UPPSALA
UNIVERSITET

Cross-lingual and Multilingual Automatic Speech Recognition for Scandinavian Languages

Rafal Černiavski

Uppsala University
Department of Linguistics and Philology
Master Programme in Language Technology
Master's Thesis in Language Technology, 30 ECTS credits

June 23, 2022

Supervisors:
Assistant Professor Sara Stymne, Uppsala University
Carl Dehlin, Conversy

Abstract

Research into Automatic Speech Recognition (ASR), the task of transforming speech into text, remains highly relevant due to its countless applications in industry and academia. State-of-the-art ASR models are able to produce nearly perfect, sometimes referred to as human-like transcriptions; however, accurate ASR models are most often available only in high-resource languages. Furthermore, the vast majority of ASR models are monolingual, that is, only able to handle one language at a time. In this thesis, we extensively evaluate the quality of existing monolingual ASR models for Swedish, Danish, and Norwegian. In addition, we search for parallels between monolingual ASR models and the cognition of foreign languages in native speakers of these languages. Lastly, we extend the Swedish monolingual model to handle all three languages.

The research conducted in this thesis project is divided into two main sections, namely monolingual and multilingual models. In the former, we analyse and compare the performance of monolingual ASR models for Scandinavian languages in monolingual and cross-lingual settings. We compare these results against the levels of mutual intelligibility of Scandinavian languages in native speakers of Swedish, Danish, and Norwegian to see whether the monolingual models favour the same languages as native speakers. We also examine the performance of the monolingual models on the regional dialects of all three languages and perform qualitative analysis of the most common errors. As for multilingual models, we expand the most accurate monolingual ASR model to handle all three languages. To do so, we explore the most suitable settings via trial models. In addition, we propose an extension to the well-established Wav2Vec 2.0-CTC architecture by incorporating a language classification component. The extension enables the usage of language models, thus boosting the overall performance of the multilingual models.

The results reported in this thesis suggest that in a cross-lingual setting, monolingual ASR models for Scandinavian languages perform better on the languages that are easier to comprehend for native speakers. Furthermore, the addition of a statistical language model boosts the performance of ASR models in monolingual, cross-lingual, and multilingual settings. ASR models appear to favour certain regional dialects, though the gap narrows in a multilingual setting. Contrary to our expectations, our multilingual model performs comparably with the monolingual Swedish ASR models and outperforms the Danish and Norwegian models.

The multilingual architecture proposed in this thesis project is fairly simple yet effective. With greater computational resources at hand, further extensions offered in the conclusions might improve the models further.

Contents

Acknowledgments

I would like to express my deepest gratitude to both of my supervisors, Sara Stymne and Carl Dehlin, for their continuous supervision, support, as well as countless and priceless pieces of advice. I highly cherish the opportunity to work with and learn from you. I would also like to thank Conversy for creating such a pleasurable environment to write a thesis in. I am also thankful to all the members of the Department of Linguistics and Philology at Uppsala University whom I had the opportunity to meet and learn from. This would have been much more difficult, if possible, without the moral support from family and friends. Lastly, didžiulis ačiū to Viktorija and Eva Elžbieta.

1. Introduction

Automatic Speech Recognition (ASR) is the task of transforming speech into text. The task involves processing input in the form of auditory data, which is ultimately converted into and outputted as textual data. Modern ASR systems power convenient human-computer interaction through voice input, offer a cheap alternative for tedious and costly transcription work, and make it possible to retrieve information from spoken data (Furui, 1999). In Natural Language Processing (NLP), ASR is not only a major field in itself but also a tool used to conduct research into spoken language. For instance, textual representations of auditory data enable named entity recognition (NER) in speech (Caubrière et al., 2020). In addition, it has been argued that ASR has the potential to become an invaluable tool in language teaching (Carrier, 2017), with a recent study revealing that students who use ASR-based tools to learn English as a foreign language tend to acquire a more extensive vocabulary, have lower speaking anxiety, and even enjoy the learning process more as compared to traditional classroom teaching (Bashori et al., 2021). The benefits of ASR systems are thus increasingly enjoyed in business, computational linguistics, education, and numerous other fields.

The success of ASR systems can be partially accredited to their outstanding accuracy. The leaderboard of the LibriSpeech benchmark (Panayotov et al., 2015a) for English ASR may serve as a vivid example. A study by Amodei et al. (2016) estimates that the word error rate (WER), a standard ASR performance metric covered in Section 4.5.1, in human-produced transcriptions on the benchmark is roughly 5.83%. For comparison, DeepSpeech 2 (Amodei et al., 2016), an End-to-End ASR model, achieves a WER of just 5.33% on clean test set, though it is outperformed by humans on noisy data (13.25% versus 12.69%). Since the introduction of DeepSpeech 2, the field of ASR has progressed even further, as the current leaderboard of the benchmark contains ten models with a WER below 2%[1].

Despite the considerable accuracy in high-resource languages, ASR models are currently unavailable for the vast majority of the world's languages. Arguably the principal reason is the lack of training data to train such models. Deep Neural Networks (DNN)-based models, such as the aforementioned DeepSpeech 2, largely rely on the availability of substantial training datasets in order to reach high quality. With virtually no annotated data being available in most languages, DNN models are simply not viable. Another drawback of most ASR models is their ability to process just one language. As argued by Cohen et al. (1997), monolingual ASR models are unfeasible for bilingual communities and multi-cultural environments where languages are often used interchangeably.

In this thesis, we explore the possibility of addressing both limitations by training multilingual ASR models. We choose to examine the case of the Scandinavian languages, namely Swedish, Danish, and Norwegian due to the considerable similarities. According to Smith et al. (2018) and Wu and Dredze (2019), shared typological features and other similarities between languages have been proven to boost the quality of multilingual models. We firstly take a closer look at the quality of monolingual Scandinavian ASR models in search of parallels between the performance of

[1] Availableat:https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean; Last accessed: February 20th, 2022

ASR models in a cross-lingual setting and the cognition of Scandinavian languages in native speakers. As such, we explore the success of zero-shot transfers across monolingual ASR models for Swedish, Danish, and Norwegian to learn whether the patterns resemble the mutual intelligibility of the three languages. In addition, we compare the performance of monolingual ASR models across the dialects and perform qualitative analysis of the most common errors made by the models. We utilize the gathered insight about the performance of monolingual models to train a joint multilingual ASR model for the Scandinavian languages: a single end-to-end system able to transcribe Swedish, Danish, and Norwegian. Lastly, we evaluate the multilingual model on its ability to recognize and transcribe the three languages and their regional dialects.

1.1. Research Questions

The research questions tackled in this thesis are the following:

1. Are monolingual Wav2Vec 2.0-based ASR models transferable across Scandinavian languages?
2. Are there parallels between zero-shot transferability of monolingual ASR models and mutual intelligibility of Swedish, Danish, and Norwegian?
3. Are Wav2Vec 2.0-based ASR models biased towards one of the dialects of their target language?
4. Can a monolingual Wav2Vec 2.0-based ASR model be extended to handle all three Scandinavian languages without significant loss of quality?

1.2. Outline

We firstly provide a brief overview of linguistic properties and dialect varieties of the Scandinavian languages in section 2. We then offer a theoretical background on monolingual and multilingual Automatic Speech Recognition models throughout the years in Section 3. In Section 4, we describe the methodology pursued in this thesis alongside the experimental settings, datasets, models, and evaluation metrics. We present and analyse the results in Section 5. Lastly, we offer conclusions and possible directions for future work in Section 6.

2. Scandinavian Languages

The term Scandinavia is a subject for an ongoing debate in terms of what is and what is not part of Scandinavia. As formulated by Sanders (2021), the ambiguity stems in part from the discourse in which the term is used. The geographic point of view, for instance, draws the boundaries of Scandinavia along the Scandinavian peninsula (Wabakken et al., 2001). In this regard, Scandinavia is considered to be along the modern borders of Sweden and Norway. Sociologists, historians, and anthropologists, on the other hand, often extend the term to also include Denmark (Jensen et al., 2017; Lindahl-Jacobsen et al., 2004; Robinson, 2003). Arguably the most broad is the definition used in linguistics, as it includes the countries whose main official language originate from Old Norse (Sanders, 2021). Such definition of Scandinavia generally includes Sweden, Norway, Denmark, and Iceland. Lastly, Finland, a neighbouring Nordic country, is sometimes included into the definition as well (Nygård et al., 2017). The definitions of Scandinavia, as well as Scandinavian languages, are thus ambiguous. This thesis abstains from taking a stance on the boundaries of Scandinavia. The definition of Scandinavian languages followed in this thesis is limited to Swedish, Danish, and Norwegian (Bokmål written standard) simply due to the scope of the thesis as well as limited resources in Icelandic and Faroese.

2.1. Linguistic Properties of the Scandinavian Languages

Haugen (1982) states that the Scandinavian languages as well as the region itself can be analysed as a single speech continuum. This is likely due to the fact that the development of Scandinavian languages has largely been concurrent and influenced by the historical ties between the countries of Scandinavia. As such, Swedish, Danish, and Norwegian originate from Old Norse and belong to the North Germanic branch of the Indo-European languages. According to Harbert (2006), the similarity of the North Germanic branch to other Germanic languages is evident from features such as verb fronting, definite and indefinite articles, and the formation of perfect tense from auxiliary *have* and past participle. The similarities between Swedish, Danish, and Norwegian, however, stretch far beyond the general features of Germanic languages.

The three Scandinavian languages share numerous aspects of grammar and syntax, while phonetics and phonology are more distinctive. Examples of the similarities in the grammar of the three languages include morphologically marked definiteness (with the definite article being attached to the root) (Skrzypek, 2009) and the absence of a morphological case system in nouns, as they usually only appear in either nominative or genitive (Cinque and Kayne, 2005). In terms of syntax, the three languages follow the verb-second rule, which allows for no more than one phrase to precede the main verb (Cinque and Kayne, 2005). In addition, a general Subject-Verb-Object (SVO) structure of clauses is the most common across all three languages (Van Riemsdijk, 2011). The differences between the phonology and phonetics of the Scandinavian languages are often viewed as most significant. This is mostly due to the fact that the pronunciation in Danish differs greatly from Swedish and Norwegian, since, according to Grønnum (1998), spoken Danish has undergone rapid change, which brought reduction of post-tonic syllables (Grønnum, 1998), less distinctive tonal accent as compared to Swedish and Norwegian (Riad, 2006), and replacement of plosives with approximants (Basbøll, 2005). The Scandinavian languages thus share numerous similarities in terms of linguistic properties that are most evident from written language.

2.2. Dialects of Scandinavia

The view of Scandinavian languages as a single speech continuum implies a lack of explicit language borders separating the Scandinavian languages, as the three languages, as well as the societies, share roots and history. Nevertheless, differences can be observed not only across the languages, but also within them. Sahlgren et al. (2021) argue that the divergence between some dialects of Swedish, Danish, and Norwegian are arguably as major as the ones between the languages. For a better overview, the following sections briefly describe the dialects of Swedish, Norwegian, and Danish.

2.2.1. Swedish

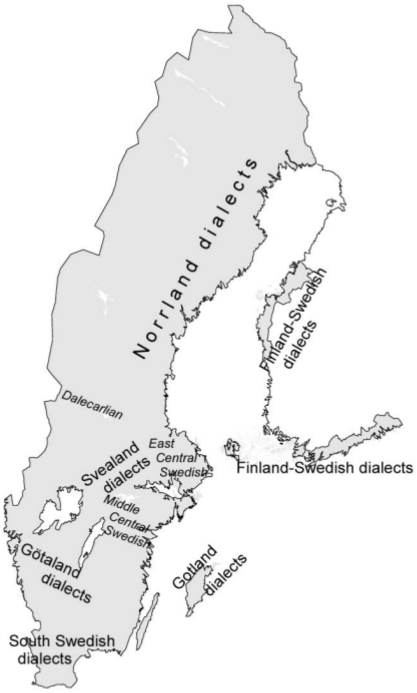


Figure 2.1.: The dialects of Swedish in Sweden according to Wessén (1954), adapted from Leinonen (2010)

The classification of the Swedish dialects proposed by Wessén (1954) distinguishes six groups: Gotland dialects, Norrland dialects, Svealand dialects, Götaland dialects, South Swedish dialects, and Finland-Swedish dialects. An approximate representation of regions where these and other dialects are spoken in Sweden is shown in Figure 2.1.

As argued by Leinonen (2010), the division is based on sentence intonation, vowel pronunciation, and distinct pronunciation of consonants. Arguably starkest are the distinctions between South Swedish dialects and East-Central Swedish of the Svealand group, which is often considered to be the standard of spoken Swedish. Schötz and Bruce (2009) observes that the sentence intonation in standard Swedish is phrase-initial and rising, whereas the intonation in South Swedish dialects is varied. Elert (2000) also reports that the long vowels in standard Swedish are instead pronounced as rising diphthongs in South Swedish dialects. Lastly, apical /r/ is common in most Swedish dialects, except for South Swedish dialects, where dorsal /r/ is used instead.

2.2.2. Norwegian

The Norwegian language and its dialects are arguably the most diverse out of the three Scandinavian languages. Influenced by the political and socio-economic situation in Norway, the Norwegian language has two written norms, Nynorsk and Bokmål, which are sometimes recognized as different languages (Kristoffersen, 2000a). According to Kristoffersen (2000b), a traditional division of the Norwegian dialects differentiates between four varieties: Nordnosk (North Norwegian), Trøndersk (Trønder Norwegian), Austnorsk (East Norwegian), and Vestnorsk (West Norwegian). Neither dialect is the official standard; however, Austnorsk (also referred to as Urban East Norwegian) and more specifically Bokmål are considered to be the unofficial national standard, as they are spoken by the majority (Kristoffersen, 2000a). The mapping of the dialects across the regions of Norway is visualized in Figure 2.2.

Johannessen et al. (2020) name retroflex flap as well as tonal accent patterns and syllabic melody as some of the most distinctive differences between the Norwegian dialects. Kristoffersen (2000a) notes that retroflex flaps are common in Austnorsk, yet generally avoided in other dialects. The author also reports that a distinction can be made between Trøndersk with Nordnosk and the other two dialect groups in terms of the syllabic melody in the pronunciation. In Trøndersk and Nordnosk, words with two syllables are often pronounced with the melody and tone common to mono-syllabic words, which is generally not the case in Austnorsk and Vestnorsk; Kristoffersen (2000a) believes this to be caused by the dropping of vowels at the end of such words, which is also common in Trøndersk and Nordnorsk.

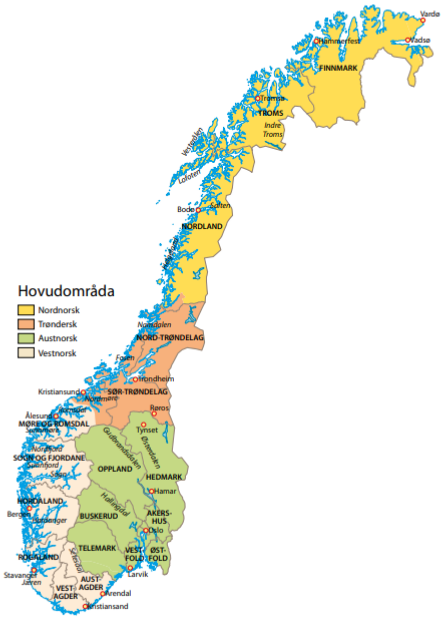


Figure 2.2.: The dialects of the Norwegian language in Norway according to Mæhlum and Røyneland (2012)

2.2.3. Danish

In contrast to Swedish and Norwegian, modern Danish is highly homogeneous. According to Basbøll (2005, p. 13), Danish is one of the most standardized languages, and the local dialects are experiencing extinction. As argued by the author, standard Danish is spoken throughout Denmark; nevertheless, regions tend to have local features that make the speech distinctive. As such, *stød* distinguishes six regional varieties of Danish corresponding to six towns and cities: Copehnagen, Næstved, Bornholm, Aalborg, Tønder, Sønderborg. Approximate locations of the regions are visualized in Figure 5.2.

Differences across the regional standards often include presence or lack of *stød* as well as divergent prosodic stress patterns. Grønnum et al. (2013, p. 67) defines stød as ‘kind of creaky voice i.e. non-modal voice with aperiodic vibrations and irregular amplitude’. The phenomenon is often said to highly resemble glottal stop. In Figure 5.2, the regions above the line, namely Aalborg, Tønder, Næstved, and Copenhagen, use stød, whereas Sønderborg and Bornholm do not. In addition, Grønnum et al. (2013) reveals divergent tone patterns following stressed vowels; prosodic stress most often raises after a stressed vowel in the Copenhagen standard, drops and then raises resembling a V-shape in the Bornholm variant, and steadily drops in other regional standards.

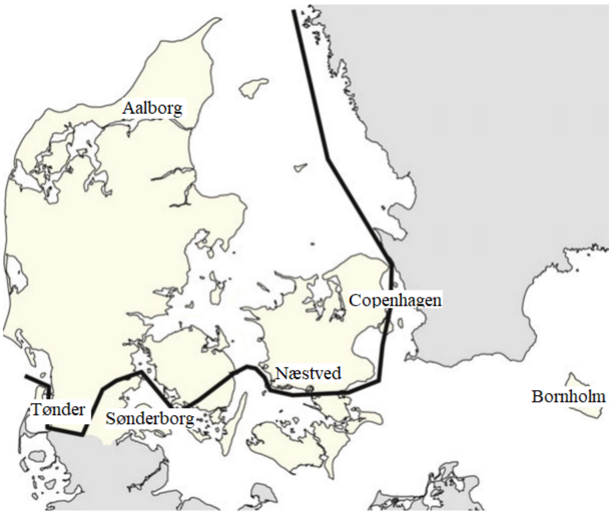


Figure 2.3.: Approximate mapping of the regional standards of Danish in Denmark. Adapted from Clausen and Kristensen (2015)

2.3. Mutual Intelligibility

Due to the close historical ties and numerous shared features, the three languages are often considered to be some of the most vivid examples of mutually intelligible languages. The mutual intelligibility though differs both across and within the languages. According to Grønnum (1998), the Scandinavian languages are highly mutually intelligible in the written form; however, when it comes to spoken language, speakers of Swedish and Norwegian may struggle to understand Danish. The results of the study conducted by Gooskens (2007) confirm the hypothesis, as the level of comprehension between Danish and Swedish was found to be not mutual. Danes were able to answer over 50% of the questions asked in Swedish, whereas

Swedes managed to answer only slightly over 24% of questions in Danish on average (Gooskens, 2007). For comparison, Danes reportedly answered roughly 57% questions in Norwegian, whereas Swedes managed to answer more than 82% asked in Norwegian. Lastly, Norwegians scored over 75% on Danish and roughly 89% on Swedish. The mutual intelligibility of spoken language across the Scandinavian languages in the spoken form thus seems strongest between Swedish and Norwegian, and weakest between Swedish and Danish. Similar patterns were also previously reported by Delsing (2005).

In addition, Gooskens (2007) and Delsing (2005) argue that the comprehension levels vary across the regions of each country. Delsing (2005) reports that most notable is the variance in the comprehension of Danish across Swedes. More specifically, the study revealed that Swedes who live in Malmö, a city located in the South of Sweden right across the Öresund bridge from Denmark, understand the Danish language significantly better than Swedes living in Stockholm, the capital city of Sweden roughly in the Center and on the East coast of the country. Similarly, Delsing (2005) observes better comprehension of Swedish in Norwegians from Oslo, the capital that is relatively close to Sweden, over Bergen, which is the second-largest city located in the West of Norway. When it comes to Danes, the comprehension of Swedish and Norwegian are comparable across Copenhagen, the capital, and Aarhus, though citizens of the latter appear to understand Norwegian better than the citizens of the capital.

The three observations seem to suggest that a closer geographic placement of the city leads to better exposure to other Scandinavian languages and thus better comprehension in the citizens. Similar conclusion was made by Gooskens (2007). The author reports signs of correlation between the mutual intelligibility levels and lexical as well as phonetic distances, which are based on the geographic distances between the regions across the three countries. It however remains an open question whether the relationship between the geographic location and the mutual intelligibility is causal.

