiskahti	ja da var eg der ute
English	I could hit with my forehead, blood spilled from the nose	yes, I was there then

## 2.1. Samples of spoken Finnish

The SKN corpus (fi. *Suomen kielen näytteitä*) [7] contains interviews from 50 Finnish-speaking locations (99 interviews in total). The informants have been interviewed in the 1960s and chosen according to dialectological tradition: they are non-mobile, old, and rural speakers. The corpus is freely available in the Language Bank of Finland and consists of audio recordings, transcription layers in the Uralic Phonetic Alphabet and a translation to standard Finnish<sup>2</sup>. After segmenting the data and removing the interviewers’ turns, there are 82 hours of labeled data in the corpus.

We experiment with two transcription layers in this work: one including non-Finnish characters and pause markers, and another without pause markers and only characters from the Finnish alphabet (plus  $\eta$  for which there is no character pair in standard Finnish). The pause markers are comma (,) for short breaks and dot (.) for longer breaks. There are examples of detailed and simple transcriptions in Table 1.

## 2.2. LIA Norwegian

The LIA Norwegian corpus of historical dialect recordings [8] was a joint effort of four Norwegian universities to collect and annotate old dialectal interviews [9]. There are 1382 informants from 227 locations in the corpus. The audio files, phonetic transcriptions and normalization to nynorsk (one of the Norwegian standard languages) are publicly available for download<sup>3</sup>.

In addition to segmenting the data and excluding the interviewers’ turns, the Norwegian data needed further preprocessing. In the original data, references to the speakers are anonymized in the transcription. This means that the uttered speech does not match the written text. Utterances that had references to anonymized names were excluded, as were utterances that included digits instead of written out numbers. Punctuation markers and transcribers’ comments were also excluded. After these steps, the corpus includes 248 hours of labeled speech data.

As for Finnish, we use two transcription layers. The first one includes diacritics and special characters as well as pause markers, while the second one does not. The pause marker used in the LIA dataset is #. There are examples of the two transcription layers in Table 1.

## 3. Methodology

In our automatic dialectal transcription experiments, we use wav2vec 2.0 [10] based models. For both languages, we experiment with the multilingual XLS-R [11] base model which

<sup>2</sup><https://www.kielipankki.fi/download/SKN/>  
License: CC-BY 4.0.

<sup>3</sup><https://tekstlab.uio.no/LIA/filer/> License: CC-BY-NC-SA 4.0.

includes 300M parameters. The languages in question are not presented evenly in the model, as it has been trained on 13,981 hours of Finnish and only 130 hours of Norwegian speech. However, the closely related languages of Swedish and Danish are present in the training data, with almost 30,000 hours combined, which could be helpful for Norwegian as well.

Besides the multilingual model, we also use models with continued pre-training in the target (or related) language. For Finnish, we experiment with a model based on the Lahjoita puhetta (LP) -corpus [12], which includes contemporary colloquial Finnish. We utilize the large wav2vec 2.0 model trained on 2600 hours of the LP data<sup>4</sup> [13].

For Norwegian, we could not obtain models with continued pre-training in Norwegian. We instead opted to use a Swedish model that also formed the basis for a standard Norwegian ASR model in [14]. The VoxRex model includes 11,100 hours of Swedish speech, mostly from local public radio<sup>5</sup> [15].

We finetune all of the above models with our transcribed dialectal data, presented in Section 2, using connectionist temporal classification (CTC) loss. We finetuned the models for a maximum of 20 epochs on a single NVIDIA V100 GPU with the Huggingface Transformers toolkit [16]. We freeze the feature extractor before finetuning, use a learning rate of 0.0005 and a batch size of 4 with gradient accumulation of 8.

Both of the aforementioned monolingual models also have off-the-shelf finetuned versions for the target languages. The Finnish LP model has been finetuned on 1500 hours of transcribed contemporary colloquial Finnish<sup>6</sup> while the Norwegian VoxRex model has been finetuned on 150 hours of Norwegian parliamentary discussions<sup>7</sup>. We use these models as a baseline to determine how close the existing ASR models are to automatic dialectal transcription.

We evaluate the systems only on character error rate (CER). Even though word error rate (WER) is often used in ASR evaluation, we chose to exclude it from our work. As dialectal transcription aims to recognize the phonetic units and their character-level counterparts, we argue that transcribing complete words is not as relevant. This is especially true for Finnish, since linguistic features such as consonant gemination or assimilation often happen at the word boundary (*menes sinne* instead of written *mene sinne* ‘go there’). We think that missing just one phonetic element from a word (or misplacing it to the leading or trailing word) penalizes the models too greatly to offer fine-grained evaluation. This decision also corresponds to machine

<sup>4</sup><https://huggingface.co/GetmanY1/wav2vec2-large-fi-lp-cont-pt>

<sup>5</sup><https://huggingface.co/KBLab/wav2vec2-large-voxrex>

<sup>6</sup><https://huggingface.co/GetmanY1/wav2vec2-large-fi-lp-cont-pt-1500h>

<sup>7</sup><https://huggingface.co/NbAiLab/nb-wav2vec2-300m-bokmaal>

Table 2: Results of the experiments. The models are presented on the left, results for the Finnish data (SKN) in the middle and results for the Norwegian data (LIA) on the right. We present weighted average CER scores (%) for the development set, while the test results also include 95 % confidence intervals. The off-the-shelf models are only compared to the simple transcriptions as they were not trained with the special characters present in the detailed transcriptions.

	SKN (fi)		LIA (no)	
	Dev	Test	Dev	Test
<b>Off-the-shelf</b>	Weighted Av. CER %		Weighted Av. CER %	
Simple	13.83	13.69 (12.63-14.84)	46.93	46.81 (45.89-47.78)
<b>Finetuned monolingual</b>				
Simple	6.59	<b>6.51</b> (5.92-7.12)	19.54	19.56 (18.89-20.31)
Detailed	7.31	<b>7.22</b> (6.65-7.81)	20.51	20.51 (19.86-21.26)
<b>Finetuned XLS-R</b>				
Simple	9.69	9.68 (8.95-10.47)	17.37	<b>17.43</b> (16.83-18.12)
Detailed	10.93	10.91 (10.21-11.67)	17.51	<b>17.62</b> (17.03-18.28)

translation evaluation, where more lenient metrics such as chrF [17] or BLEU [18] are used. The CER scores are averaged over speakers and weighted on reference length in characters. For the test set, we also report the 95% confidence intervals [19].

## 4. Results

The results of the experiments are presented in Table 2. We notice a considerable gap in performance between the two languages, with Finnish models achieving scores below and around 10 % CER, while Norwegian models are closer to 20 % CER. We argue that this is the result of the available training data for the models, as both the off-the-shelf and the monolingual model are not trained on colloquial Norwegian data, but on parliamentary speech on the one hand and on Swedish on the other. Moreover, even though the multilingual XLS-R outperforms other Norwegian models, the amount of Norwegian in the model training is low, as discussed in Section 3. We also see only a minor difference between the detailed and simple transcriptions for Norwegian. The two styles are thus quite close to each other.

Based on our results, reaching reliable performance in automatic dialectal transcription for Norwegian would entail other techniques than basic finetuning of existing models. A possible solution would be to continue the pretraining of wav2vec 2.0 models, which has led to good results in low resource settings, for instance in [20]. We chose to abstain from this option in this work because of computational constraints, but find it intriguing for further studies.

For Finnish however, the results look more promising. The off-the-shelf model achieves reasonably good performance, even if it has not been trained for the task of automatic dialectal transcription<sup>8</sup>. Nonetheless, the finetuned multilingual XLS-R outperforms the off-the-shelf model considerably. The multilingual model is in turn even more clearly outperformed by the monolingual model, trained on the LP dataset [12].

We can thus see that the colloquial Finnish used for continued pretraining of the wav2vec 2.0 model in [13] is greatly beneficial for the automatic dialectal transcription, even though

<sup>8</sup>The  $\eta$  in the reference transcription has been transposed to  $n$  or  $ng$  accordingly for the evaluation of the off-the-shelf model, which only uses standard Finnish alphabet.

contemporary colloquial Finnish does not necessarily correspond to the old rural dialects used in the SKN data. This is however an encouraging finding that the domain of the continued pretraining data can also be somewhat different from the task data.