2.4. Scandinavian Languages for Multilingual Research

The Scandinavian languages are some of the most suitable candidates for multilingual research. Firstly, the languages are relatively similar in both spoken and written forms. The three languages share comparable phonesets, or sets of phonetic units used in speech, as well as identical alphabets with the exception of two letters which are the same in Danish and Norwegian yet different in Swedish. These similarities allow avoiding numerous technical issues one might face while working with languages that have different numbers of characters or even completely different scripts. Secondly, shared roots of the three languages enable inherent multilingual cues extending across Swedish, Danish, and Norwegian. The numerous similarities mentioned in Section 2.1 and most notably a largely similar lexicon make it possible for the native speakers of the three languages to be able to communicate with each other without a mediator language, such as English. With these and other arguments in mind, Sahlgren et al. (2021) make a case in favour of multilingual models suited for all Scandinavian languages (also including Icelandic and Faroese) in order to use computational resources more efficiently, to ensure transparency, and equal access to powerful models across the region.

3. Theoretical Background

This section is divided into six parts. It firstly describes the research area of Automatic Speech Recognition (ASR) in Section 3.1, followed by an overview of statistical (3.1.1) versus neural ASR models (3.1.2). Section 3.2 addresses the architecture and framework of Wav2Vec 2.0. Section 3.3 offers a summary of the domain, accent, and language-based research on Wav2Vec 2.0. Lastly, Section 3.4 provides an overview of research focused on multilingual ASR.

3.1. Automatic Speech Recognition

Automatic Speech Recognition (ASR) is a complex multimodal task that facilitates signal processing and language modelling. The goal of the task is to transcribe; that is, to transform speech into text. Due to the lack of inherent connection between the two modalities, traditional ASR models rely on extensive external knowledge about signal processing, language modelling, and the parallels connecting them. As such, the sources of knowledge often include lexicons of existing words in the language, a phoneset or a set of phonemes, a language model, and possibly numerous other tools, databases, and rules (Povey et al., 2011).

In most simple terms, an ASR tool receives an input in the form of an audio signal, usually, an audio file, which it encodes and extracts the auditory features from; afterwards, an ASR model maps the auditory features to either characters, morphemes, or words, thus resulting in a transcription of the audio. In traditional ASR systems, this is achieved by passing the input through a handful of modules that are responsible for discrete aspects of signal processing or language modelling. Benesty et al. (2008) list four components that are common in traditional ASR systems, namely a feature extraction module, an acoustic model, a language model, and a hypothesis search module. The main purpose of the feature extraction module is to transform the audio signal from a file-specific encoding (e.g. wav) to discrete numeric representations; however, as noted by Benesty et al. (2008), the module can also remove noise and channel distortion among other tasks. The acoustic model then uses the encoded numeric representations of auditory signals to deduce the phonemes or other phonetic units that make up the encoded speech. In other words, it transforms the encoded numeric representations of sounds into representations of characters, morphemes, phonemes, or words. The language model component stores the knowledge about word or character co-occurrence patterns in the language that is being transcribed. Lastly, the hypothesis search module combines the predictions made by the acoustic model with the knowledge about the language from the language model and outputs the most likely text corresponding to the audio signal, which was the initial input. A visualization of the modules and their interaction can be seen in Figure 3.1 below.

3.1.1. Statistical ASR

There are numerous divergent implementations of each of the four components of traditional ASR models. In the traditional view of ASR, most of the components employ statistical methods. Sen et al. (2018) name Linear Prediction Coding (LPC),

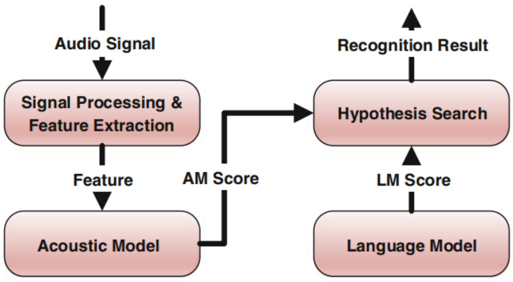


Figure 3.1.: Common architecture of traditional ASR models (Benesty et al., 2008)

Mel-Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP), and Discrete Wavelet Transform (DWT) as some of the numerous techniques used in the feature extraction step. The four methods combine the theory of auditory cognition and mathematics resulting in an effective sound encoding and discretization. The variety of techniques can in part be explained by the diversity of applications of the feature extraction module, as, in addition to the transformation of audio into numeric representations, it can be used to, for instance, extract phonetic features of the speaker. Such features are mainly used in speech classification or speaker identification tasks; however, as noted by Sen et al. (2018), the aforementioned features are key components of the raw audio and thus need to be accurately processed in the module to ensure high-quality discrete numeric representation. This is relevant to the task because the pronunciation of a word *speech* in low and high pitch should ideally receive a similar numeric representation by the feature extraction module. When it comes to the acoustic model component, Benesty et al. (2008) refer to Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) as some of the most conventional solutions in traditional ASR. The acoustic features encoded by the feature extraction module most commonly constitute a large set of numeric representations of individual sounds much shorter than most phonemes. According to Benesty et al. (2008), a combination of GMMs and HMMs offers a probabilistic solution for such problem, as the former can be said to accurately classify individual sounds, while the latter groups them into phonemes or potentially longer sequences. As for the language model component, the canonical probabilistic approach employs n-grams, a frequency-based co-occurrence rate of linguistic units, most often words. As argued by Habeeb et al. (2021), n-grams are especially effective in eradicating grammatical errors and choosing between alternative spellings, which can pose a challenge to even state-of-the-art ASR models. Lastly, the hypothesis search module of a traditional ASR model ranges from linear tree lexicons (Ortmanns and Ney, 2000) that decode each word as a unit in isolation to stack decoding (Sturtevant, 1989), which models lengthy sequences whilst potentially also evaluating partial hypotheses. In traditional ASR, hypothesis search is exceptionally computationally demanding and the choice of the module implementation is thus often influenced by the access to (or lack of) computational power. Overall, the combinations of component implementations in traditional statistical ASR are numerous. The implementation of a model is commonly tailored to the availability of the resources needed to train each component as well as computational resources.

3.1.2. Neural ASR

Contrary to the distribution and probability theory used in statistical ASR, neural ASR models exploit the ability of modern technology to learn countless features as well as the connections between these features through Deep Neural Networks (DNNs) and other deep learning strategies. Similarly to the applications in other Natural Language Processing (NLP) tasks, neural ASR models tend to require copious amounts of training data to achieve competitive results (J. Li et al., 2018; Perero-Codosero et al., 2022). However, if the data is available, deep learning methods such as DNNs contribute to considerable improvements in ASR systems by enhancing individual components (Dahl et al., 2012) and enabling end-to-end models (Wang and G. Li, 2019).

A common application of deep learning strategies in ASR is an extension to the aforementioned GMM-HMM acoustic model with the addition of Recurrent Neural Networks (RNNs). As formulated by Kamath et al. (2019), Deep Neural Networks, unlike GMMs, effectively learn non-linear features. Tanaka et al. (2019) add that such ability enables the HMMs to then draw accurate parallels across frame-level context-dependent states. To paraphrase, a hybrid DNN and HMM acoustic model makes it possible for an ASR model to learn the cues linking the modalities of speech and text through countless frame-level patterns that add up to sequences such as phonemes and words. A somewhat more recent neural re-consideration of conventional ASR components can be observed in the language model module. There appears to be an increasing interest in replacing the reliance on n-grams with Transformer-based language models such as BERT (Lee et al., 2021; Yu and Chen, 2021). In brief, the Transformer architecture makes use of self-attention, a technique to shift the attention of the model towards more relevant elements regardless of their position in the sequence (Vaswani et al., 2017). Therefore, as opposed to the n-sized windows in n-grams, Transformer-based language models learn the contextual cues between virtually all linguistic units it encounters. In ASR, contextual embeddings prove to be exceptionally useful for transcription in a domain the model has not been trained on, for Transformer language models such as BERT operate on sub-word level and can therefore construct words even if they have never encountered them during training. The Transformer architecture has also seen applications in other components of ASR models, such as the encoder (Baevski et al., 2020) and acoustic model (Haidar et al., 2021), resulting in faster inference and lower error rates.

As argued by Kamath et al. (2019), advancements in neural ASR brought about by deep learning methods such as DNNs and Attention contributed to a shift in the architecture of ASR models towards end-to-end models. Traditional ASR models are complex and difficult to produce, as every component is mostly trained or compiled individually (Wang and G. Li, 2019). This requires considerable knowledge about the language, the data one has to work with, and the structure of the component in addition to the technical knowledge needed to build or train the component in question. End-to-end ASR models offer a convenient solution for most of these challenges. The term *end-to-end* denotes the simplified structure of neural ASR models where the transformation of speech into text is carried out without intermediate states and all the individual components of a traditional ASR model are combined in a single neural network pipeline (B. Li et al., 2020). In other words, the speech is virtually mapped directly to the corresponding text without having to firstly map speech representations to, for example, phonemes. Wang and G. Li (2019) argue that such pipeline is more straightforward for both development and use, as the neural networks learn direct relations between speech and text. B. Li et al. (2020) further elaborates that end-to-end models have less parameters overall, making the ASR systems less computationally demanding and thus more accessible on all sorts of devices.

Considering the numerous advantages of end-to-end models, it can be argued that end-to-end architecture has become the standard for modern ASR models. The architecture is common among ASR models that dominate the ASR benchmarks such as LibriSpeech (Panayotov et al., 2015b). Most notable examples are Wav2Vec (Schneider et al., 2019) and its successor Wav2Vec 2.0 (Baevski et al., 2020). Further enhanced by several extensions and training strategies such as masked language modelling (Y.-A. Chung et al., 2021) and self-training (Xu et al., 2021), the latter currently holds the five best word error rates on the LibriSpeech benchmark[1] in English. The following section provides and overview of the architecture.

3.2. Wav2Vec 2.0

In 2020, Baevski et al. (2020) proposed Wav2Vec 2.0, a novel signal processing framework that is capable of learning accurate speech representations with no supervision. A visual representation of the architecture of Wav2Vec 2.0 can be seen in Figure 3.2 below. The framework firstly uses convolutional neural networks to transform input (X) in the form of raw audio into latent speech representations (Z). The encoded speech representations are then quantized; that is, the continuous set of sounds captured in the latent space is transformed into a smaller, discrete set (Q). Quantization of speech representation can arguably be compared to the comprehension of speech units by the human brain, as the set of sounds in every language is defined with a finite set of phonetic units, such as the International Phonetic Alphabet (IPA). As reported by Baevski et al. (2020), quantization results in more accurate contextual representations (C). The context representations are learned by passing the encoded inputs to a Transformer architecture, where certain portions of the audio are masked and have to be reconstructed from a set of real quantized representation and a set of distractors, together referred to as codebooks.

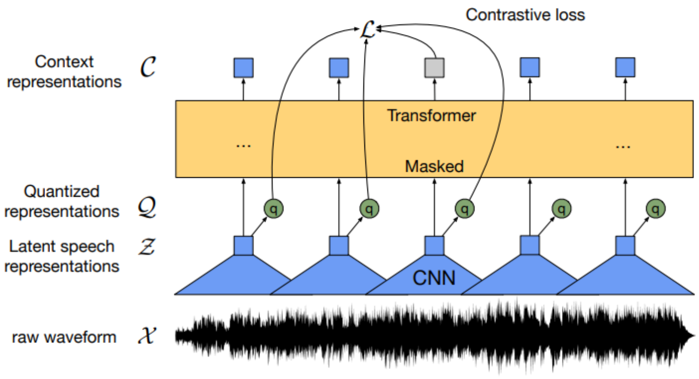

Contrastive loss
$\mathcal{L}$
Context representations $\mathcal{C}$
Transformer
...
Masked
...
Quantized representations $\mathcal{Q}$
q
Latent speech representations $\mathcal{Z}$
CNN
raw waveform $\mathcal{X}$

Figure 3.2.: The architecture of Wav2Vec 2.0 (Baevski et al., 2020)

The Wav2Vec 2.0 architecture learns extensive contextual representations of speech units by jointly minimizing contrastive loss and diversity loss. More specifically, Baevski et al. (2020) define the objective function as

$$\mathcal{L} = \mathcal{L}_m + \alpha\mathcal{L}_d \tag{3.1}$$

where $\mathcal{L}_m$ is the contrastive loss, α is diversity penalty, and $\mathcal{L}_d$ is the diversity loss. Contrastive loss can be defined as the task where the model has to recognize the

[1] https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean; Last accessed: February 20th, 2022

true quantized representation from a set of one true quantized representation and a number distractors, which are other masked segments of the same sequence. To give an example, if an entry passed to the model corresponds to the pronunciation of the phrase *colourless green ideas*, the model might mask the segments corresponding to *co*, *rl*, *re*, and *ea*, and, in turn, try to predict the right segment for *_olourless* from a set consisting of the true segment (*co*), and a set of distractors (*rl*, *re*, *ea*). More formally, the contrastive loss function can be defined as

$$\mathcal{L}_m = -\log \frac{\exp(sim(c_t, q_t)/k)}{\sum_{\tilde{q}} \exp(sim(c_t, \tilde{q})/k)} \tag{3.2}$$

where c_t is a context speech representation predicted by the Transformer, q_t is the quantized speech representation at time step t, and k is the number of distractors.

Baevski et al. (2020) introduce diversity loss to ensure that the model makes use of as many of the the learned quantized representations as possible. As such, diversity loss is defined as

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \overline{p}_{g,v} \log \overline{p}_{g,v} \tag{3.3}$$

where V is the number of entries in G codebooks, and l is the mean softmax distribution over the entries from each codebook $\overline{p}_g$.

The main objective of the aforementioned Wav2Vec 2.0 framework is to accurately encode speech signals into quantized and contextualized representations; in other words, it mainly serves as an encoder. The term Wav2Vec 2.0 is, however, often used to refer to end-to-end models that, among other things, perform ASR. This is made possible by combining a Wav2Vec 2.0 encoder with a Connectionist Temporal Classification (CTC) decoder head, which transforms the encoded representations (stored as values) into characters, resulting in text. CTC (Graves et al., 2006) can be very roughly summarized as a sequence modelling algorithm that enables mapping between sequences of different lengths. This is of utmost importance when working with two different modalities, as is the case in ASR. In most, if not all languages, the lengths of text and speech sequences are independent, as one letter does not necessarily correspond to one phoneme, the smallest phonetic unit (consider, for instance, the word *queue* and its pronunciation annotation in IPA, namely */kju:/*). Moreover, some phonemes are pronounced longer than others even if they correspond to the same grapheme, the smallest graphological unit (for instance, both */ʌ/* (short *a*) and */a:/* (long *a*) generally correspond to the grapheme *a*). CTC breaks away from the pre-condition of equal lengths by firstly predicting the token sequences (with tokens corresponding to letters), merging the repetitions, and removing the grapheme boundary markers. As formulated by Hannun (2017), when used for ASR, CTC is an algorithm that learns to infer the most likely sequence of graphemes given the encoded sequence of sounds by minimizing CTC loss (Graves et al., 2006).

CTC is an efficient complement to the Wav2Vec 2.0 framework; unsurprisingly, a successful combination of the two leads to state-of-the-art ASR performance in a variety of settings. The initial combination of Wav2Vec 2.0 encoder with CTC decoder was proposed by Baevski et al. (2020) already in the original Wav2Vec 2.0 paper. Somewhat resembling the traditional ASR models, both components require training, which is carried out individually. The authors thus proposed different names for the two training procedures; namely, they referred to the process of training the Wav2Vec 2.0 encoder as *pre-training*, whereas they use the term *fine-tuning* to refer to the training of the CTC decoder. In their initial study, Baevski et al. (2020) pre-trained the model on 960 hours[2] of unlabeled data. Afterwards, they fine-tuned the CTC decoder on annotated data of different sizes to predict characters, ultimately transforming the voice input into text. The results demonstrated competitive performance and resembled those of semi-supervised (Synnaeve et al., 2019) and fully supervised (Gulati et al., 2020) models which require extensive additional data in the form of acoustic and language models. In addition, with just 10 minutes of annotated data, the authors of Wav2Vec 2.0 reached solid performance: 9.1 and 15.6 Word Error Rate (WER) on the *clean* versus *other* subsections of the LibriSpeech benchmark (Panayotov et al., 2015a).

The excellent quality of a Wav2Vec 2.0-based model even in low-resource settings is, however, not the only reason why the framework seems to have taken over modern ASR. Arguably of equal importance is the potential of Wav2Vec 2.0 to be further extended, enhanced, and combined with other tools and models. This can be observed from the original paper as well, where Baevski et al. (2020) proposed a so-called LARGE setup for the encoder, which increases the number of Transformer blocks, attention heads, model dimensions, and inner dimensions in the encoder. The authors also demonstrated the possibility of enhancing the CTC character-level decoder with n-grams and Transformer language models. Both extensions require additional resources in the form of computing power or data; however, they lead to considerable improvements as compared to the BASE model with no extensions.

3.3. ASR across Domains, Accents/Dialects, and Languages

Wav2Vec 2.0 has drawn tremendous amount of attention from the NLP community, as research into Wav2Vec 2.0 has evidently become a direction in ASR in itself. The direction of such research is fruitful for the ASR community because it often deviates from the pursuit of new state-of-the-art and instead focuses on aspects such as more efficient usage of data as well as the feasibility of models for low-resource languages. As such, one of the arguments made by Baevski et al. (2020) in favour of Wav2Vec 2.0 is the ability to pre-train the model on unlabeled data. This has been further extended by Hsu et al. (2021), who proved that the speech features learned during the pre-training step transfer across domains relatively effectively. Nonetheless, in-domain pre-training on unlabeled data leads to consistent and substantial improvement in the performance of a Wav2Vec 2.0-based ASR model. Furthermore, Hsu et al. (2021) revealed that exposure to more domains in the pre-training step increases the robustness of the model and thus betters the performance even in an out-of-domain setting. Similarly, Turan et al. (2020) investigated the effects of exposing the model to different dialects of a language in both the pre-training and fine-tuning steps. In their experiments, the authors used a dataset with 21 dialects of English, both native and non-native, for pre-training and fine-tuning a Wav2Vec 2.0 model. As reported by Turan et al. (2020), such approach enabled the model to learn accent-invariant features, leading to a consistent improvement over all accents of English. It can therefore be argued that Wav2Vec 2.0 models benefit from the diversity in the training data, as exposure of models to other domains and accents improves the speech recognition capabilities.

Exposure of a Wav2Vec 2.0-based model to multiple languages in the pre-training step has also been proven to be an effective approach. Multilingual language models such as XLM (Conneau and Lample, 2019) and mBERT (Devlin et al., 2019) are eminently suitable for multilingual and cross-lingual NLP tasks; nevertheless, when

[2] Alternatively 860 hours in certain setups

it comes to monolingual settings in high-resource languages such as English, monolingual counterparts such as BERT generally appear to be more effective Conneau et al. (2021) and Conneau and Lample (2019). According to Conneau et al. (2021), this is not necessarily the case with ASR models. The authors pre-trained a model on a combination of 53 languages (thus the name XLSR-53) and compared it against a model trained on monolingual data only, with the same amount of training data and training time. The WER achieved by the multilingual model was at least 2 times lower on all ten languages that it was tested on, e.g. 2.9 (XLSR-53) vs 6.8 (Monolingual) WER on Spanish, 63.6 (XLSR-53) vs. 12.2 (Monolingual) WER on Swedish. Conneau et al. (2021) additionally experimented with fully multilingual model that was both pre-trained and fine-tuned on ten languages. When compared to the monolingual baselines (pre-trained and fine-tuned only in the target language), the results differ across the languages. In the case of Spanish, the fully multilingual model performed worse than the monolingual baseline and the XLSR-53 setup (9.4 versus 6.8 and 2.9 WER respectively); when it comes to Swedish, the XLSR-53 model was also observed to be most accurate (12.2 WER), yet the fully multilingual model (21.0 WER) outperformed the monolingual baseline (63.6 WER). The deviations can perhaps be linked to the differences in the sizes of data per language used for pre-training. Similarly to the conclusions drawn by Bhable and Kayte (2020), in relatively high-resource settings, e.g. Spanish with 168 hours, the fully multilingual setup likely introduced noise, whereas for low-resource languages, e.g. Swedish with just 3 hours, the monolingual data were insufficient to train an accurate monolingual model (thus a WER of 63.6), and data in other languages seemed to compensate the deficit.

3.4. Multilingual ASR

The developments in the field of Automatic Speech Recognition have continuously kindled interest in the multilingual aspect of ASR. The primary goal of multilingual ASR is being able to accurately handle two or more languages with a single system. In light of this, a canonical architecture of multilingual ASR models has a language identification component stacked on top of multiple monolingual models. In other words, a classification model firstly aims to correctly identify which language is being spoken, and then forward the speech input to the corresponding monolingual model. Such architecture has been proposed by D.-C. Lyu and R.-Y. Lyu (2008), Mabokela and Manamela (2013), and Barroso et al. (2010). Due to its simplicity and effectiveness, the architecture remains a common solution for numerous non-language-specific ASR models in the industry (Liu et al., 2021). As pointed out by D.-C. Lyu and R.-Y. Lyu (2008), the architecture also excels at accurately processing instances of code-switching, which is the phenomenon of switching between two or more languages on a phrase or sentence level as one speaks. The classification-based architecture is thus a relatively common solution for multilingual ASR in that it builds upon and overcomes some of the main limitations of the monolingual systems. However, the practicality of such architecture is questionable.

Firstly, classification-based multilingual ASR is costly in terms of computational resources. This is mainly due to the fact that multiple systems, a classifier and several speech-to-text engines, are ran in parallel. Perhaps even more challenging is the requirement of multiple accurate monolingual ASR systems, as they generally require copious amounts of training data (Hannun et al., 2014; Schultz and Waibel, 2000). Adding an additional monolingual model thus linearly increases the amount of data needed to produce such multilingual ASR system, which further increases the computational costs.

An alternative solution for multilingual ASR systems involves combining some of the components of monolingual models, while others remain detached. In traditional ASR, multilingual approaches tend to rely on modifying or extending the acoustic model across multiple languages. As such, Lin et al. (2009) merge phone models and phonetic representations across multiple languages, whereas Bhable and Kayte (2020) use Subspace Gaussian Mixture Models (SGMM) with shared parameters, which are then joined through HMMs. The results achieved by such models are mixed; the first approach appears to be only effective in low-resource setting, yet harmful for resource-rich languages such as English. The second approach, however, reportedly improved the performance on all languages involved.

Shared acoustic models have also drawn attention in the field of neural ASR. Here, the multilingual structure of a model can be achieved through shared hidden layers, as described by Benesty et al. (2008). Implementations of such architecture have been proposed by Huang et al. (2013) and Heigold et al. (2013). As shown in Figure 3.3, such architecture commonly has a shared input layer as well as hidden layers that jointly learn acoustic features. The softmax layers, which constitute the decoder of such model, are, however, individual for every language. This is done to encourage the model to use language-specific structures when outputting the transcription. Benesty et al. (2008) argues that the benefits of this architecture are numerous. Most notably, the data in foreign languages improves the model's ability to generalize and reduces its bias. Furthermore, it makes it possible for the model to deduce common cues shared across languages and thus aids the model's learning of low-resource languages. The results reported by Huang et al. (2013) seem to support these claims, for a model with shared hidden layers reduced the WER over monolingual DNN baselines by at least 3% on average across French, German, Spanish, and Italian, all of which are relatively high-resource languages.

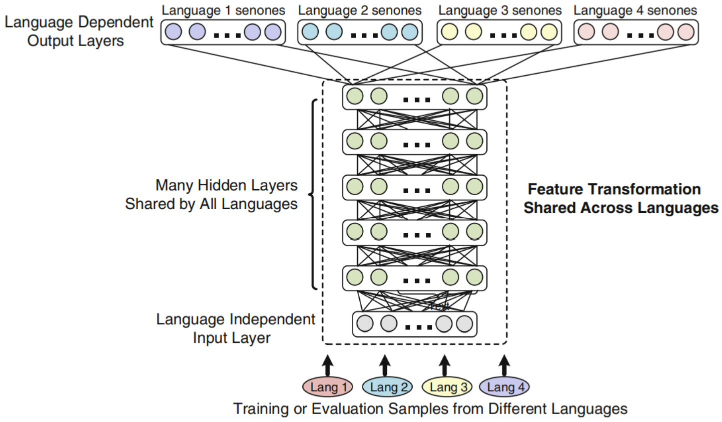

Language Dependent Output Layers
Language 1 senones
Language 2 senones
Language 3 senones
Language 4 senones
Many Hidden Layers Shared by All Languages
Feature Transformation Shared Across Languages
Language Independent Input Layer
Lang 1
Lang 2
Lang 3
Lang 4
Training or Evaluation Samples from Different Languages

Figure 3.3.: The architecture of a multilingual ASR model with shared hidden layers (Benesty et al., 2008; Huang et al., 2013)

It can be argued that most straightforward is the multilingual integration in end-to-end models. Without any additional pre-processing or multilingual resources, end-to-end models can be directly trained on multiple languages at once. An example of such approach has been proposed by Conneau et al. (2021). The authors used the aforementioned Wav2Vec 2.0 framework to jointly pre-train and fine-tune a multilingual model on 10 languages at once. Full integration did not, however, lead to best results. Most notable improvements were observed with the previously mentioned XLSR-53 approach, where only the pre-training step was carried out jointly, whereas the fine-tuning was carried out in isolation. Evidently, joint training on all ten languages introduces substantial noise due to the typological differences between the languages. Pratap et al. (2020) report somewhat similar observations with a sequence-to-sequence ASR model: the improvements in terms of the word error rate differ substantially across languages, as low-resource languages benefit the most, and high-resource languages lag behind monolingual baselines. Nevertheless, Pratap et al. (2020) introduce a mostly universally-functional solution involving input language embedding as well as script-based language clusters. The language-based embedding is a 10-dimensional vector that encodes the features of a language and is fed to the encoder along with the input. A language cluster, on the other hand, is a simple grouping of languages based on their script, which determines which languages will be jointly learned by the ASR model. As with other multilingual approaches, this benefits low-resource languages. However, the authors reported up to almost 50% increase in the word error rates for high-resource languages.

To summarize, multilingual ASR is an active research topic that has seen approaches ranging from monolingual systems joined by a classifier to fully connected multilingual end-to-end models. Following the work of Conneau et al. (2021) and Pratap et al. (2020), current multilingual ASR models almost exclusively benefit low-resource languages. In addition, fully connected multilingual models suffer from noise when handling typologically different languages that, for instance, use different scripts.

4. Methodology

In this thesis project, we firstly analyse the performance of monolingual ASR models for the Scandinavian languages. We then train and evaluate a multilingual end-to-end ASR model for Swedish, Danish, and Norwegian. The research conducted in this thesis project can be divided into two parts and sets of experiments, namely monolingual and multilingual.

This section firstly covers the experimental settings alongside the hypotheses for monolingual (4.1) and multilingual (4.2) models. Section 4.3 describes the two datasets used in this research. Section 4.4 covers the models used. Section 4.5 addresses the evaluation metrics and practices used in this project.

4.1. Monolingual Scandinavian ASR Models

As explored by Delsing (2005) and Gooskens (2007), the spoken language comprehension levels differ across the native speakers of Swedish, Danish, and Norwegian. Both authors suggest that the extent of comprehension is affected by one's native language as well as their exposure to other Scandinavian languages, or lack thereof. The first aim of this research is to explore whether similar patterns can be observed in the monolingual ASR models for Swedish, Danish, and Norwegian.

Research into comprehension of foreign languages where the subjects of study have received no education in that foreign language highly resembles zero-shot transfer setting for models. In zero-shot setting, a model trained on one language or task is used to process data in another language or perform a different task. An example of such setting is explored by Wu and Dredze (2019), who use an English part of speech tagging model to tag sequences in French.

Zero-shot transfer is especially effective for low-resource languages, for which the training data is sparse. Prior research also suggests that the success of zero-shot transfers depends on the quality of features learned from the source language or languages (Lauscher et al., 2020). On the other hand, the transferability of the learned features in zero-shot settings also depends on how similar the source and target languages are in their typologies, vocabulary overlap, and other aspects (Wu and Dredze, 2019). As covered in Section 2, the Scandinavian languages are alike in a range of aspects varying from the alphabet to grammar. We therefore seek to investigate whether these similarities enable successful cross-lingual zero-shot transfers of the monolingual ASR models. We also investigate whether differences in the performance of ASR models on Scandinavian languages correlate with the cross-lingual comprehension levels in native speakers of the Scandinavian languages.

In order to learn whether the monolingual ASR models for Swedish, Danish, and Norwegian are mutually transferable, we additionally compare their performance against an accurate English model. By doing so, we seek to ensure that the good or poor performance of the monolingual models stems from the language it was trained on rather than the overall quality of the model. In addition, we compare the performance of the models across the regional dialects of the Scandinavian languages. We seek to learn whether monolingual ASR models are biased towards certain regional dialects; that is, if they perform better on certain regional dialects than on others.

Lastly, we evaluate the effectiveness of complementing monolingual ASR models with statistical language models across all regional dialects and in a cross-lingual setting. 4-gram and 5-gram language models have been proven to consistently boost the performance of ASR models in monolingual settings (Baevski et al., 2020; Tian et al., 2022). It has also been studied from the perspective of in-domain and out-of-domain language model applicability (Håkansson and Hoogendijk, 2020). To the best of our knowledge, the applicability of n-grams language models for cross-lingual ASR is yet to be studied. We assume that n-grams, and likely other types of language models, are highly effective in cross-lingual settings because they offer knowledge about the target language, which the ASR models are not exposed to.

4.1.1. Hypotheses

We hypothesize that the monolingual Wav2Vec 2.0-based ASR models for the Scandinavian languages partially exhibit patterns similar to native speakers in terms of the comprehension of other Scandinavian languages. In addition, we expect to find bias in the monolingual ASR models towards one regional dialect. We also believe that the language-specific supervision offered by a 4-gram language model substantially enhances the performance of ASR models in both monolingual and cross-lingual settings. More specifically, our hypotheses are the following:

(**Mono1**)**:** The Swedish ASR model performs better on Norwegian than on Danish;

(**Mono2**)**:** The Danish ASR model performs better on Norwegian than on Swedish;

(**Mono3**)**:** The Norwegian ASR model performs better on Swedish than on Danish;

(**Mono4**)**:** The English ASR model performs worse than the Swedish, Danish, and Norwegian models on all three Scandinavian languages;

(**Mono5**)**:** In-domain 4-gram language model contributes towards improving the performance of all ASR models;

(**Mono6**)**:** Monolingual ASR models favour one dialect of the target language;

4.1.2. Experimental Setup

We investigate the cross-lingual comprehension of monolingual ASR models for the Scandinavian languages in two settings. Firstly, we locate the most accurate Wav2Vec 2.0-based ASR models for Swedish and Norwegian Bokmål available on Hugging Face. No Danish end-to-end ASR model was available on the platform, so we fine-tune one ourselves. All three models are described in greater detail in section 4.4. We then evaluate the ability of the three models to transcribe data in all three languages; that is, each model is tested on Swedish, Danish, and Norwegian.

In the first setting, we retain the parameters of model decoders unchanged with the exception of the vocabulary (set of legal characters): we replace the character *æ* with *ä* and *ø* with *ö* while testing the Danish and Norwegian models on Swedish, and vice versa in the opposite direction. These replacements allow us to ensure that the vocabulary of each model contains all the characters used in the target language. In the second setting, we provide language-specific supervision by boosting the decoders of the monolingual models with statistical language models in the target language.

We use 4-gram[1] language models which we compile with the *KenLM*[2] library on the train subsets of the NST datasets described in Section 4.3. We keep the standard parameters in both the language model and decoder. For comparison, we transcribe the Swedish, Danish, and Norwegian test sets using a monolingual ASR model in both settings, without any modifications to the vocabulary of the model.

We split up the entries in the three test sets according to the regional dialects the speakers belong to, combine the entries with the same regional dialect, and compute word and character error rates (the metric is explained in Section 4.5 per each regional dialect. Lastly, we use the *asr_evaluation* library to compile the most common transcription errors in all monolingual and cross-lingual settings.

4.2. Multilingual Scandinavian ASR Model

Multilingual training is often beneficial only to low-resource languages (Conneau and Lample, 2019; Devlin et al., 2019). As proven by Conneau et al. (2021), the case of multilingual ASR is slightly different, for a jointly pre-trained encoder can boost the performance of a multilingual model. Nevertheless, a joint decoder, i.e. one trained on several languages, leads to worse performance across most if not all languages (Conneau et al., 2021). It should be noted that the authors pre-trained and fine-tuned their multilingual model on ten highly divergent languages: Spanish, French, Italian, Kyrgyz, Dutch, Russian, Swedish, Turkish, Tatar, and Chinese. It is possible that the poor performance of their multilingual end-to-end model is, on part, due to the noise introduced from using ten highly divergent languages. In the case of syntactic parsing, multilingual models excel when the languages a model is trained on are similar. For instance, Smith et al. (2018) observed that joint training on Estonian, Finish, and North Sami treebanks leads to considerable improvements on all three languages in sentence and word segmentation, part-of-speech tagging, annotation of morphological features, and dependency parsing. Furthermore, Smith et al. (2018) report that joint training of Swedish, Danish, Norwegian, and Faroese mostly boosted the performance on all four languages as compared to monolingual baselines. We therefore suspect that if multilingual end-to-end ASR models are to be trained on typologically similar languages, they could demonstrate performance comparable to monolingual baselines.

We identify the inability to use language models in multilingual ASR models as a major drawback. To address this, we propose a novel architecture for multilingual ASR models, which is shown in Figure 4.1. Our modification to the standard Wav2Vec 2.0 model with CTC decoder involves adding language classification component. The additional component predicts the language from the encoded values and allows to boost the CTC decoder with the appropriate language model. The usage of the classification module as well as language models is though optional, as the proposed architecture has two settings. The first setting uses the default Wav2Vec 2.0 encoder and CTC decoder, with the latter being fine-tuned on the three Scandinavian languages, yet no language model. In the second setting, the information is firstly encoded through Wav2Vec 2.0 encoder, forwarded to the language classification module, and then decoded with the corresponding language model in the CTC decoder.

Lastly, we believe that the existing monolingual ASR models can be used to initialize multilingual models. Monolingual models already have high accuracy on

[1] We stick to 4-gram rather than 5-gram due to technical constraints, as the library was failing to produce a 5-gram language model for Swedish.

[2] https://github.com/kpu/kenlm

one of the languages; hence, the task becomes one of training a model to additionally transcribe in two more languages while retaining the ability to transcribe in the language the monolingual model is trained on. We first explore this idea by training multiple trial models and comparing their result against monolingual baselines as well as against a multilingual ASR model trained from scratch. We then choose the most effective setting to train our final multilingual model.

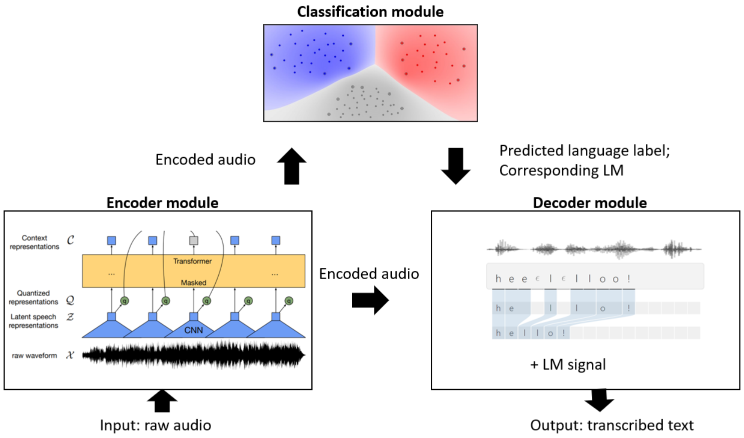

Classification module
Encoded audio
Predicted language label;
Corresponding LM
Encoder module
Decoder module
Context representations $\mathcal{C}$
Transformer
Masked
Quantized representations $\mathcal{Q}$
Latent speech representations $\mathcal{Z}$
raw waveform $\mathcal{X}$
CNN
Encoded audio
+ LM signal
Input: raw audio
Output: transcribed text

Figure 4.1.: The proposed multilingual, language-aware end-to-end ASR architecture with a language classification component. The images for the encoder and decoder components are borrowed from Baevski et al. (2020) and Graves et al. (2006) respectively.

4.2.1. Hypotheses

We hypothesize that the numerous similarities between Swedish, Danish, and Norwegian enable effective joint multilingual training. As such, we expect the approach with initializing the multilingual model with a monolingual one to outperform the training from scratch, since it allows to leverage the existing high quality of the model on one of the languages. We though expect the multilingual models to perform worse than the monolingual baselines. Our hypotheses can be formulated as follows:

(Multi1): Among the trial models, monolingual baselines are expected to achieve best results;

(Multi2): A multilingual model can be extended to handle all three languages by additionally fine-tuning it on the remaining two languages;

(Multi3): Without additional exposure to the data a monolingual model is trained on, i.e. by additionally fine-tuning the monolingual model only on the other two languages, the monolingual-turned-multilingual model's ability to transcribe its initial target language will degrade significantly;

(Multi4): The performance of the multilingual model across the dialects of Swedish, Danish, and Norwegian is less skewed than that of monolingual models.

(Multi5): Without the language classification module and language-specific language models, the multilingual ASR model is likely to make cross-lingual errors; for instance, it is likely to predict Norwegian words with a Swedish spelling.

4.2.2. Experimental Setup

We explore two approaches for training multilingual ASR models for the Scandinavian languages. Firstly, we train a multilingual decoder from scratch on all three languages at once. For the encoder, we use the available pre-trained VoxRex Wav2Vec 2.0 encoder trained for Swedish. We randomly initialize the decoder with parameters identical to those of the Wav2Vec 2.0 large VoxRex model described in section 4.4.1, as these parameters have reportedly led to good results for multiple ASR models published on the Hugging Face platform, whereas a grid search of the hyperparameters for Wav2Vec 2.0 models is computationally heavy. We then fine-tune the decoder with a shared vocabulary on a subset of 15,000 randomly chosen samples from each language, 45,000 samples in total, for 5 epochs.

In the second approach, we attempt second language acquisition in an ASR model; that is, we further fine-tune the monolingual Wav2Vec 2.0 large VoxRex Swedish model to see whether it can acquire a second and third language. We choose the Swedish model since it has been trained for most updates and has the lowest WER in its target language. We attempt three settings: firstly, we fine-tune the Swedish model on 15,000 randomly sampled entries per language in Danish and Norwegian, without feeding it any Swedish data. In the second setting, we feed it with 15,000 entries per language in Danish and Norwegian, and 7,500 entries in Swedish. In the third setting, we fine-tune the model on 15,000 entries in each language. In all settings, we fine-tune the model for 5 hours with early stopping at 24 hours.

In all approaches, we use CTC loss as the objective function. We evaluate the model on a held-out validation dataset of 2,000 entries per language every 1,000 updates. We then compare the performance of the multilingual models across all three languages to select the most suitable approach, which we adopt to fine-tune the final multilingual model for three days. We also compare the performance of multilingual models to monolingual baselines trained with the exact same parameters and settings, but on one language at a time.

As for the classification component, we randomly initialize a linear layer which we then train to predict the language label from the encoded audio representations. We train the classifier on randomly selected and shuffled 15,000 entries from the train set per language. We train the classifier for two days with the following parameters: train and evaluation batch sizes of 4, 2 gradient accumulation steps, learning rate of 1e-4, mean pooling, and maximum number of epochs set to 100. We then individually evaluate the classification module on a combination of all three test sets.

Lastly, add the classification module to the multilingual ASR model by re-routing the output of the encoder to the classification module. The module predicts the language and selects the correct language model to further boost the decoder of the ASR model. We evaluate the performance of the proposed architecture across the test sets for all three languages.

4.3. Datasets

4.3.1. Common Voice (CV)

Common Voice (Ardila et al., 2020) is an open-source multilingual dataset of speech and text pairs. The data is gathered by volunteers who are offered to contribute by submitting text, speech, or both. The submitted texts and recordings are additionally validated. At present, the dataset contains 18,243 hours of recordings and their corresponding texts, 14,122 hours of which have been validated. The recently released 8.0 version contains data in 87 languages including numerous low-resource languages, such as Votic, Erzya, and Baasa. The distribution of the data across the languages is though highly skewed, as English, German, and French constitute roughly 29% of all the validated hours available in the dataset. When it comes to the Scandinavian languages, CV 8.0 contains data in Swedish, Danish, and Norwegian Nynorsk, yet amount of data available per language, which is shown in Table 4.1, differs greatly. Nevertheless, the majority of these data are manually verified and thus high-quality. Furthermore, data in Swedish, Danish, and Norwegian available in CV 8.0 mirror or even exceed the amounts of data available in most low-resource languages.