There is also a bigger difference between the transcription layers for Finnish, with the detailed transcription generally producing worse results. This is natural as the detailed transcription includes more characters and thus more possibilities for errors than the simple transcription. Anyhow, the difference is not large and shows that even the detailed transcription is achievable automatically.

## 5. Discussion and analysis

To analyze the results further, we take a closer look at our best models' character errors in the detailed transcription. The finetuned XLS-R model for Norwegian produced a lot of errors, given the weighted average character error rate was 17.62%. The most often mistaken character of the reference transcriptions in the Norwegian data was  $e$ , which was often predicted as  $a$  or  $i$ . These mistakes are to be expected, given the close proximity of the phonemes the characters correspond to. Often mistaken pairs were also  $o$  and  $\ddot{a}$ , as well as  $n$  and  $m$ . The incorrect predictions of the Norwegian model are thus anticipated, but their frequency is very high.

For the finetuned colloquial Finnish model the character error rate was 7.22%. The most often misrecognized characters in the model's predictions were  $i$  and  $e$ , which were often confused with each other. The character  $\ddot{a}$  was also regularly missed and substituted with  $e$  or  $a$ . All in all, the vowels tend to get misrecognized considerably more often than consonants in the Finnish predictions.

For the finetuned colloquial Finnish model, we also analyze the dialectal difficulties of the recognition. The character error rates per speaker in the test data are presented alongside Finnish dialects in Figure 2. We notice that none of the areas appear as very difficult or very easy for the model since there are big differences inside dialect areas and even inside locations. However, the dialect areas in the middle of the country (Tavastia, Southern Ostrobothnia, Savo) have the lowest CER scores on average (around 0.06). This can also be visually observed from the Figure, as there are more locations with white dots in these

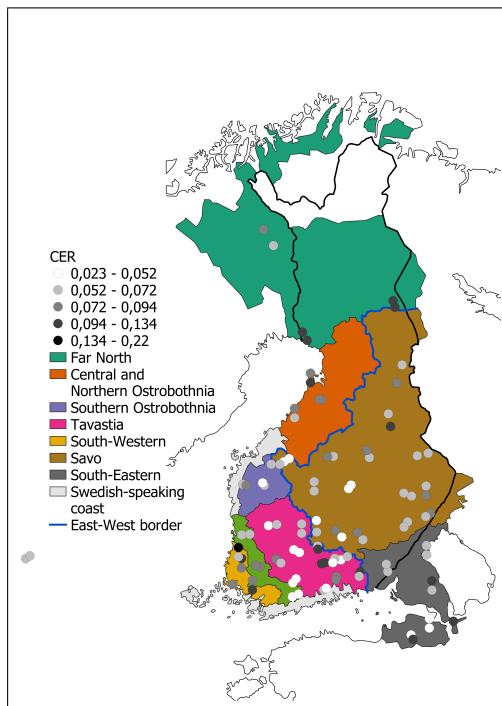


Figure 2: *The dialect areas of Finnish [21] as base colors and the CER scores of the monolingual finetuned model predictions per speaker presented as gradient black and white dots. The areas present the situation before WWII, while current dialects are spoken mostly inside Finland (borders in black). The two dots in the far West are interviewees from Värmland, Sweden. Map is made with QGIS.*

areas. All of these three dialects have widespread characteristic features (given their central geographical location), which might make them easier for the model to transcribe.

The worst CER scores appear in the South-West and in the Far North (average around 0.10). We argue that the South-Western dialects have many features which are not present in other dialects (e.g. the  $\delta$  ([ $\delta$ ] in IPA) and  $\vartheta$  [ $\theta$ ] in Table 1), making their dialectal transcription more difficult for the model. The South-Western dialects also exhibit several phoneme shortenings and lengthenings which are sometimes automatically transcribed with the wrong quantity.