Language	**Overall**	**Validated**
Swedish	48	40
Danish	6	6
Norwegian Nyrorsk	0.38	0.3

Table 4.1.: Hours of data per language available in the Common Voice 8.0 dataset.

4.3.2. Nordisk Språkteknologi (NST)

NST is an extensive ASR dataset available in Swedish, Danish, and Norwegian Bokmål. The data consist of speech and transcription pairs manually compiled, mainly for the purpose producing ASR models. Data additionally include transcribed telephone conversations as well as recordings and transcriptions with the recordings of conversations at an office. The data has been manually validated for the most part, and the audio quality is high despite being recorded in early 2000s. Most recordings in the dataset are available in two channels recorded with two separate microphones, one placed right next to the speaker, and one across the room. The dataset also includes richly annotated metadata for every speaker, disclosing their age, gender, region of birth, and regional dialect. The region of birth and regional dialect indicate the geographic region a speaker was born in and the regional dialect they speak; both pieces of information were self-reported by the speaker.

Following a recent modernisation of the dataset, it is now available in a reader-friendly format; however, the modernisation of the Swedish test set is still in progress. To avoid possible errors in parsing the old versus new version of the dataset, 20% of the Swedish train set with no overlap are used as the test in this study, and the original test set is omitted. In addition, due to an already large dataset and lack of notable differences between the audio recorded with two microphone, the second channel, namely the audio recorded with a distant microphone, is omitted. The amounts of data in each language used in this study are shown in Table 4.2 below.

Language	**Train**	**Test**
Swedish	303.7	73.2
Danish	126.9	77.1
Norwegian Bokmål	436.3	115.3

Table 4.2.: Hours of data per language in the NST dataset.

The regional dialects represented in the dataset as well as amounts of data available in the dataset per dialect are reported in Tables 4.3, 4.4, and 4.5 below. The variables region of birth and regional dialect match in most cases; yet for cases where the region of birth differed from the self-reported regional dialect, the latter was chosen for the dialect label. All audio and text pairs were discarded where both the region of birth and regional dialect were unspecified or specified as 'other'.

The regions of birth and regional dialects reported in the dataset differ from the dialectal boundaries drawn by linguists, which are addressed in Section 5.1.2. It is therefore debatable whether the dataset is representative of the dialectal varieties of the Scandinavian languages. It is, however, representative of the major regions and the language variety spoken in that region, which corresponds to the term regiolect. As shown in section 5.1.2, the dialects of the Scandinavian languages can, for the most part, be mapped to a certain region. Therefore, the terms dialect and regiolect are henceforth used synonymously in this thesis.

Region of birth and regional dialect match in most cases; yet for cases where the region of birth differed from the self-reported regional dialect, the latter was chosen for the dialect label. All audio and text pairs were discarded where both the region of birth and regional dialect were unspecified or specified as 'other'.

Table 4.3.: The size of the Swedish subset of the NST dataset. The number of hours is rounded to one decimal place. The names of regions are directly adopted from the dataset.

	Train		**Test**	
	Hours	Speakers	Hours	Speakers
Norrland	44.0	109	10.7	26
Stockholm med omnejd	42.3	137	11.7	6
Västra sydsverige	38.1	97	10.2	28
Mellansverige	30.7	87	9.7	26
Göteborg med omnejd	29.9	97	6.7	43
Västergötland	27.6	72	5.5	14
Östergötland	25.5	66	4.8	11
Östra sydsverige	25.3	61	6.3	16
Dalarna med omnejd	23.6	62	5.1	13
Västsverige	16.6	45	2.4	7
Total	**303.7**	**833**	**73.2**	**190**

Table 4.4.: The size of the Danish subset of the NST dataset. The number of hours is rounded to one decimal place. The names of regions are taken directly from the dataset.

	Train		**Test**	
	Hours	Speakers	Hours	Speakers
Storkøbenhavn	29.5	185	27.8	20
Fyn	17.1	88	8.5	7
Vest- og Sydsjælland	16.7	87	9.2	6
Vestjylland	16.5	87	6.9	5
Nordjylland	16.0	82	8.1	6
Østjylland	15.5	86	8.9	7
Sønderjylland	15.2	82	7.6	6
Total	**126.4**	**697**	**77.1**	**57**

Table 4.5.: The size of the Norwegian subset of the NST dataset. The number of hours is rounded to one decimal place. The names of regions are directly adopted from the dataset.

	Train		**Test**	
	Hours	Speakers	Hours	Speakers
Bergen og Ytre Vestland	54.0	110	8.3	6
Oslo-området	52.1	128	25.3	24
Ytre Oslofjord	49.2	116	7.3	6
Sør-Vestlandet	46.3	98	10.3	7
Trøndelag	39.4	92	9.3	8
Sørlandet	38.7	96	9.0	6
Voss og omland	33.6	85	9.4	6
Hedmark og Oppland	31.3	86	8.5	9
Nordland	31.0	74	8.8	6
Sunnmøre	30.6	65	9.3	7
Troms	30.2	65	9.6	5
Total	**436.3**	**1015**	**115.3**	**90**

4.3.3. Dataset Usage

We use the CV dataset for validation of trial models in different settings. More specifically, we use the data to evaluate the multilingual models during training. To reduce the computational costs, we limit the size of the dataset to 2,000 randomly sampled entries per language, amounting to just over an hour of data per language. For Norwegian, we use a random held out sample of 2,000 entries from the NST training subset instead. This is to ensure that the training, evaluation, and testing subsets contain the same language variety, namely Norwegian Bokmål.

We use the NST dataset to both train and evaluate the models. When it comes to the latter, we use the entire test sets in the the three languages. For training, we randomly sample smaller portion of the train subsets due to limited access to computational power as well as space constraints. More specifically, we train the monolingual Danish ASR model on 30,000 random entries from the Danish train subset. Samples of training data for the multilingual models are limited to 15,000 random entries per language.

4.3.4. Text and Audio Pre-processing

Wav2Vec 2.0-based ASR models are highly susceptible to changes in the sampling rate of the audio as well as the vocabulary (character set) that constitutes the data. This is due to the fact that Wav2Vec 2.0-based models are trained to map audio signals of specified dimensionality to a pre-defined set of characters. Should the sampling rate differ from the one encountered in training, the model's encoder would likely fail to interpret the dimensionality of the data. Similarly, a trained Wav2Vec 2.0 model is unable to predict characters that are outside of its vocabulary. To accurately evaluate the performance of an ASR model, it is therefore crucial to ensure that the sampling rate and the vocabularies match across the train and test set.

To address the above, we additionally pre-process both audio and text. We down-sample the audio files from the CV dataset to 16 kHz, which is the standard sampling rate for Wav2Vec 2.0 ASR models. Lower sampling rate generally leads to lower audio quality; nevertheless, a lower sampling rate also requires less computational power, which allows for more efficient training. We carry out downsampling using the built-in resampling feature available in the Hugging Face Dataset class. As for the text, we lowercase all characters and remove all non-alphanumeric characters, such as punctuation markers. We keep the numerals in the data unchanged, as excluding numerals from the transcription yet leaving it in the audio would distort the alignment of the data.

4.4. Models

The quality of ASR models trained with Wav2Vec 2.0 largely depends on the availability of training data as well as computational resources. The amount of data described in 4.3 is considerable and sufficient to train accurate models; however, the access to computational resources devoted to this thesis was limited, as at most 1 GPU was available at all times. In light of this, whenever possible, we use the models available on Hugging Face instead of training all models from scratch. As such, we directly adopt five open-source models from Hugging Face: Swedish pre-trained model, Swedish end-to-end model, Norwegian end-to-end model, Danish pre-trained model, and English end-to-end model. These models, alongside the multilingual trial models, are further discussed in the following subsections.

4.4.1. Swedish

We use two Swedish ASR models: a pre-trained model and an end-to-end model. Firstly, we use *large VoxRex (version C)*[3] (Malmsten et al., 2022) as the pre-trained Wav2Vec 2.0 model for Swedish. As reported by the authors, the model has been pre-trained for 400,000 updates on 10,000 hours of Swedish radio records and 1,500 hours of data owned by the National Library of Sweden. For the end-to-end model, we use *Wav2vec 2.0 large VoxRex Swedish (C)*[4], which a fine-tuned version of the pre-trained model. The end-to-end model has been fine-tuned for 120,000 updates on a combination of the CV and NST datasets. The model demonstrates outstanding performance as, with a 5-gram language model, it achieves a WER of 3.73% on the Swedish test subset of CV. Both models are trained, maintained, and shared by the National Library of Sweden. To the best of our knowledge, the two models are the best pre-trained and end-to-end models for Swedish at the moment of writing this thesis.

4.4.2. Norwegian

We use an end-to-end Norwegian ASR model trained and shared by the National Library of Norway, namely the *Norwegian Wav2Vec2 Model - 1B Bokmål* [5]. As reported by the authors, the model is fine-tuned on top of the Swedish *large VoxRex* for roughly 18,560 updates on the Norwegian Parliamentary Speech Corpus. When boosted with an in-domain 5-gram language model, the model achieves a WER of 12.22% on the test subset of the NPSC dataset. To the best of our knowledge, this is state-of-the-art end-to-end ASR model for Norwegian Bokmål.

[3] Available at: https://huggingface.co/KBLab/wav2vec2-large-voxrex; Last accessed: May 9th, 2022

[4] Available at: https://huggingface.co/KBLab/wav2vec2-large-voxrex-swedish; Last accessed: May 9th, 2022

[5] Available at: https://huggingface.co/NbAiLab/nb-wav2vec2-1b-bokmaal; Last accessed: May 9th, 2022

4.4.3. Danish

We use a pre-trained ASR model for Danish trained and made public on Hugging Face by Alvenir.ai[6]. As reported by the authors, the model has been pre-trained on roughly 1,300 hours of audio books and podcast recordings. The number of updates and epochs are not reported. Since no end-to-end Wav2Vec2.0 Danish models were available at the moment of writing the thesis, we train such model ourselves. We combine the aforementioned pre-trained Danish model with a randomly initialized CTC decoder, which we then train on one GPU for roughly 5,800 updates (10 epochs) on randomly sampled 20% of the Danish NST train set. We retain the parameters of the pre-trained model for the encoder. In the decoder, we set the batch size to 10, gradient accumulation steps to 3, learning rate to 1e-4, and weight decay to 0.005. The complete parameters of the Encoder and Decoder modules are shown in Tables A.1 and A.2 in the appendix. With a 4-gram language model, the end-to-end Danish ASR model achieves a WER of 13.82% on the Danish NST test set.

4.4.4. English

We use an end-to-end English ASR model proposed in the original Wav2Vec 2.0 paper (Baevski et al., 2020), namely *Wav2Vec 2.0 base 960h*, which we accessed from Hugging Face[7]. According to Baevski et al. (2020), the model is both pre-trained and fine-tuned on 960 hours of data (roughly 250,000 samples) from the Librispeech dataset for 400,000 training steps. The model is then fine-tuned on 100 hours of same data for an unreported number of updates. With a 5-gram language model, the end-to-end English model achieves a WER of 3.4% and 8.0% on the Librispeech *clean* versus *other* test set respectively.

4.4.5. Trial Multilingual Models

We examine the most suitable multilingual setting by training four trail multilingual models and three monolingual baselines. Our experiments involve training only the CTC decoder component of the model. Due to the demanding nature of pre-training a Wav2Vec 2.0 encoder, we use the *large VoxRex (version C)* pre-trained model for Swedish (Malmsten et al., 2022) in all setting. In the most straightforward approach, we fine-tune a multilingual model from scratch on 15,000 randomly sampled entries from the NST test set in each language. We further refer to this model as *From Scratch DA+NO+SE*. In the remaining multilingual approaches, we initialize the multilingual models from *Wav2vec 2.0 large VoxRex Swedish (C)*, which is both pre-trained and fine-tuned on Swedish. We attempt additionally fine-tuning the Swedish model on 15,000 entries in Danish and Norwegian (*Retraining DA+NO*), 15,000 entries in Danish and Norwegian as well as 7,500 entries in Swedish (*Retraining DA+NO+SE_half*), and 15,000 entries per language in all three languages (*Retraining DA+NO+SE_full*). Lastly, we train monolingual baselines by initializing a CTC decoder on top of the *large VoxRex (version C)* pre-trained Swedish model and fine-tune it on 15,000 entries in the corresponding language. All trial models are trained for 24 hours but not more than 5 epochs. The results of the trial models are covered in Section 5.2.1.

[6] Available at: https://huggingface.co/Alvenir/wav2vec2-base-da; Last accessed: May 9th, 2022

[7] Available at: https://huggingface.co/facebook/wav2vec2-base-960h; Last accessed: May 9th, 2022

4.5. Evaluation metrics

We evaluate the performance of ASR models using Word Error Rate (WER), Character Error Rate (CER) and qualitative analysis of the errors. We estimate WER and CER using the Hugging Face implementation of the metrics, whereas we conduct qualitative analysis using the asr-evaluation toolkit[8].

4.5.1. Word Error Rate

Word Error Rate (WER) is an ASR evaluation metric that estimates the number of errors in transcription relative to the number of words. The calculation of WER is shown in equation 4.1, where three types of errors, namely substitutions (S), insertions (I), and deletions (D) are summed and divided by the number of tokens in the reference (N, usually words or characters).

$$WER = \frac{S+I+D}{N} \tag{4.1}$$

Substitution is an instance when the spelling of a word predicted by the model deviates from the reference transcription. Insertion is an erroneous prediction of an additional word that is absent in the reference transcription. Lastly, deletion is a type of error that occurs when a word present in the reference transcription is omitted in the predicted transcription. Examples of word-level substitution, insertion, and deletion are shown in Table 4.6.

Table 4.6.: Word-level examples of Substitution, Insertion, and Deletion.

	Reference	**Prediction**	WER
Substitution	*This is an error*	*This is an eror*	0.25
Insertion	*This is an error*	*This is an an one error*	0.33
Deletion	*This is an error*	*This (is an) error*	0.5

WER is arguably the most common ASR evaluation metrics used in numerous benchmarks, such as LibriSpeech (Panayotov et al., 2015c), LibriLight (Kahn et al., 2020), and TIMIT (Garofolo et al., 1992). As suggested by Ali and Renals (2018), WER is most accurate when a model is evaluated on at least two hours of data. All test subsets used in this research, including the distributions of dialects in the test sets, contain at least two hours of data; therefore, we use WER as the core evaluation parameter in this research.

4.5.2. Character Error Rate

Character Error Rate (CER) is an ASR evaluation metric that estimates the number of erroneous characters (letters) relative to the number of correct characters. The metric is calculated using the same equation as WER (4.1); however, the substitutions, insertions, and deletions occur on a character level. An insertion is thus an incorrectly predicted character rather than an incorrectly predicted word. Examples of character-level substitution, insertion, and deletion are shown in Table 4.7.

CER is a common secondary evaluation metric which is often used alongside WER, as is the case with the aforementioned benchmarks (Garofolo et al., 1992; Kahn et al., 2020; Panayotov et al., 2015c). CER addresses a major limitation of the word-level metric, which is that WER penalizes the score equally regardless of whether there is one or more mistakes in the spelling of a word. In other words, a

[8] Available at: https://github.com/belambert/asr-evaluation; Last accessed: May 14th, 2022

Table 4.7.: Character-level examples of Substitution, Insertion, and Deletion. The number of tokens used in calculating the CER includes spaces. CER scores are rounded to three decimal places.

	Reference	**Prediction**	CER
Substitution	*This is an error*	*This is am error*	0.063
Insertion	*This is an error*	*This is ann errror*	0.125
Deletion	*This is an error*	*Thi(s) is a(n) er(r)or*	0.188

completely random sequence of character is considered to be as severe a mistake as a spelling with just one character off. We therefore use CER as a secondary evaluation metric for better insight into the quality of predicted transcriptions.

4.5.3. Qualitative Analysis

We additionally supplement the evaluation of the quality of ASR models with brief qualitative analysis of errors and predicted transcriptions. We report most common errors and their types (substitutions, insertions, and deletions) alongside their frequencies. We use the results of qualitative analysis for better overview of the challenges faced by the models as well as for verification of whether the differences captured by the WER and CER are indeed errors, or possible alternative spellings. Lastly, we include examples of the transcriptions predicted by models for better understanding of whether the evaluation metrics allow to fully judge the quality of models.

5. Results and Discussion

This section is divided into two parts, namely Monolingual Models (5.1) and Multilingual Models (5.2). In the first part, we present and discuss the performance of the monolingual Scandinavian ASR models in a zero-shot transfer setting (5.1.1), perform a brief qualitative analysis of some of the most common transcription errors (5.1.3), and compare their performance across regional dialects (5.1.2). In the second part, we present and analyse the performance of the trial multilingual models (5.2.1) and the language classification component (5.2.2), followed by an overview of the final multilingual model (5.2.3).

5.1. Monolingual Models

5.1.1. Zero-shot Transferability and Mutual Intelligibility

The performance of the monolingual Swedish, Danish, and Norwegian ASR models across the three languages is reported in Table 5.1. The table also includes the performance of the English end-to-end model on the three languages for comparison.

Table 5.1.: The performance of the monolingual models on the Scandinavian languages, no language model versus a 4-gram language model. The percentages are rounded to two decimal places. The left-most column denotes the language on which the model was tested.

		No LM		With LM	
Model	Test Set	WER	CER	WER	CER
Swedish	Swedish	2.19%	0.98%	2.74%	1.07%
	Danish	78.58%	39.66%	72.69%	40.67%
	Norwegian	61.78%	21.29%	52.06%	19.64%
Danish	Swedish	120.10%[1]	61.03%	98.93%	56.94%
	Danish	19.14%	5.45%	13.82%	4.33%
	Norwegian	104.56%	52.59%	90.06%	49.52%
Norwegian	Swedish	83.51%	26.34%	73.82%	24.70%
	Danish	83.79%	36.82%	75.05%	36.33%
	Norwegian	16.47%	3.62%	12.03%	2.81%
English	Swedish	110.06%	50.26%	93.59%	48.53%
	Danish	99.50%	54.11%	88.71%	54.75%
	Norwegian	102.52%	49.79%	90.35%	48.22%

[1] WER and CER can be above 100% when the number of insertions and/or deletion errors outnumbers the number of words in the reference.

Analysis of the results with focus on the applicability of zero-shot transferability of the monolingual ASR models for the Scandinavian languages reveals that the performance differs greatly depending on the source (model) and target (test set) language. A monolingual Swedish model achieves a lower WER and CER on Norwegian than on Danish. A Danish model, on the other hand, performs better on

Norwegian than on Swedish. Lastly, the performance of the Norwegian model is similar on Swedish and Danish in terms of WER, with a lower error rate on the former. The difference is though more noticeable on the CER parameter.

The first three hypotheses for the monolingual models expected the monolingual ASR models for the Scandinavian languages to follow the mutual intelligibility patterns of native speakers; that is, much like a native speaker of Swedish, a Swedish ASR model would do better job at transcribing Norwegian than Danish, a Danish model would perform better on Norwegian than on Swedish, and a Norwegian model would perform better on Swedish than on Danish. The results appear to support the three hypotheses, as the performance of monolingual ASR models for the Scandinavian languages seems to match the patterns of mutual intelligibility of native speakers of the Scandinavian languages reported by Delsing (2005) and Gooskens (2007). It should be noted that in both studies, the task for native speakers was to answer questions asked in their non-native language (response was given in their native language). This task differs from transcription, as transcription arguably does not necessarily require one to understand the question asked, whereas answering a question does not require one to know the correct spelling of the words present in the questions. Nevertheless, albeit the somewhat different natures of the task, the levels of comprehension of the foreign Scandinavian languages in Scandinavians seemingly resemble the ability to transcribe the foreign Scandinavian languages in zero-shot transfer setup for monolingual ASR models.

Comparison of the performance of monolingual Scandinavian models with an English model offers further insight into the subject. The English model is outperformed by the Swedish and Norwegian models across all three languages; nevertheless, the English model appears to perform better on Swedish and Norwegian as compared to the Danish model, on both parameters. However, when boosted with a language model, the Danish model outperforms the English counterpart on Norwegian in terms of WER. The overall applicability of zero-shot transfers for Scandinavian ASR models is thus questionable. In best setting, namely when the Swedish model is used to transcribe Norwegian, WER remains above 60% without a language model and 52% with a language model.