The difficulties with the Far North dialect, on the other hand, do not seem to be tied to any single linguistic features. The predictions however include features from other dialects, even if they are not present in the Far North. We interpret this as slight overfitting to other dialects. A possible avenue to follow in further work would be to build dialect-specific models to avoid such issues, even though data scarcity would be a major obstacle.

## 6. Conclusion

In this paper we have presented automatic dialectal transcription as a distinct ASR task besides standard language ASR, rich (or verbatim) transcribing, and phoneme recognition. The task includes elements from all of the above. We exhibited the task with data from two unrelated languages, namely Finnish and Norwegian, which have long dialectological traditions. We used two levels of transcription precision (detailed and simple) and

experimented with existing self-supervised models based on the wav2vec 2.0 architecture. We used the models off-the-shelf and with finetuning with our own labeled data.

We found major differences between the two languages in terms of model performance and argued this to be a result of training data for the Norwegian models. We achieved reasonably good performance for Finnish models, on both levels of transcription precision. We further evaluated the character errors in both languages and dialectal differences for Finnish. We found the South-Western and Far North dialects of Finnish to be the most difficult, while the dialects that are geographically central achieved the best scores. We argued that this is a result of shared linguistic features of the central dialects.

We encourage new studies and applications in automatic dialectal transcription, which would be greatly beneficial for several linguistic fields, but especially for those that are interested in language variation (namely dialectology and sociolinguistics). We also hope to see further evaluation on other languages besides Finnish and Norwegian, which could highlight new strengths and issues of ASR applications in dialectal transcription.

## 7. Limitations

Since this work is focused on presenting the task of automatic dialectal transcription, our experimental setup can be seen as fairly basic. Moreover, computational constraints made it difficult to experiment with continued pretraining, for instance. We hope that other scholars will build further and more nuanced models on these datasets, and this task.

The setup of the task in this work might not represent actual dialectological or sociolinguistic fieldwork. We split each interview into training, validation, and test set, meaning the models had heard all speakers during training. In real-world sociolinguistics or dialectology, it would be more likely that some interviews were transcribed in full, while some interviews were not transcribed at all. This could be mirrored in this task as well, splitting the data so that some speakers (or even dialects) would be left out of training completely.

Given the page limit, we deemed it best not to include any further languages to the study, even though there are many more dialectal datasets from different languages that could have been used. Even though the languages chosen present two different language families, they are nonetheless rather large Northern European languages which use the Latin script. This is a potential bias of the work, and the task could be greatly different for languages that transcribe in a different script.

## 8. Acknowledgements

This work is supported by the Research Council of Finland through project No. 360356 “Speech as speech – acoustic modeling in variational linguistics”. The author also wishes to acknowledge CSC – IT Center for Science, Finland, for generous computational resources.

The Interspeech 2025 organisers would like to thank ISCA and the organising committees of past Interspeech conferences for their help and for kindly providing the previous version of this template.

## 9. References

- [1] C. Taguchi, Y. Sakai, P. Haghani, and D. Chiang. “Universal Automatic Phonetic Transcription into the International Phonetic Alphabet,” in *Proc. INTERSPEECH 2023*, 2023, pp. 2548–2552.