The results do not seem to support the fourth hypothesis for the monolingual models, which expected the English model to have the poorest performance out of the four models tested on Swedish, Danish, and Norwegian. More specifically, the hypothesis breaks with the Danish model. A plausible explanation is the overall limited fine-tuning of the Danish end-to-end model. As mentioned in the model description in Section 4.4.3, the decoder is fine-tuned for only 4,000 updates, which is a relatively low number in comparison to the other models tested in this project. It could be the case that with further fine-tuning on data in Danish, the model would perform better not only on Danish, but also on Swedish and Norwegian. In the case of the English model, which has been fine-tuned for 400,000 steps already, further fine-tuning on English data is highly unlikely to affect the performance on the Scandinavian languages.

Lastly, the results reveal the effectiveness of boosting Wav2Vec 2.0-based ASR models with simple n-gram language models. An in-domain 4-gram language model appears to improve both WER and CER in most settings. In the case of Danish and Norwegian ASR models, a language model steadily improves the performance on both the language a model is trained on and foreign languages. When it comes to the Swedish model, however, a 4-gram language model evidently hinders the model's ability to transcribe Swedish data, as both WER and CER become higher. A language model nevertheless improves the Swedish models' ability to transcribe

Danish and Norwegian. Lastly, an in-domain 4-gram also boosted the performance of the English model on all three Scandinavian languages.

The fifth hypothesis for monolingual models expected n-gram language models to improve the performance of ASR models on all languages. The hypothesis cannot be confirmed, as a 4-gram language model leads to higher WER and CER in the Swedish ASR model when used on Swedish data. In all other cases, however, the addition of an n-gram language model leads to improvement. The effectiveness of n-gram language models for monolingual ASR systems has been vastly researched (Baevski et al., 2020; Conneau et al., 2021; Håkansson and Hoogendijk, 2020; Khassanov, 2020) and was therefore an expected observation. The steady gains in terms of the WER and CER in most setting are likely, on part, due to the matching domains of the train and test sets, which is described by Khassanov (2020) as a deciding factor of whether a statistical language model is an effective addition to an ASR model. As seen from the results, the in-domain knowledge seemingly transfers across languages as well. However, the case of the Swedish model suggests that there might be extent to which statistical in-domain language models are efficient for ASR models. The Swedish ASR model has extremely low WER and CER to begin with (2.19% and 0.98% respectively), so a 4-gram language model only misleads the end-to-end model. Arguably, this might be due to the nature of statistical language models, as it implausible for them to capture all the possible combinations of tokens, no matter their size. To summarize, statistical language models appear to be an overall useful addition to ASR models in both monolingual and cross-lingual settings; however, they might mislead high-quality ASR models.

5.1.2. Performance across the Regional Dialects

The performance of the Swedish end-to-end model across the Swedish regional dialects is shown in Figure 5.1, the Danish model on the Danish regional dialects in Figure 5.2, and the Norwegian model on the Norwegian regional dialects in Figure 5.3. A complete list of the scores obtained in cross-lingual settings (e.g. Swedish model on Danish regional dialects) and monolingual (Swedish model on Swedish regional dialects) settings is shown in Table B.4 in the Appendix.

The results reveal that the performance of the monolingual ASR models on the languages they were trained on varies across the regional dialects. The Swedish model performs best on *Östergötland* (1.68% WER and 0.66% CER without a language model) and worst on *Västra sydsverige* (2.35% WER 0.91% CER with a language model) regional dialects. The observed standard deviation between the Swedish regional dialects on the WER parameter constitutes 0.202% without a language model and 0.240% with a language model. The Danish model performs best on *Fyn* (11.13% WER and 3.17% CER with a language model) and worst on *Østjylland* (15.37% WER and 4.84% CER with a language model) regional dialects. The standard deviation between the Danish regional dialects on the WER parameter is estimated at 1.528% without a language model and 1.237% with a language model. The Norwegian model performs best on *Hedmark oh Oppland* (10.40% WER and 2.29% CER with a language model) and worst on *Bergen og Ytre Vestland* (15.20% WER and 4.03% CER with a language model) regional dialects. The standard deviation on the WER parameter across the regional dialects of the Norwegian language is roughly 2.211% without a language model and 1.554% with a language model.

The low variance in the quality of the Swedish model across the Swedish regional dialects may stem from overall high quality of the model. The high variance in the case of Norwegian, in addition to a poorer quality of the model, may be linked to the dialect-rich nature of the language, with major differences being present on

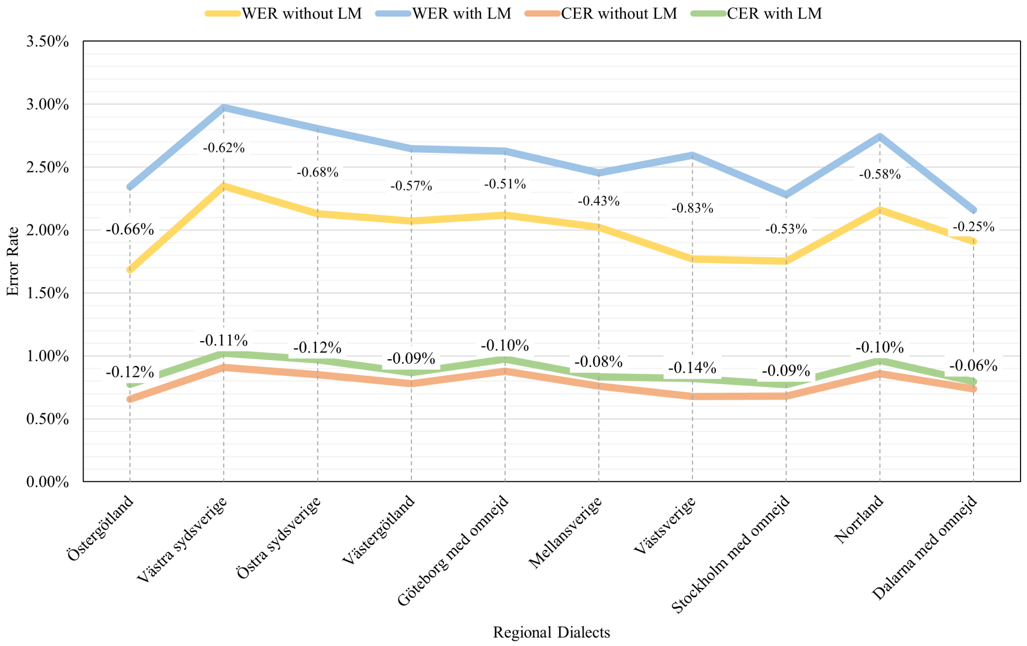


Figure 5.1.: The performance of the monolingual Swedish ASR model across the Swedish regional dialects, no language model versus a 4-gram language model. The numbers alongside pairs of lines indicate the improvement in the WER made by the addition of a language model. In the case of Swedish, the addition of a statistical language model has negative effect on the performance, thus the numbers are negative.

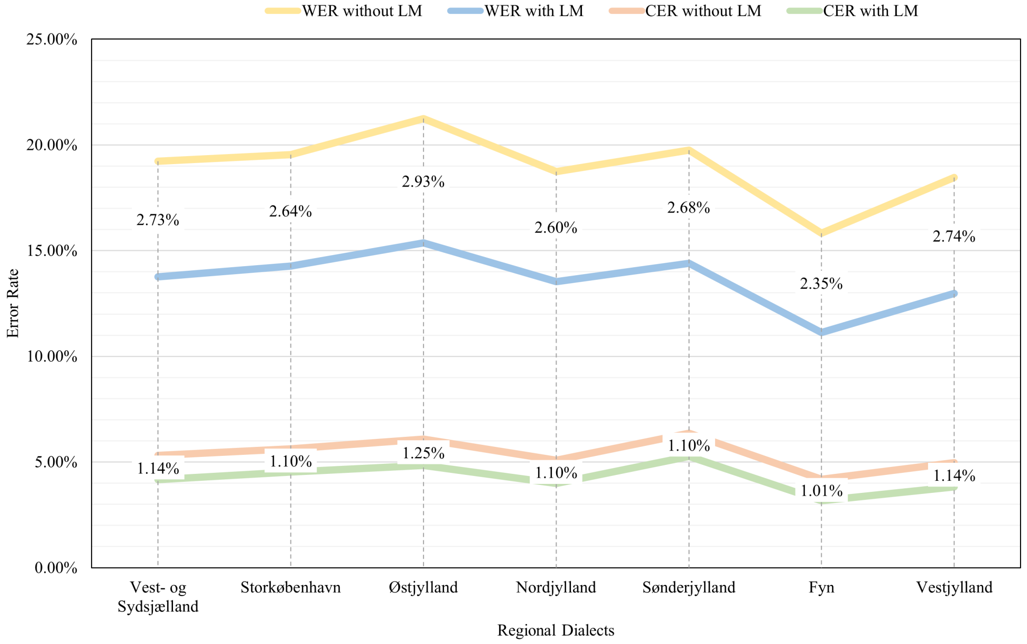


Figure 5.2.: The performance of the monolingual Danish ASR model across the Danish regional dialects, no language model versus a 4-gram language model. The numbers between the two lines indicate the difference in the WER made by the addition of a language model.

spoken and written level (Kristoffersen, 2000a). Most notably, the addition of a statistical language model consistently improves the quality of transcription in all

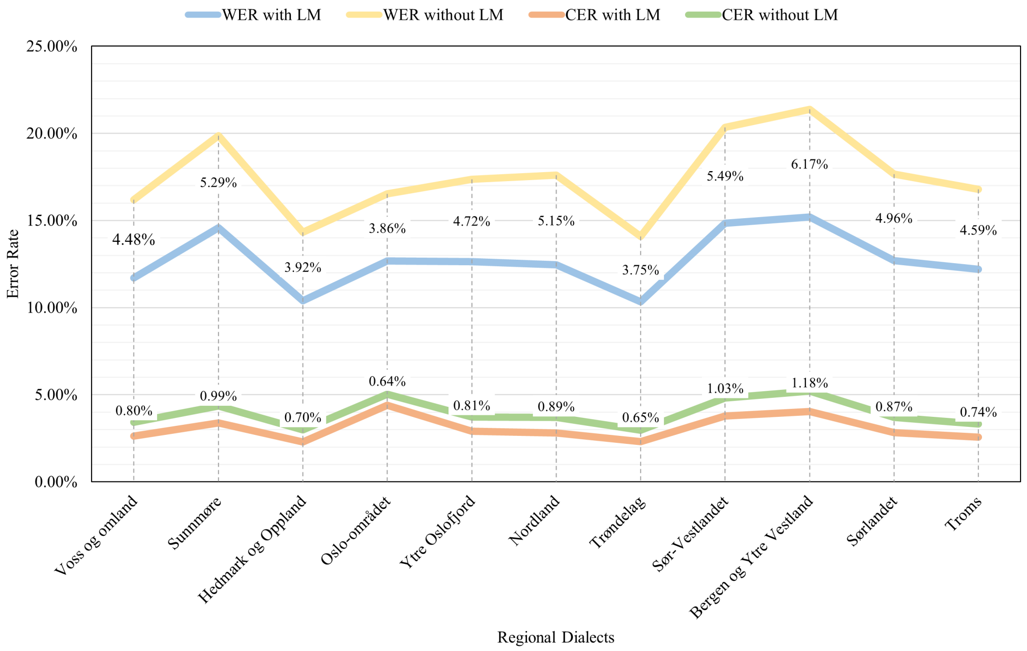


Figure 5.3.: The performance of the monolingual Norwegian ASR model across the Norwegian regional dialects, no language model versus a 4-gram language model. The numbers alongside pairs of two lines indicate the difference in the WER made by the addition of a language model.

regional dialects in Danish and Norwegian. In the case of Swedish, it leads to worse performance in all dialects.

As seen from Table B.4 in the Appendix, the performance of the monolingual models in a cross-lingual setting differs as well. Both WER and CER parameters remain high on all dialects in all cross-lingual settings. It should be noted that the addition of a statistical language model consistently improve both parameters in cross-lingual settings on all dialects. The results can be further analysed with focus on the cross-lingual similarities of the dialects to other Scandinavian languages; however, such extensive analysis is out of the scope of this thesis.

The last hypothesis for the monolingual models expected the ASR models to be biased towards one dialect, likely the one they are most exposed to during the training. While results reveal that not all dialects are transcribed equally successfully, the evidence is arguably insufficient to confirm the hypothesis. As reported in the numbers above, The Danish and Norwegian models seem to exhibit striking difference in the performance across the dialects. Even though it is possible to identify the regional dialect on which these models perform the best, the difference is arguably insignificant. Furthermore, it cannot be excluded that the difference is due to skewed representation of the dialects in both train and test set.

5.1.3. Qualitative Analysis of Errors

The most common insertion, deletion, and substitution errors in the transcriptions of the test produced by monolingual models on their target language are shown in Table 5.2. Examples of the transcriptions produced by the monolingual models can be seen in Tables B.1, B.2, and B.3.

The error patterns reveal that the models commit most errors while transcribing function words – words that ensure grammatical rather than semantic soundness of

Table 5.2.: A side-by-side comparison of the most common insertion, deletion, and substitution errors in the transcriptions produced by the Swedish, Danish, and Norwegian models, with and without language model, on their target languages (e.g. Swedish on Swedish). The number in the parentheses is the raw frequency of the error. For substitutions, the token left of the arrow is the target and right of the arrow is the prediction.

No Language Model			With Language Model		
Insertion	Deletion	Substitution	Insertion	Deletion	Substitution
SWEDISH					
i (191)	i (50)	de → det (193)	i (170)	i (82)	de → det (198)
en (25)	och (46)	det → de (141)	mitt (50)	och (45)	det → de (185)
för (25)	det (31)	skall → ska (114)	en (28)	läder (30)	skall → ska (128)
det (20)	en (39)	istället → stället (54)	det (19)	en (29)	skröplig → skröpplig (94)
så (19)	att (28)	idag → dag (44)	att (17)	det (28)	börje → börjar (79)
DANISH					
er (292)	i (726)	er → af (357)	er (376)	i (607)	er → af (381)
en (238)	er (441)	af → er (289)	i (319)	er (411)	as → a (288)
i (236)	en (219)	as → a (245)	en (310)	at (216)	af → er (287)
at (194)	af (196)	i → e (224)	at (244)	af (203)	åbn → åben (200)
et (109)	at (195)	e → i (176)	for (183)	en (163)	fra → for (193)
NORWEGIAN					
i (271)	r (1897)	r → er (2196)	i (203)	e (1305)	r → er (1640)
for (187)	e (1839)	e → er (503)	for (201)	r (1276)	én → en (524)
er (159)	s (1138)	én → en (493)	er (164)	i (817)	er → r (410)
det (145)	t (1106)	n → en (421)	det (142)	t (758)	er → har (246)
til (129)	i (1060)	l → el (328)	til (141)	s (635)	sju → syv (239)

sentences. Examples of such errors include prepositions (*i, för, for*), pronouns (*en, det, er*), and conjunctions (*och, så, for*). The high frequency of such errors is partially due to the high frequency of function words in general. C. Chung and Pennebaker (2007) estimated that roughly every fifth word is a function word in the English language, with similar frequencies being observable in most other languages. In addition, function words tend to be short in both pronunciation and spelling. Given their high frequency and short pronunciation time, it is arguably not surprising that the models occasionally miss or transcribe erroneously some of the tokens. When it comes to Norwegian, most frequent errors seem to occur due to the model's inability to transcribe individual letters, which are from instances of spelled out words in the dataset. It is rather plausible that the model is under-exposed to words being spelled out during training, since, as mentioned in Section 4.4.2, the model is trained on a different dataset.

It can be argued that not all of the insertions, deletions, and substitutions captured by the algorithm are errors, and not all errors are classified correctly. The third most common substitution error in Swedish states that the model incorrectly transcribes *skall* as *ska*, both of which are alternative spellings of the same auxiliary verb, with the latter being more frequent in modern Swedish. Both words have the same pronunciation (*/ska/*), so a model is likely to learn whichever spelling is more frequent in the training data. Nevertheless, most ASR evaluation metrics consider this to be an error, as the spelling of the reference and prediction do not match. Furthermore, WER and CER error metrics appear to over-generate the number of errors in the Scandinavian (and possibly other) languages, as they are unable to identify instances of incorrect conjunction or disjunction. As a result, the disjunction of *istället* into *i* and *stället* is considered to be two errors, namely an instance of substitution of *istället* with *stället* and an insertion of *i*, even though both options are grammatically correct to begin with. An example of such an error is shown below:

(1) **Reference**: *många talar indianspråk istället för spanska*
Prediction: *många talar indianspråk i stället för spanska.*

Lastly, the analysis of error patterns offers additional insight into the effects of a language model on the quality of the produced transcriptions. An n-gram language model causes the Swedish ASR model to under-generate the preposition *i*, which leads to more deletion errors, yet fewer insertion errors. In addition, a language model leads to over-generation of the personal pronoun *mitt* as well as incorrect spelling of *skröplig*. For Danish, a language model leads to a more frequent prediction of function words, such as the preposition *i*, resulting in fewer deletions yet more insertion errors of such words. It, however, appears to consistently resolve the erroneous substitution of *i* with *e*. Lastly, when it comes to Norwegian, the addition of a language model appears to be a very consistent solution for the deletion of one character tokens, such as *e*, *r*, and *i*, and incorrect substitutions of one character tokens with two-character words, such as *r* with *er*; however, it increases the number of erroneous insertions of a few function words, such as *for*, *er*, and *til*. Thus, from an overview of the most frequent errors, it can be concluded that an n-gram language model offers a solution to some errors; nevertheless, plausibly due to the nature of statistical language models, it also causes errors, mostly by over-generating function words.

5.2. Multilingual Model

We train two components of the multilingual end-to-end Scandinavian ASR model, namely a CTC decoder and a language classification module. The former are covered in Section 5.2.1 and the latter in Section 5.2.2. We then combine the two components into a single multilingual end-to-end Scandinavian ASR model. The model is evaluated and discussed in Section 5.2.3.

5.2.1. Trial Multilingual Models

We determine the most suitable settings for the CTC decoder by firstly training four trial models alongside monolingual baselines. The results achieved by the trial models are shown in Figures 5.4 and 5.5. It should be noted that we trained the models with 1,000 warmup steps, which are not visualized in the figures.

The results of the trial models shed light on the optimal data settings and the effects of bootstrapping the ability to transcribe Danish and Norwegian from an existing Swedish model. Most noticeable is the degrading performance on the Swedish data when the model is re-trained only on Danish and Norwegian data. Similarly, when a smaller portion of data is used for Swedish than for the other two languages, the model exhibits unstable degrading curve. However, when the model already fine-tuned for Swedish is re-trained on all three languages with equal proportions of data, it is able to retain the ability to transcribe the data in Swedish. Slight decline can still be observed in terms of the performance on Swedish. Nevertheless, the model reaches nearly identical accuracy as compared to the monolingual setting. Surprisingly, the model trained from scratch performs comparably with the two aforementioned settings as well.

When it comes to Norwegian and Danish, all settings lead to steady improvement over training steps. The monolingual baselines demonstrate clear advantage over the multilingual models throughout the training. Initializing the multilingual model from a model already fine-tuned for Swedish seems to be rather effective overall, as all three re-training settings outperform the multilingual model trained from

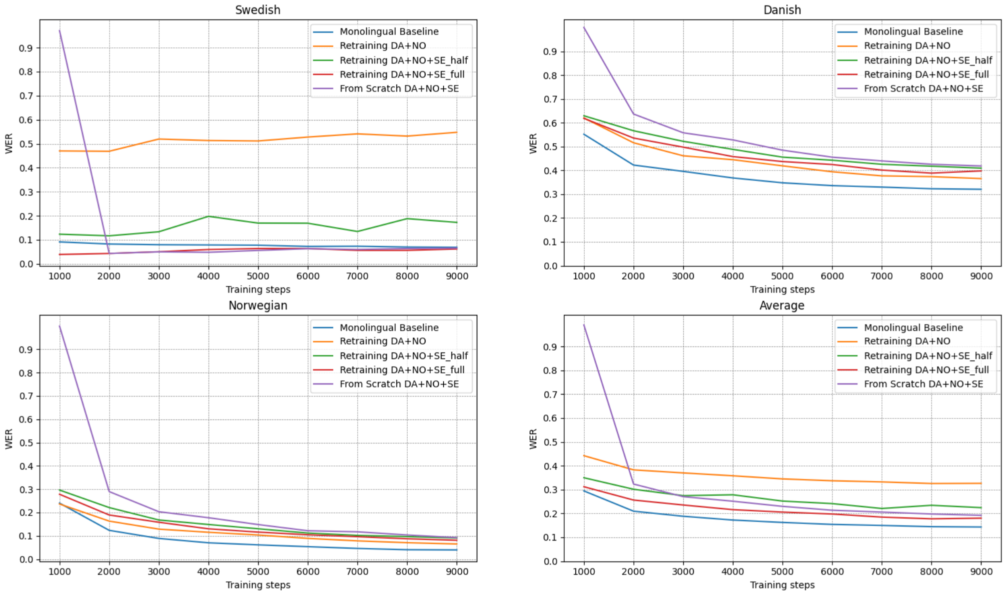


Figure 5.4.: Word Error Rates of the multilingual trial models and a monolingual baseline on the evaluation set, mapped over training steps. The model IDs are explained in Section 4.4.5.

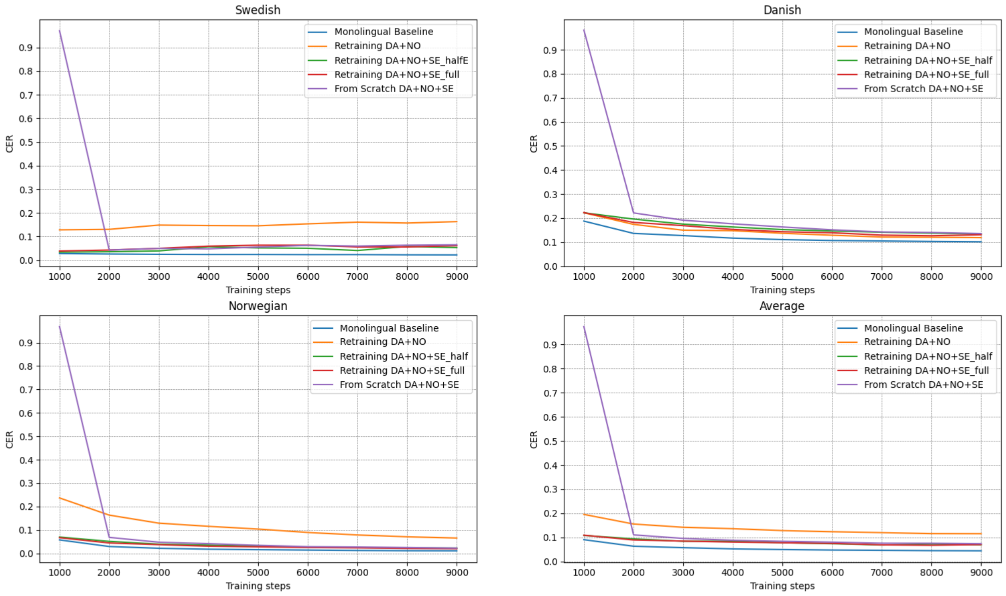


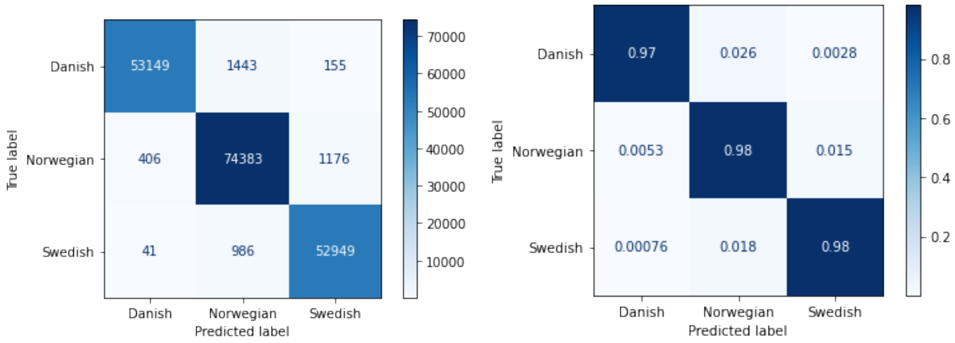
Figure 5.5.: Character Error Rates of the multilingual trial models and a monolingual baseline on the evaluation set, mapped over training steps. The model IDs are explained in Section 4.4.5.

scratch on both languages. Lastly, re-training the Swedish model on just Danish and Norwegian appears to be the most effective multilingual setting, likely to the least amount of cross-lingual noise encountered during training.

The first hypothesis for multilingual models expected the monolingual baselines to outperform the multilingual models in all languages. The statement holds true for Danish and Norwegian, yet the pattern breaks on Swedish. This is likely due to the overall high quality of the Swedish end-to-end model used to initialize the multilingual models. Even though its ability to transcribe Swedish degrades due to additional training on two more languages, it still remains slightly better than the monolingual baseline. The second hypothesis suggested that monolingual end-to-end models could be extended to handle multiple languages. The hypothesis is confirmed, as we the multilingual models initialized from monolingual Swedish end-to-end model reach quality comparable to that of monolingual models. Finally, the third hypothesis expected that the ability of a monolingual-turned-multilingual model to transcribe its initial target language would degrade should it not be exposed to data in the target language during training. The hypothesis can be confirmed, for we observed a decline in the ability to transcribe Swedish in the *Retraining DA+NO* and *Retraining DA+NO+SE_half* settings. This suggests that it is essential to ensure that the model is exposed to training data in all three languages throughout the training in order to enable accurate performance on the three directions. We therefore use this setting (*Retraining DA+NO+SE_full*) for our final end-to-end model.

5.2.2. Language Classification Module

We train a language classification module on randomly sampled and equally distributed subsets of the train sets in three languages for 41,000 updates. Evaluation of the language classification module is shown in Figure 5.6.



(a) Raw counts

(b) Normalized

Figure 5.6.: Confusion matrices of the language classifier's predictions on the test sets from all three languages.

The overall quality of the classification module is surprisingly high. It is able to identify the correct language tag with an accuracy of roughly 98%, yet the accuracy differs greatly depending on the duration of the audio sequence. As shown in Figure 5.7, the accuracy is especially low on very short sequences, approximately up to 1.5 seconds, whereas it is nearly flawless on sequences of at least 5.25 seconds. The short Sequences mostly contain the pronunciation of just one word or even one letter. The inability to predict the language from such short sequences might be linked to the lexical and phonetic similarities between the languages. Due to a considerable overlap in the lexicon of the three languages, prediction of the language from just one word might become a challenging task even to a human speaker. Nevertheless, perhaps because of a lesser overlap in the phonology and phonetics of the three languages, the classification module is still able to deduce the language more accurately than by chance. In addition, the errors made by the classifier appear more or less systematic, as the module confuses Swedish and Norwegian as well as Danish and Norwegian roughly ten times more frequently than Swedish and Danish. This could be due to the fact that the overlap of phonetic features between Danish and Swedish is smaller than the overlap between Swedish and Norwegian as well as Norwegian and Danish (Gooskens, 2007).

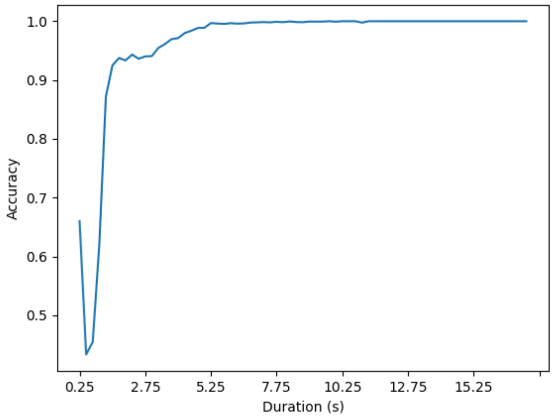


Figure 5.7.: The accuracy of the language classification module mapped over the quantized lengths of the entries in the test set.

5.2.3. Multilingual Scandinavian End-to-end ASR Model

The multilingual Scandinavian end-to-end ASR model combines the large VoxRex (version C) encoder described in Section 4.4.4, jointly fine-tuned CTC decoder (5.2.1), and the classification module (5.2.2). The complete parameters of the final end-to-end model are presented in Tables A.3 and A.4. We further evaluate the model on its ability to transcribe Swedish, Danish, and Norwegian, alongside its performance across the dialects of all three languages. Lastly, we perform a brief qualitative analysis of the most common errors made by the multilingual model in each language.

Performance on Swedish, Danish, and Norwegian

The performance of the joint multilingual end-to-end model on all three Scandinavian languages is shown in Table 5.3 below. The multilingual model is able to retain the ability to accurately transcribe the data in Swedish with some losses, in addition to learning to transcribe both Danish and Norwegian. Surprisingly, the multilingual model outperforms the monolingual counterparts on numerous parameters. For Danish and Norwegian, the multilingual model significantly outperforms the monolingual models in terms of WER and CER, regardless of whether it is boosed by a statistical language model or not. As for Swedish, the multilingual model performs better than the monolingual in terms of CER when a language model is used.

Given that the multilingual model is initialized on top of the Swedish end-to-end model, the variation in the performance across the languages does not come as a surprise. The low error rate on Swedish is mainly due to the fact that a high-quality Swedish end-to-end model is used as the base. The lower error rate on the

Table 5.3.: The performance of the multilingual end-to-end model versus the three monolingual end-to-end models described in Section 4.4. A side-by-side comparison of the results obtained with and without 4-gram language models in both settings. The better score between the monolingual and multilingual models is marked in bold.

		No LM		**With LM**	
Test Set	Model	WER	CER	WER	CER
Swedish	Monolingual[1]	**2.19%**	**0.98%**	**2.74%**	1.07%
	Multilingual	4.61%	1.16%	3.26%	**0.96%**
Danish	Monolingual[2]	19.14%	5.45%	13.82%	4.33%
	Multilingual	**12.69%**	**3.74%**	**10.43%**	**3.31%**
Norwegian	Monolingual[3]	16.47%	3.62%	12.03%	2.81%
	Multilingual	**9.64%**	**2.77%**	**6.51%**	**2.18%**

[1] *Wav2vec 2.0 large VoxRex Swedish (C)* described in Section 4.4.1.
[2] Our own monolingual end-to-end ASR model for Danish described in Section 4.4.3.
[3] *Norwegian Wav2Vec2 Model - 1B Bokmål* described in Section 4.4.2.

Norwegian as compared to Danish can perhaps be linked to the better performance of the Swedish model on the Norwegian language even in a zero-shot setting.

The competitive performance of the multilingual model as compared to the monolingual models is an unexpected outcome. There are several plausible explanations behind such optimistic results. Firstly, when it comes to Norwegian, the monolingual end-to-end Norwegian model was trained on data in a different domain. The multilingual model, on the other hand, is both trained and tested on the NST datasets, with the addition of the CV dataset used for validation. It can thus be argued that the better performance on Norwegian is made possible due to the training on in-domain data. When it comes to Danish, it is rather likely that the quality of the model improves simply due to the fact that the model is trained for more updates, as the monolingual model was trained for 4,000 updates, whereas the multilingual model is trained for 41,000 updates. In other words, the model is exposed to more data in general, including more Danish data. It should not be excluded that the joint training of a multilingual model improves the model's ability to generalize. Since the model is constantly exposed to batches of data in all three languages, it is less prone to overfitting to the training data.

The addition of a language model improves the average WER and CER for all three languages. Interestingly, in the multilingual setting, a language model no longer harms the quality of transcription in Swedish, as was the case with the monolingual model. Perhaps this is because the language model becomes more effective when the decoder of the ASR model is trained on more varied data.

All in all, the fourth hypothesis, which expected the multilingual model to lag behind the monolingual models, can be rejected. Despite the potential cross-lingual noise introduced from additional languages, the multilingual model not only matches, but also outperforms the monolingual models on Danish and Norwegian.

Performance on the Regional Dialects

The quality of the multilingual ASR model as compared to the monolingual models across the regional dialects of Swedish, Danish, and Norwegian is shown in Table 5.4. The improvement in the quality of the multilingual model upon Danish and Norwegian monolingual counterparts leads to better results across all regional dialects of the two languages. Without the addition of a language model, the multilingual ASR model is outperformed by the monolingual Swedish model on all dialects. However,

Table 5.4.: The performance of the monolingual models described in Section 4.1 versus the multilingual model across the regional dialects of Swedish, Danish, and Norwegian. The lowest WER per regional dialect is in bold, and the lowest CER is underlined.

	MONOLINGUAL MODELS				**MULTILINGUAL MODEL**			
	No LM		With LM		No LM		With LM	
	WER	CER	WER	CER	WER	CER	WER	CER
Östergötland	**1.68%**	0.66%	2.34%	0.77%	3.78%	0.85%	2.43%	0.62%
Västra sydsverige	**2.35%**	0.91%	2.97%	1.02%	7.74%	1.86%	5.21%	1.43%
Östra sydsverige	**2.13%**	0.85%	2.81%	0.97%	4.81%	1.15%	3.30%	0.91%
Västergötland	**2.07%**	0.78%	2.65%	0.87%	4.16%	0.93%	2.79%	0.71%
Göteborg med omnejd	**2.12%**	0.88%	2.63%	0.97%	3.90%	1.00%	2.81%	0.81%
Mellansverige	**2.02%**	0.76%	2.45%	0.84%	3.85%	0.86%	2.76%	0.70%
Västsverige	**1.77%**	0.68%	2.59%	0.82%	3.35%	0.73%	2.41%	0.59%
Stockholm med omnejd	**1.75%**	0.68%	2.28%	0.77%	3.05%	0.67%	2.21%	0.54%
Norrland	**2.16%**	0.86%	2.74%	0.96%	4.31%	1.05%	3.10%	0.86%
Dalarna med omnejd	**1.91%**	0.74%	2.16%	0.80%	3.92%	0.88%	2.74%	0.68%
Standard Deviation (σ)	0.202%	0.089%	0.240%	0.088%	1.238%	0.317%	0.805%	0.242%
Vest- og Sydsjælland	19.23%	5.31%	13.76%	4.16%	12.81%	3.63%	**10.51%**	3.17%
Storkøbenhavn	19.54%	5.62%	14.26%	4.52%	13.44%	4.03%	**11.10%**	3.58%
Østjylland	21.24%	6.08%	15.37%	4.84%	14.00%	4.13%	**11.52%**	3.64%
Nordjylland	18.73%	5.08%	13.53%	3.98%	12.14%	3.31%	**9.92%**	2.91%
Sønderjylland	19.76%	6.36%	14.40%	5.27%	12.80%	4.34%	**10.49%**	3.90%
Fyn	15.83%	4.19%	11.13%	3.17%	10.20%	2.77%	**8.27%**	2.41%
Vestjylland	18.45%	4.97%	12.98%	3.82%	11.37%	3.16%	**9.31%**	2.77%
Standard Deviation (σ)	1.528%	0.675%	1.237%	0.643%	1.192%	0.531%	1.022%	0.496%
Voss og omland	16.20%	3.43%	11.71%	2.63%	7.19%	1.66%	**4.59%**	1.17%
Sunnmøre	19.87%	4.36%	14.58%	3.37%	9.28%	2.18%	**5.94%**	1.54%
Hedmark og Oppland	14.32%	2.99%	10.40%	2.29%	9.33%	2.31%	**6.19%**	1.70%
Oslo-området	16.53%	5.03%	12.67%	4.39%	10.50%	4.19%	**7.85%**	3.73%
Ytre Oslofjord	17.36%	3.72%	12.64%	2.91%	11.41%	2.99%	**7.67%**	2.25%
Nordland	17.61%	3.70%	12.45%	2.81%	9.30%	2.32%	**5.87%**	1.65%
Trøndelag	14.08%	2.96%	10.34%	2.31%	7.62%	1.83%	**4.90%**	1.31%
Sør-Vestlandet	20.34%	4.82%	14.85%	3.78%	10.89%	2.79%	**7.21%**	2.10%
Bergen og Ytre Vestland	21.37%	5.21%	15.20%	4.03%	11.07%	3.10%	**6.92%**	2.29%
Sørlandet	17.65%	3.69%	12.70%	2.82%	9.78%	2.40%	**6.49%**	1.76%
Troms	16.79%	3.32%	12.19%	2.57%	7.64%	1.75%	**5.10%**	1.28%
Standard Deviation (σ)	2.211%	0.764%	1.554%	0.677%	1.400%	0.703%	1.066%	0.686%

when boosted by a language model, the multilingual model achieves better CER on most Swedish regional dialects, and lower WER on *Västsverige* and *Stockholm med omnejd*.

As formulated in the fourth hypothesis for the multilingual models, we expected the performance of the multilingual model to be more stable across the regional dialects of the Scandinavian languages. This appears to hold true for Danish and Norwegian, as the standard deviations between the regional dialects are smaller in the case of multilingual models. The multilingual setting though leads to a substantial increase in the variance of the performance on the regional dialects of Swedish. This is mostly due to a single outlier, namely the *Västra sydsverige* regional dialect, which appears to be most challenging to both monolingual and multilingual models. The formulated hypothesis thus cannot be confirmed, as the results differ across languages.

Qualitative Analysis

The most common errors made by the multilingual model in Swedish, Danish, and Norwegian are shown in Table 5.5. Examples of the transcriptions produced by the multilingual model in Swedish, Danish, and Norwegian, with and without a language model, are show in Table B.5. The preposition *i* appears to be by far the

Table 5.5.: A side-by-side comparison of the most common insertion, deletion, and substitution errors in the transcriptions produced by the multilingual model, with and without a language model. The number in the parentheses is the raw frequency of the error. For substitutions, the token left of the arrow is the target and right of the arrow is the prediction.

No Language Model			With Language Model		
Insertion	Deletion	Substitution	Insertion	Deletion	Substitution
SWEDISH					
i (183)	i (118)	det → de (263)	i (192)	i (95)	de → det (237)
det (47)	att (41)	de → det (261)	det (56)	och (54)	det → de (219)
en (34)	är (40)	är → er (132)	en (36)	att (50)	skall → ska (119)
är (30)	och (38)	skall → ska (108)	för (31)	är (48)	är → er (75)
för (28)	en (36)	ska → skal (101)	är (31)	en (44)	istället → stället (52)
DANISH					
er (201)	i (358)	åbn → åben (238)	er (239)	i (381)	åbn → åben (235)
i (130)	er (279)	as → a (162)	i (174)	er (289)	as → a (223)
at (130	en (242)	det → de (153)	en (160)	en (221)	har → er (172)
en (128)	at (167)	har → er (140)	at (119)	at (209)	æ → a (146)
det (76)	det (165)	i → e (139)	for (89)	af (138)	det → de (137)
NORWEGIAN					
i (184)	i (475)	én → en (263)	i (137)	i (438)	har → er (220)
for (104)	til (341)	ett → et (212)	for (101)	er (315)	et → ett (127)
en (83)	er (313)	det → de (188)	en (96)	til (304)	én → en (124)
det (65)	de (285)	og → å (174)	og (60)	det (286)	og → å (122)
of (60)	en (236)	har → er (173)	det (58)	excel (242)	det → de (111)

most problematic, as it is the most common deletion in all three languages both with and without a language model. In addition, it is the most common insertion error in both Swedish and Norwegian, both with and without the addition of a language model. In the case of Danish, however, the auxiliary verb *er* is incorrectly inserted even more frequently than the aforementioned preposition. In addition, the verb is frequently erroneously deleted by the multilingual model.

Overall, the most frequent errors made by the multilingual model appear highly similar to those made by the monolingual models on their target languages reported in Table 5.2. The model still mostly struggles with function words, namely prepositions, conjunctions, and auxiliary verbs. In addition, some of the most common the substitution errors are only considered to be errors due to the limited ability of the evaluation metric to permit alternative spellings of the word, e.g. *skall* versus *ska*; *åbn* versus *åben*; *én* versus *en*. Interestingly, the multilingual setting also resolves the inability of the Norwegian model to predict individual characters, mostly occurring in spelled out words.