- [2] V. N. Vitale, L. Schettino, and F. Cutugno, “Rich speech signal: exploring and exploiting end-to-end automatic speech recognizers’ ability to model hesitation phenomena,” in *Interspeech 2024*, 2024, pp. 222–226.
- [3] M. Plüss, M. Hürlimann, M. Cuny, A. Stöckli, N. Kapotis, J. Hartmann, M. A. Ulasik, C. Scheller, Y. Schraner, A. Jain, J. Deriu, M. Cieliebak, and M. Vogel, “SDS-200: A Swiss German speech to Standard German text corpus,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 3250–3256. [Online]. Available: <https://aclanthology.org/2022.lrec-1.347>
- [4] A. Waheed, B. Talafha, P. Sullivan, A. Elmadany, and M. Abdul-Mageed, “VoxArabica: A robust dialect-aware Arabic speech recognition system,” in *Proceedings of ArabicNLP 2023*, H. Sawaf, S. El-Beltagy, W. Zaghouni, W. Magdy, A. Abdelali, N. Tomeh, I. Abu Farha, N. Habash, S. Khalifa, A. Keleg, H. Haddad, I. Zitouni, K. Mrini, and R. Almatham, Eds. Singapore (Hybrid): Association for Computational Linguistics, Dec. 2023, pp. 441–449. [Online]. Available: <https://aclanthology.org/2023.arabicnlp-1.38/>
- [5] E. Dolev, C. Lutz, and N. Aepli, “Does Whisper understand Swiss German? an automatic, qualitative, and human evaluation,” in *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, Y. Scherrer, T. Jauhainen, N. Ljubešić, M. Zampieri, P. Nakov, and J. Tiedemann, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 28–40. [Online]. Available: <https://aclanthology.org/2024.vardial-1.3/>
- [6] I. Nigmatulina, T. Kew, and T. Samardžić, “ASR for non-standardised languages with dialectal variation: the case of Swiss German,” in *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, M. Zampieri, P. Nakov, N. Ljubešić, J. Tiedemann, and Y. Scherrer, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL), Dec. 2020, pp. 15–24. [Online]. Available: <https://aclanthology.org/2020.vardial-1.2/>
- [7] Institute for the Languages of Finland, “Samples of Spoken Finnish, Downloadable Version,” 2014-01-01. [Online]. Available: <http://urn.fi/urn:nbn:fi:ib-2020112937>
- [8] Norwegian University of Science and Technology, University of Bergen, University of Oslo, and The Arctic University of Norway, “Lia norsk - korpus av eldre dialektopptak,” 2019. [Online]. Available: <https://www.tekstlab.uio.no/LIA/korpus.html>
- [9] K. Hagen, G. Kristoffersen, Ø. A. Vangsnes, and T. A. Áfarli, Eds., *Språk i arkiva: Ny forskning om eldre talemål frå LIA-prosjektet*. Novus forlag, Dec. 2021. [Online]. Available: <https://omp.novus.no/index.php/novus/catalog/book/19>
- [10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [11] A. Babu, C. Wang, A. Tjandra, K. Lakhota, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, “XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale,” in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.
- [12] A. Moisisio, 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.
- [13] Y. Getman, T. Grosz, and M. Kurimo, “What happens in continued pre-training? analysis of self-supervised speech models with continued pre-training for colloquial finnish asr,” in *Interspeech 2024*, 2024, pp. 5043–5047.
- [14] J. De La Rosa, R.-A. Braaten, P. Kummervold, and F. Wetjen, “Boosting Norwegian automatic speech recognition,” in *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, T. Alumäe and M. Fishel, Eds. Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 555–564. [Online]. Available: <https://aclanthology.org/2023.nodalida-1.55>
- [15] M. Malmsten, C. Haffenden, and L. Börjeson, “Hearing voices at the national library – a speech corpus and acoustic model for the swedish language,” 2022. [Online]. Available: <https://arxiv.org/abs/2205.03026>
- [16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: <https://aclanthology.org/2020.emnlp-demos.6/>
- [17] M. Popović, “chrF: character n-gram F-score for automatic MT evaluation,” in *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 392–395. [Online]. Available: <https://aclanthology.org/W15-3049>
- [18] M. Post, “A call for clarity in reporting BLEU scores,” in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: <https://aclanthology.org/W18-6319>
- [19] L. Ferrer and P. Riera, “Confidence intervals for evaluation in machine learning.” [Online]. Available: <https://github.com/luferrer/ConfidenceIntervals>
- [20] Y. Getman, T. Grosz, K. Hiiovain-Asikainen, and M. Kurimo, “Exploring adaptation techniques of large speech foundation models for low-resource asr: a case study on northern sami,” in *Interspeech 2024*, 2024, pp. 2539–2543.
- [21] T. Itkonen, *Nurmijärven murrekirja*, ser. Suomalaisen Kirjallisuuden Seuran toimituksia ; 498. Helsinki: Suomalaisen kirjallisuuden seura, 1989.