As per the fifth hypothesis for multilingual models, it was expected that the multilingual setting would lead to frequent confusion of the spellings across languages. Judging from the error patterns, the hypothesis can arguably be rejected, as the vast majority of errors remain language-specific.

6. Conclusions

We examined monolingual, cross-lingual, and multilingual models and applications of ASR systems for Scandinavian languages. We extensively evaluated the performance of state-of-the-art Swedish, Danish, and Norwegian Wav2Vec 2.0-based ASR models on their target language as well as other Scandinavian languages. In addition, we analysed the effectiveness of statistical language models on the quality of transcriptions. We also proposed an architecture of a multilingual ASR model for Scandinavian languages. The proposed architecture incorporates a language classification module and thus enables the usage of language-specific language models. Lastly, we trained and evaluated a multilingual ASR model for Swedish, Danish, and Norwegian.

Key findings offer insight into the applicability of zero-shot transfers and bias towards dialects in ASR models, as well as the effectiveness of a multilingual approach. We learned that in a zero-shot setting, ASR models trained a Scandinavian language serve as a better source language than a highly accurate English model, except for the Danish model. From a practical standpoint, however, zero-shot transferability of ASR models is questionable, for the quality of the produced transcriptions is poor. We also found common patterns between the mutual intelligibility levels of the Scandinavian languages in native speakers and monolingual ASR models. Overall, monolingual ASR models perform better on the languages that are easier to comprehend for native speakers. We identified differences in the performance of both monolingual and multilingual ASR models across regional dialects of the three languages. In a multilingual setting, the performance across the regional dialects is more stable overall. Lastly, we trained an accurate multilingual end-to-end ASR model for Swedish, Danish, and Norwegian. Lastly, we confirmed that multilingual ASR models can perform comparably with their monolingual counterparts. Our multilingual model performed worse than the monolingual Swedish model. Nevertheless, the multilingual model outperformed the monolingual Danish and Norwegian ASR models.

The following conclusions can be made in addressing the four research questions formulated for this thesis project:

1. *Are monolingual Wav2Vec 2.0-based ASR models transferable across Scandinavian languages?*

   The word and character error rates observed in a zero-shot setting suggest that Wav2Vec 2.0-based ASR models are indeed transferrable across Scandinavian languages. The Swedish and Norwegian ASR models performed better on all three languages than the English baseline; however, the English baseline outperformed the Danish model on Swedish and Norwegian. Given the relatively poor quality of the Danish ASR model, it can be concluded that a high quality of ASR models serves as a pre-condition for the transferability of ASR models across the Scandinavian languages.

2. *Are there parallels between zero-shot transferability of monolingual ASR models and mutual intelligibility of Swedish, Danish, and Norwegian?*

Monolingual Scandinavian ASR models, when used in a cross-lingual setting, exhibit patterns similar to the comprehension levels in native speakers. Similarly to native speakers of Swedish, a Swedish ASR model achieves better results on Norwegian data than on Danish. The Danish model also favours Norwegian over Swedish, as do native speakers of Danish. Native speakers of Norwegian reportedly understand the Swedish language better than Danish, which seems to also be the case with Norwegian ASR models.

3. Are Wav2Vec 2.0-based ASR models biased towards one of the dialects of their target language?

   The performance of Wav2Vec 2.0-based ASR models differs across the regional dialects of the Scandinavian languages. The degree of variance differs across the three languages. The Swedish ASR model performed best on the *Östergötland* regional dialect, the Danish model achieved the lowest error rates on *Fyn*, and the Norwegian model performed nearly identically well on *Hedmark oh Oppland* and *Trøndelag* regional dialects. It is though difficult to conclude whether any of the models is strictly biased towards one dialect, such as the standard speech. It cannot be excluded that the differences in the observed performance stem not only from the quality of models, but also the imbalances in the test sets. Thus, more extensive research is required to answer this question with more confidence.

4. *Can a monolingual Wav2Vec 2.0-based ASR model be extended to handle all three Scandinavian languages without significant loss of quality?*

   The multilingual Scandinavian ASR model trained and evaluated in this thesis project outperforms the monolingual models on Danish and Norwegian, and performs relatively comparably with the monolingual Swedish model. It can therefore be concluded that monolingual ASR models can be extended to handle multiple languages. Such practice allows to utilize and extend existing high-quality monolingual models rather than scraping them and training new models from scratch. Nevertheless, it cannot be ruled out that this approach is only viable due to the fact that the Scandinavian languages are highly similar. In other words, extending a monolingual ASR model to also handle typologically different language might lead to substantially different results.

We based our research on Swedish, Danish, and Norwegian due to the considerable similarities shared between Scandinavian languages. We cannot rule out the possibility that the positive results observed in the multilingual setting would not hold should the methodology be applied to other, less similar language groups, such as the Slavic or Indian languages. Furthermore, we tested the models on comparable datasets from similar domains in the three languages, which might have also affected the results positively. Further research could explore the methodology in a more realistic setting. More directions for further research are formulated in the following subsection.

6.1. Future Work

The work carried out in this thesis can be expanded though both linguistic and technical aspects. As such, further research could explore the possibility of further expansion of the multilingual Scandinavian ASR models to also include Faroese and Icelandic. Alternatively, the models can also be expanded to include a less similar language, possibly with a different script, to learn how this would affect the acquisition of a new language as well as retention of the already learned languages.

From a technical point of view, the possibilities for further research are truly endless, given the access to the computational resources. An almost certainly successful extension could replace the statistical language model with a Transformer-based language model, as done by Baevski et al. (2020). The Transformer-based language model can also be turned multilingual, thus eliminating the need for a language classification module. Furthermore, it would be interesting to learn how exposure of the model to all three languages in the pre-training step would affect the end results, as the encoder model used by us is only pre-trained on Swedish. Lastly, one could exploit the language classification module to enable accurate processing of code-switching, i.e. instances where a speaker switches from one language to another. This could perhaps be done by training the module to classify smaller chunks of audio and select the corresponding language model.

Bibliography

Ali, Ahmed and Steve Renals (2018). “Word Error Rate Estimation for Speech Recognition: e-WER”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 20–24. DOI: 10.18653/v1/P18-2004. URL: https://aclanthology.org/P18-2004.

Amodei, Dario, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Vaino Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu (2016). “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin”. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML’16. New York, NY, USA: JMLR.org, pp. 173–182.

Ardila, Rosana, 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”. English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.520.

Baevski, Alexei, 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, pp. 12449–12460.

Barroso, Nora, Odei Barroso, Unai Susperregi, Aitzol Ezeiza, and Karmele López de Ipiña (2010). “Language identification oriented to Multilingual Speech Recognition in the Basque context”. In: *2010 IEEE 15th Conference on Emerging Technologies Factory Automation (ETFA 2010)*, pp. 1–8. DOI: 10.1109/ETFA.2010.5641193.

Basbøll, Hans (2005). *The phonology of Danish*. Oxford: OUP Oxford.

Bashori, Muzakki, Roeland van Hout, Helmer Strik, and Catia Cucchiarini (2021). “Effects of ASR-based websites on EFL learners’ vocabulary, speaking anxiety, and language enjoyment”. *System* 99, p. 102496. ISSN: 0346-251X. DOI: https://doi.org/10.1016/j.system.2021.102496. URL: https://www.sciencedirect.com/science/article/pii/S0346251X21000506.

Benesty, J, J Chen, and Y Huang (2008). *Automatic speech recognition: A deep learning approach*. Berlin: Springer.

Bhable, Suvarnsing G. and Charansing N. Kayte (2020). “Review: Multilingual Acoustic modeling of Automatic Speech Recognition(ASR) for low resource languages". In: *2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI)*, pp. 1–4. DOI: 10.1109/ICATMRI51801.2020.9398431.

Carrier, Michael (2017). "Automated Speech Recognition in language learning: Potential models, benefits and impact". *Training Language and Culture* 1 (Feb. 2017), pp. 46–61. DOI: 10.29366/2017tlc.1.1.3.

Caubrière, Antoine, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin (2020). "Where are we in Named Entity Recognition from Speech?" In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4514–4520. URL: https://aclanthology.org/2020.lrec-1.556.

Chung, Yu-An, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu (2021). "w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training". *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 244–250.

Chung, Cindy and James W Pennebaker (2007). "The psychological functions of function words". *Social communication* 1, pp. 343–359.

Cinque, Giglielmo and Richard S Kayne (2005). *The Oxford handbook of comparative syntax*. Oxford: Oxford University Press.

Clausen, Sara J. and Line B. Kristensen (2015). "The cognitive status of stød". English. *Nordic journal of linguistics* 38.2, pp. 163–187.

Cohen, P., S. Dharanipragada, J. Gros, M. Monkowski, C. Neti, S. Roukos, and T. Ward (1997). "Towards a universal speech recognizer for multiple languages". In: *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pp. 591–598. DOI: 10.1109/ASRU.1997.659140.

Conneau, Alexis, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli (2021). "Unsupervised Cross-Lingual Representation Learning for Speech Recognition". In: *Proc. Interspeech 2021*, pp. 2426–2430. DOI: 10.21437/Interspeech.2021-329.

Conneau, Alexis and Guillaume Lample (2019). "Cross-lingual Language Model Pretraining". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

Dahl, George, Dong Yu, li Deng, and Alex Acero (2012). "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition". *Audio, Speech, and Language Processing, IEEE Transactions on* 20 (Feb. 2012), pp. 30–42. DOI: 10.1109/TASL.2011.2134090.

Delsing, Lars-Olof (2005). *Håller språket ihop Norden?: en forskningsrapport om ungdomars förståelse av danska, svenska och norska [Does the Language Keep the Nordic Countries Together? A Research Report on How Well Young People Understand Danish, Swedish and Norwegian]*. Copenhagen: Nordic Council of Ministers. DOI: 10.6027/TN2005-573.

Devlin, Jacob, 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Elert, Claes-Christian (2000). *Allmän och svensk fonetik [General and Swedish Phonetics]*. Stockholm: Almqvist Wiksell.

Furui, Sadaoki (1999). "Automatic Speech Recognition and Its Application to Information Extraction". In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland, USA: Association for Computational Linguistics, June 1999, pp. 11–20. doi: 10.3115/1034678.1034680. url: https://aclanthology.org/P99-1002.

Garofolo, J., Lori Lamel, W. Fisher, Jonathan Fiscus, D. Pallett, N. Dahlgren, and V. Zue (1992). "TIMIT Acoustic-phonetic Continuous Speech Corpus". *Linguistic Data Consortium* (Nov. 1992).

Gooskens, Charlotte (2007). "The contribution of linguistic factors to the intelligibility of closely related languages". *Journal of Multilingual and multicultural development* 28.6, pp. 445–467.

Graves, Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber (2006). "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376.

Grønnum, Nina (1998). "Danish". *Journal of the International Phonetic Association* 28.1-2, pp. 99–105. doi: 10.1017/S0025100300006290.

Grønnum, Nina, Miguel Vazquez-Larruscan, and Hans Basbøll (2013). "Danish stød: laryngealization or tone". *Phonetica* 70.1-2, pp. 66–92.

Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang (2020). "Conformer: Convolution-augmented Transformer for Speech Recognition". *ArXiv* abs/2005.08100.

Habeeb, Imad Qasim, Hanan Najm Abdulkhudhur, and Zeyad Qasim Al-Zaydi (2021). "Three N-grams Based Language Model for Auto-correction of Speech Recognition Errors". In: *International Conference on New Trends in Information and Communications Technology Applications*. Springer, pp. 131–143.

Haidar, Md. Akmal, Chao Xing, and Mehdi Rezagholizadeh (2021). "Transformer-Based ASR Incorporating Time-Reduction Layer and Fine-Tuning with Self-Knowledge Distillation". In: *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*. Ed. by Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlcek. ISCA, pp. 2102–2106. doi: 10.21437/Interspeech.2021-1743. url: https://doi.org/10.21437/Interspeech.2021-1743.

Håkansson, Alexander and Kevin Hoogendijk (2020). "Transfer learning for domain specific automatic speech recognition in Swedish: An end-to-end approach using Mozilla's DeepSpeech". MA thesis. Chalmers tekniska högskola / Institutionen för data och informationsteknik.

Hannun, Awni (2017). "Sequence Modeling with CTC". *Distill*. doi: 10.23915/distill.00008.

Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng (2014). *Deep Speech: Scaling up end-to-end speech recognition*. doi: 10.48550/ARXIV.1412.5567. url: https://arxiv.org/abs/1412.5567.

Harbert, Wayne (2006). *The Germanic Languages*. Cambridge: Cambridge University Press.

Haugen, Einar Ingvald (1982). *Scandinavian language structures: a comparative historical survey*. Vol. 5. Niemeyer.

Heigold, G., V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean (2013). "Multilingual acoustic models using distributed deep neural networks". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8619–8623. DOI: 10.1109/ICASSP.2013.6639348.

Hsu, Wei-Ning, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli (2021). "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training". In: *Interspeech*.

Huang, Jui Ting, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong (2013). "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers". *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7304–7308.

Jensen, Anita, Theodore Stickley, Wenche Torrissen, and Kjerstin Stigmar (2017). "Arts on prescription in Scandinavia: a review of current practice and future possibilities". *Perspectives in Public Health* 137.5. PMID: 27852837, pp. 268–274. DOI: 10.1177/1757913916676853. eprint: https://doi.org/10.1177/1757913916676853. URL: https://doi.org/10.1177/1757913916676853.

Johannessen, Janne, Andre Kåsen, Kristin Hagen, Anders Nøklestad, and Joel Priestley (2020). "Comparing methods for measuring dialect similarity in Norwegian". In: *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 5343–5350.

Kahn, J., M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux (2020). "Libri-Light: A Benchmark for ASR with Limited or No Supervision". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673. DOI: 10.1109/ICASSP40776.2020.9052942.

Kamath, Uday, John Chih Liu, and James Whitaker (2019). "Deep Learning for NLP and Speech Recognition". In: *Springer International Publishing*.

Khassanov, Yerbolat (2020). "Language model domain adaptation for automatic speech recognition systems". PhD thesis. Nanyang Technological University.

Kristoffersen, Gjert (2000a). *The phonology of Norwegian*. English. New York: Oxford University Press. ISBN: 9780198237655;0198237650;9780191543937;0191543934;

Kristoffersen, Gjert (2000b). *The phonology of Norwegian*. Oxford: OUP Oxford.

Lauscher, Anne, Vinit Ravishankar, Ivan Vuli, and Goran Glava (2020). "From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4483–4499. DOI: 10.18653/v1/2020.emnlp-main.363. URL: https://aclanthology.org/2020.emnlp-main.363.

Lee, Taewoo, Min-Joong Lee, Tae Gyoon Kang, Seokyeoung Jung, Minseok Kwon, Yeona Hong, Jungin Lee, Kyoung-Gu Woo, Ho-Gyeong Kim, Jiseung Jeong, Jihyun Lee, Hosik Lee, and Young Sang Choi (2021). "Adaptable Multi-Domain Language Model for Transformer ASR". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7358–7362. DOI: 10.1109/ICASSP39728.2021.9413475.

Leinonen, Therese Nanette (2010). "An acoustic analysis of vowel pronunciation in Swedish dialects". PhD thesis. Rijksuniversiteit Groningen.

Li, Bo, Shuo-yiin Chang, Tara N. Sainath, Ruoming Pang, Yanzhang He, Trevor Strohman, and Yonghui Wu (2020). "Towards Fast and Accurate Streaming End-To-End ASR". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6069–6073. DOI: 10.1109/ICASSP40776.2020.9054715.

Li, Jason, Ravi Teja Gadde, Boris Ginsburg, and Vitaly Lavrukhin (2018). “Training Neural Speech Recognition Systems with Synthetic Speech Augmentation”. *ArXiv* abs/1811.00707.

Lin, Hui, Li Deng, Dong Yu, Yi-fan Gong, Alex Acero, and Chin-Hui Lee (2009). “A study on multilingual acoustic modeling for large vocabulary ASR”. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4333–4336. DOI: 10.1109/ICASSP.2009.4960588.

Lindahl-Jacobsen, Rune, My Euler, Merete Osler, Elsebeth Lynge, and Niels Keiding (2004). “Women’s death in Scandinavia - What makes Denmark different?” *European journal of epidemiology* 19 (Feb. 2004), pp. 117–21. DOI: 10.1023/B:EJEP.0000017834.35943.bd.

Liu, Danyang, Ji Xu, Pengyuan Zhang, and Yonghong Yan (2021). “A unified system for multilingual speech recognition and language identification”. *Speech Communication* 127, pp. 17–28. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2020.12.008. URL: https://www.sciencedirect.com/science/article/pii/S0167639320303125.

Lyu, Dau-Cheng and Ren-Yuan Lyu (2008). “Language identification on code-switching utterances using multiple cues”. In: *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*. Jan. 2008, pp. 711–714.

Mabokela, Koena Ronny and Madimetja Jonas Manamela (2013). “An integrated language identification for code- switched speech using decoded-phonemes and support vector machine”. In: *2013 7th Conference on Speech Technology and Human - Computer Dialogue (SpeD)*, pp. 1–6. DOI: 10.1109/SpeD.2013.6682661.

Mæhlum, Brit and Unn Røyneland (2012). *Det norske dialektlandskapet [The Norwegian Dialect Regions]*. Oslo, Norway: Cappelen Damm Akademisk.

Malmsten, Martin, Chris Haffenden, and Love Börjeson (2022). “Hearing voices at the National Library–a speech corpus and acoustic model for the Swedish language”. *arXiv preprint arXiv:2205.03026*.

Nygård, Mikael, Camilla Härtull, Annika Wentjärvi, and Susanne Jungerstam (2017). “Poverty and old age in Scandinavia: A problem of gendered injustice? Evidence from the 2010 GERDA Survey in Finland and Sweden”. *Social Indicators Research* 132.2, pp. 681–698.

Ortmanns, S. and Hermann Ney (2000). “The time-conditioned approach in dynamic programming search for LVCSR”. *Speech and Audio Processing, IEEE Transactions on* 8 (Dec. 2000), pp. 676–687. DOI: 10.1109/89.876301.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur (2015a). “Librispeech: An ASR corpus based on public domain audio books”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur (2015b). “Librispeech: An ASR corpus based on public domain audio books”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur (2015c). “Librispeech: An ASR corpus based on public domain audio books”. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210.

Perero-Codosero, Juan M., Fernando M. Espinoza-Cuadros, and Luis A. Hernández-Gómez (2022). “A Comparison of Hybrid and End-to-End ASR Systems for the IberSpeech-RTVE 2020 Speech-to-Text Transcription Challenge”. *Applied Sci-*

ences 12.2. ISSN: 2076-3417. DOI: 10.3390/app12020903. URL: https://www.mdpi.com/2076-3417/12/2/903.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. (2011). "The Kaldi speech recognition toolkit". In: *IEEE 2011 workshop on automatic speech recognition and understanding*. CONF. IEEE Signal Processing Society.

Pratap, Vineel, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert (2020). "Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters". In: Oct. 2020, pp. 4751–4755. DOI: 10.21437/Interspeech.2020-2831.

Riad, Tomas (2006). *Scandinavian accent typology*. Vol. 59. 1. Oldenbourg Wissenschaftsverlag GmbH, pp. 36–55.

Robinson, David Earle (2003). "Neolithic and Bronze Age Agriculture in Southern Scandinavia – Recent Archaeobotanical Evidence from Denmark". *Environmental Archaeology* 8.2, pp. 145–165. DOI: 10.1179/env.2003.8.2.145. eprint: https://doi.org/10.1179/env.2003.8.2.145. URL: https://doi.org/10.1179/env.2003.8.2.145.

Sahlgren, Magnus, Fredrik Carlsson, Fredrik Olsson, and Love Börjeson (2021). "It's Basically the Same Language Anyway: the Case for a Nordic Language Model". In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, May 2021, pp. 367–372. URL: https://aclanthology.org/2021.nodalida-main.39.

Sanders, Ruth H (2021). *The Languages of Scandinavia: Seven Sisters of the North*. Chicago: University of Chicago Press. URL: https://www.researchgate.net/publication/322447601_The_Languages_of_Scandinavia_Seven_Sisters_of_the_North_University_of_Chicago_Press_2017.

Schneider, Steffen, Alexei Baevski, Ronan Collobert, and Michael Auli (2019). "wav2vec: Unsupervised pre-training for speech recognition". *arXiv preprint arXiv:1904.05862*.

Schötz, Susanne and Gösta Bruce (2009). "Phrase initial accent I in South Swedish". *Proceedings of Fonetik 2009*, pp. 42–47.

Schultz, T. and A. Waibel (2000). "Polyphone decision tree specialization for language adaptation". In: *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*. Vol. 3, 1707–1710 vol.3. DOI: 10.1109/ICASSP.2000.862080.

Sen, Soumya, Anjan Dutta, and Nilanjan Dey (2018). *Audio Processing and Speech Recognition: Concepts, Techniques and Research Overviews*. Dec. 2018. ISBN: ISBN-10: 9811360979. DOI: 10.1007/978-981-13-6098-5.

Skrzypek, Dominika (2009). "The Formation of the Definite Article in the Nordic Languages". *Lingua Posnaniensis* 51 (Jan. 2009), pp. 65–76. DOI: 10.2478/v10122-009-0005-y.

Smith, Aaron, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne (2018). "82 Treebanks, 34 Models: Universal Dependency Parsing with Multi-Treebank Models". In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels: Association for Computational Linguistics, pp. 113–123. DOI: 10.18653/v1/K18-2011. URL: https://aclanthology.org/K18-2011.

Sturtevant, Dean G. (1989). "A Stack Decoder for Continous Speech Recognition". In: *Proceedings of the Workshop on Speech and Natural Language*. HLT '89. Cape Cod, Massachusetts: Association for Computational Linguistics, pp. 193–198. ISBN: 1558601120. DOI: 10.3115/1075434.1075466. URL: https://doi.org/10.3115/1075434.1075466.

Synnaeve, Gabriel, Qiantong Xu, Jacob Kahn, Edouard Grave, Tatiana Likhomanenko, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert (2019). "End-to-end ASR: from Supervised to Semi-Supervised Learning with Modern Architectures". *ICML 2020 Workshop SAS* abs/1911.08460.

Tanaka, Tomohiro, Ryo Masumura, Takafumi Moriya, Takanobu Oba, and Yushi Aono (2019). "A Joint End-to-End and DNN-HMM Hybrid Automatic Speech Recognition System with Transferring Sharable Knowledge". In: Sept. 2019, pp. 2210–2214. DOI: 10.21437/Interspeech.2019-2263.

Tian, Jinchuan, Jianwei Yu, Chao Weng, Yuexian Zou, and Dong Yu (2022). "Improving Mandarin End-to-End Speech Recognition with Word N-gram Language Model". *IEEE Signal Processing Letters*, pp. 1–1. DOI: 10.1109/LSP.2022.3154241.

Turan, Mehmet Ali Tugtekin, Emmanuel Vincent, and Denis Jouvet (2020). "Achieving Multi-Accent ASR via Unsupervised Acoustic Model Adaptation". In: *Proc. Interspeech 2020*, pp. 1286–1290.

Van Riemsdijk, Henk (2011). *Clitics in the Languages of Europe*. Vol. 20. 5. Berlin: Walter de Gruyter.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wabakken, Petter, Håkan Sand, Olof Liberg, and Anders Bjärvall (2001). "The recovery, distribution, and population dynamics of wolves on the Scandinavian peninsula, 1978-1998". *Canadian Journal of Zoology* 79.4, pp. 710–725. DOI: 10.1139/z01-029.

Wang, Song and Guanyu Li (2019). "Overview of end-to-end speech recognition". *Journal of Physics: Conference Series* 1187 (Apr. 2019), p. 052068. DOI: 10.1088/1742-6596/1187/5/052068.

Wessén, Elias (1954). *Våra folkmål [Our Dialects]*. Stockholm: Fritze.

Wu, Shijie and Mark Dredze (2019). "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 833–844. DOI: 10.18653/v1/D19-1077. URL: https://aclanthology.org/D19-1077.

Xu, Qiantong, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli (2021). "Self-Training and Pre-Training are Complementary for Speech Recognition". *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3030–3034.

Yu, Fu-Hao and Kuan-Yu Chen (2021). "Non-autoregressive Transformer-based End-to-end ASR using BERT". *arXiv preprint arXiv:2104.04805*.

A. Model Parameters

A.0.1. Monolingual Danish Model

Table A.1.: The Encoder parameters of the monolingual Danish model. The parameter names are from the Hugging Face implementation. The parameters are directly adopted from the pre-trained model for Danish described in Section 4.4.3.

Parameter Name	**Parameter**
activation_dropout	0.1
apply_spec_augmente	True
attention_dropout	0.1
bos_token_id	1
classifier_proj_size	256
codevector_dim	256
contrastive_logits_temperature	0.1
conv_bias	False
conv_dim	512, 512, 512, 512, 512, 512, 512
conv_kernel	10, 3, 3, 3, 3, 3, 2
conv_stride	5, 2, 2, 2, 2, 2, 2
ctc_loss_reduction	sum
ctc_zero_infinity	False
diversity_loss_weight	0.1
do_stable_layer_norm	False
feat_extract_activation	GeLU
feat_extract_norm	group
feat_proj_dropout	0.0
feat_quantizer_dropout	0.0
final_dropout	0.1
hidden_act	GeLU
hidden_dropout	0.1
hidden_size	768
initializer_range	0.02
intermediate_size	3072
layer_norm_eps	1e-05
layerdrop	0.1
mask_feature_length	10
mask_feature_prob	0.0
mask_time_length	10
mask_time_prob	0.05
model_type	wav2vec2
num_attention_heads	12
num_codevector_groups	2
num_codevectors_per_group	320
num_conv_pos_embedding_groups	16
num_conv_pos_embeddings	128
num_feat_extract_layers	7
num_hidden_layers	12
num_negatives	100
pad_token_id	32
proj_codevector_dim	256
torch_dtype	float32
transformers_version	4.11.3
use_weighted_layer_sum	False
vocab_size	33

A.0.2. Multilingual Scandinavian Model

Table A.2.: The Decoder parameters of the monolingual Danish model. The parameter names are from the Hugging Face implementation.

Parameter Name	**Parameter**
group_by_length	True
learning_rate	0.0001
gradient_accumulation_steps	3
per_device_train_batch_size	10
evaluation_strategy	steps
num_train_epochs	10
fp16	True
gradient_checkpointing	True
save_steps	500
eval_steps	500
logging_steps	500
weight_decay	0.005
warmup_steps	1000
attention_dropout	0.1
hidden_dropout	0.1
feat_proj_dropout	0.0
mask_time_prob	0.05
layerdrop	0.1
ctc_loss_reduction	sum
vocab_size	33

Table A.3.: The Encoder parameters of the final multilingual model. The parameter names are from the Hugging Face implementation. The parameters are directly adapted from the Swedish *large VoxRex (version C)* pre-trained model, described in Section 4.4.1.

Parameter Name	**Parameter**
activation_dropout	0.05
apply_spec_augmente	True
attention_dropout	0.1
bos_token_id	1
classifier_proj_size	256
codevector_dim	256
contrastive_logits_temperature	0.1
conv_bias	False
conv_dim	512, 512, 512, 512, 512, 512, 512
conv_kernel	10, 3, 3, 3, 3, 3, 2
conv_stride	5, 2, 2, 2, 2, 2, 2
ctc_loss_reduction	mean
ctc_zero_infinity	True
diversity_loss_weight	0.1
do_stable_layer_norm	True
eos_token_id	2
feat_extract_activation	GeLU
feat_extract_norm	layer
feat_proj_dropout	0.05
feat_quantizer_dropout	0.0
final_dropout	0.0
hidden_act	GeLU
hidden_dropout	0.05
hidden_size	1024
initializer_range	0.02
intermediate_size	4096
layer_norm_eps	0.00001
layerdrop	0.05
mask_channel_selection	static
mask_feature_length	10
mask_feature_prob	0.1
mask_time_length	10
mask_time_prob	0.05
mask_channel_selection	static
model_type	wav2vec2
num_adapter_layers	3
num_attention_heads	16
num_codevector_groups	2
num_codevectors_per_group	320
num_conv_pos_embedding_groups	16
num_conv_pos_embeddings	128
num_feat_extract_layers	7
num_hidden_layers	24
num_negatives	100
output_hidden_size	1024
pad_token_id	32
proj_codevector_dim	256
tdnn_dilation	1, 2, 3, 1, 1
tdnn_dim	512, 512, 512, 512, 1500
tdnn_kernel	5, 3, 3, 1, 1
torch_dtype	float32
transformers_version	4.17.0
use_weighted_layer_sum	False
vocab_size	46

Table A.4.: The Decoder parameters of the final multilingual model. The parameter names are from the Hugging Face implementation.

Parameter Name	**Parameter**
group_by_length	True
learning_rate	0.0001
gradient_accumulation_steps	3
per_device_train_batch_size	6
evaluation_strategy	steps
num_train_epochs	20
fp16	True
gradient_checkpointing	True
save_steps	1000
eval_steps	1000
logging_steps	500
weight_decay	0.005
warmup_steps	500
attention_dropout	0.1
hidden_dropout	0.05
feat_proj_dropout	0.05
mask_time_prob	0.05
layerdrop	0.05
ctc_loss_reduction	mean
vocab_size	46

B. Complete Results

B.1. Monolingual Models

Table B.1.: An example of the transcriptions of a Swedish entry produced by the four monolingual models described in Section 4.1, with and without a Norwegian language model.

Target		samverkande krafter kommer att hålla förändringarna på måttlig nivå
Swedish Model	no LM	samverkande krafter kommer att hålla förändringarna på måttlig nivå
	with LM	samverkande krafter kommer att hålla förändringarna på måttlig nivå
Danish Model	no LM	sarm hvad et gamte kendaf de gå men op på lo fettaanditingerne boe mort gliniveuer
	with LM	samt vad e gamle kända e går men opålofetaanditingerne bo mår gliniveuer
Norwegian Model	no LM	sanverkenee krefter kommer at tolde forendringene på måtlig nivå
	with LM	samverkan krefter kommer torde forendringene på måtlig nivå
English Model	no LM	somver cante craft e comnato la feren dring ana po motlinivo
	with LM	som var cantecrafte comnatolaferen ring anapomotlinivo

Table B.2.: An example of the transcriptions of a Danish entry produced by the four monolingual models described in Section 4.1, with and without a Danish language model.

Target		hvad synes du om den nye bank der åbnede i går dernede på hjørnet over for kirken
Swedish Model	no LM	värsund storemt nybank doch uppnigår den nu på gönare oc förkering
	with LM	väsunstemtnybank de uppnigår den nu på gönareocförkering
Danish Model	no LM	hvad synes du om den nye bank der åbnede i går dernede på hjørnet over for kirken
	with LM	hvad synes du om den nye bank der åbnede i går dernede på hjørnet over for kirken
Norwegian Model	no LM	hva synes dømt ne pank til åpnet e godne de på hjørnet a og for keangn
	with LM	ve synes dømt nu bank til op god nede på hjørnet og for kan
English Model	no LM	henlibud ililili cuginfor sit out for cavat to pierce corner and wake as he
	with LM	hen libudilililicugn for sit out for cavatopierce cornerandwake is e

Table B.3.: An example of the transcriptions of a Norwegian entry produced by the four monolingual models described in Section 4.1, with and without a Norwegian language model.

Target		maleriets integritet søkes i dets spesifikke og essensielle egenskaper
Swedish Model	no LM	måleriets integritet sökes i dess pescifike och ecensiella egenskaper
	with LM	maleriets integritet sökesidesspescifike och censiella egenskaper
Danish Model	no LM	molet eas integritat serkis i des bedste fie oracancia le aiken skar børd
	with LM	målet eies integritet serkisiedes bedstefieoracancia akenskarbørd
Norwegian Model	no LM	malleriets integritet søkes i det spesifikke og essensielle egenskaper
	with LM	maleriets integritet søkes i det spesifikke og essensielle egenskaper
English Model	no LM	moloria's integritate circus in the specific o assincera le egnscopper
	with LM	moloria's integritate circus i te specific senter leegnscopper

B.2. Multilingual Models

Table B.4.: The performance of the three Scandinavian ASR models across the regional dialects of Swedish, Danish, and Norwegian. Matching model and target language pairs are highlighed.

	SWEDISH				**DANISH**				**NORWEGIAN**			
	No Language Model		With Language Model		No Language Model		With Language Model		No Language Model		With Language Model	
	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER
Östergötland	1.68%	0.66%	2.34%	0.77%	125.33%	64.42%	102.27%	59.80%	83.03%	26.51%	73.60%	25.00%
Västra sydsverige	2.35%	0.91%	2.97%	1.02%	113.51%	55.76%	96.03%	52.65%	89.44%	30.21%	78.50%	28.45%
Östra sydsverige	2.13%	0.85%	2.81%	0.97%	118.30%	58.63%	97.70%	54.90%	85.74%	27.82%	75.77%	26.26%
Västergötland	2.07%	0.78%	2.65%	0.87%	125.54%	62.53%	101.25%	57.75%	81.04%	25.06%	72.38%	23.58%
Göteborg med omnejd	2.12%	0.88%	2.63%	0.97%	123.42%	62.13%	100.13%	57.59%	82.02%	25.90%	72.93%	24.22%
Mellansverige	2.02%	0.76%	2.45%	0.84%	121.22%	62.12%	99.66%	58.02%	82.35%	25.39%	72.78%	23.77%
Västsverige	1.77%	0.68%	2.59%	0.82%	122.32%	62.36%	100.27%	57.75%	80.29%	24.59%	71.55%	23.17%
Stockholm med omnejd	1.75%	0.68%	2.28%	0.77%	120.87%	62.85%	99.51%	58.39%	82.92%	25.76%	73.37%	24.14%
Norrland	2.16%	0.86%	2.74%	0.96%	119.23%	61.34%	98.18%	57.37%	82.11%	24.49%	72.24%	22.77%
Dalarna med omnejd	1.91%	0.74%	2.16%	0.80%	118.57%	60.30%	97.63%	56.34%	80.99%	24.21%	71.31%	22.50%
Vest- og Sydsjælland	79.03%	40.22%	73.09%	41.32%	19.23%	5.31%	13.76%	4.16%	85.16%	37.94%	76.45%	37.80%
Storkøbenhavn	79.31%	40.89%	73.68%	42.06%	19.54%	5.62%	14.26%	4.52%	84.75%	38.18%	76.09%	37.92%
Østjylland	78.95%	39.70%	73.35%	40.82%	21.24%	6.08%	15.37%	4.84%	84.16%	36.93%	76.07%	36.44%
Nordjylland	77.91%	37.81%	71.36%	38.55%	18.73%	5.08%	13.53%	3.98%	82.91%	34.79%	73.74%	33.78%
Sønderjylland	78.69%	39.41%	72.83%	40.30%	19.76%	6.36%	14.40%	5.27%	81.95%	35.16%	73.07%	34.43%
Fyn	76.40%	37.55%	70.16%	38.21%	15.83%	4.19%	11.13%	3.17%	81.32%	34.63%	72.27%	33.84%
Vestjylland	77.92%	39.08%	71.85%	40.12%	18.45%	4.97%	12.98%	3.82%	84.07%	36.94%	75.22%	36.35%
Voss og omland	61.06%	20.02%	51.61%	18.50%	99.64%	46.19%	84.77%	43.00%	16.20%	3.43%	11.71%	2.63%
Sunnmøre	61.93%	20.69%	52.32%	19.14%	108.01%	53.11%	91.09%	49.39%	19.87%	4.36%	14.58%	3.37%
Hedmark og Oppland	61.58%	20.81%	51.35%	19.06%	107.34%	56.75%	93.31%	53.81%	14.32%	2.99%	10.40%	2.29%
Oslo-området	61.45%	22.32%	52.18%	20.67%	109.45%	57.81%	94.69%	54.57%	16.53%	5.03%	12.67%	4.39%
Ytre Oslofjord	62.36%	21.71%	52.55%	20.02%	106.82%	57.02%	93.26%	54.22%	17.36%	3.72%	12.64%	2.91%
Nordland	60.83%	20.60%	51.03%	18.91%	104.38%	52.05%	90.09%	49.23%	17.61%	3.70%	12.45%	2.81%
Trøndelag	60.94%	20.41%	50.17%	18.36%	119.68%	62.83%	99.50%	58.11%	14.08%	2.96%	10.34%	2.31%
Sør-Vestlandet	63.38%	21.78%	53.40%	20.21%	94.68%	43.69%	82.32%	41.09%	20.34%	4.82%	14.85%	3.78%
Bergen og Ytre Vestland	64.14%	22.96%	54.81%	21.53%	93.39%	42.93%	81.27%	40.75%	21.37%	5.21%	15.20%	4.03%
Sørlandet	61.20%	20.53%	51.46%	18.88%	95.39%	46.13%	83.51%	43.85%	17.65%	3.69%	12.70%	2.82%
Troms	61.17%	20.16%	51.20%	18.46%	103.32%	51.70%	89.22%	48.69%	16.79%	3.32%	12.19%	2.57%

Table B.5.: Examples of the transcriptions predicted by the multilingual model in Swedish, Danish, and Norwegian, with and without a language model in the corresponding language.

Language	Setting	Transcription
Swedish	**Target**	samverkande krafter kommer att hålla förändringarna på måttlig nivå
	no LM	samverkande krafter kommer att hålla förändringarna på måttlig nivå
	with LM	samverkande krafter kommer att hålla förändringarna på måttlig nivå
Danish	**Target**	hvad synes du om den nye bank der åbnede i går dernede på hjørnet over for kirken
	no LM	hvad synes du om den nye bank der åbnede i går dernede på hjørnet over for kirken
	with LM	hvad synes du om den nye bank der åbnede i går dernede på hjørnet over for kirken
Norwegian	**Target**	maleriets integritet søkes i det spesifikke og essensielle egenskaper
	no LM	maleriets integritet søkes i det spesifikke og essensielle egenskaper
	with LM	maleriets integritet søkes i det spesifikke og essensielle egenskaper

Table B.6.: The Word Error Rates of the trial models and monolingual baselines on Swedish, Danish, and Norwegian alongside the average values throughout the training steps.

	Model \ Training steps	1000	2000	3000	4000	5000	6000	7000	8000	9000
Swedish	Monolingual Baseline	9.11%	8.23%	7.94%	7.84%	7.74%	7.23%	7.30%	6.98%	6.87%
	Retraining DA+NO	47.00%	46.85%	51.95%	51.33%	51.15%	52.77%	54.10%	53.17%	54.75%
	Retraining DA+NO+SE_full	12.32%	11.65%	13.34%	19.76%	16.94%	16.89%	13.45%	18.79%	17.24%
	Retraining DA+NO+SE_half	3.88%	4.30%	5.01%	5.94%	6.35%	6.33%	5.60%	5.60%	6.15%
	From Scratch DA+NO+SE	97.10%	4.30%	4.97%	4.79%	5.56%	6.27%	5.96%	6.34%	6.52%
Danish	Monolingual Baseline	55.17%	42.22%	39.57%	36.82%	34.78%	33.59%	32.98%	32.29%	32.06%
	Retraining DA+NO	61.98%	51.59%	46.11%	44.48%	41.90%	39.40%	37.69%	37.37%	36.54%
	Retraining DA+NO+SE_full	62.92%	56.63%	52.19%	48.81%	45.55%	44.25%	42.56%	41.75%	40.93%
	Retraining DA+NO+SE_half	61.84%	53.56%	49.75%	45.78%	43.70%	42.45%	40.12%	38.85%	39.77%
	From Scratch DA+NO+SE	99.97%	63.65%	55.78%	52.79%	48.48%	45.53%	43.97%	42.56%	41.85%
Norwegian	Monolingual Baseline	24.03%	12.36%	8.89%	7.02%	6.14%	5.37%	4.60%	4.06%	3.97%
	Retraining DA+NO	23.66%	16.33%	12.89%	11.54%	10.37%	8.91%	7.85%	7.07%	6.52%
	Retraining DA+NO+SE_full	29.64%	22.12%	16.80%	14.87%	13.04%	11.20%	10.21%	9.68%	9.10%
	Retraining DA+NO+SE_half	27.83%	18.95%	15.84%	13.02%	11.61%	10.47%	9.64%	8.75%	8.11%
	From Scratch DA+NO+SE	99.95%	28.98%	20.35%	17.77%	14.85%	12.19%	11.73%	10.37%	9.24%
Average	Monolingual Baseline	29.43%	20.94%	18.80%	17.23%	16.22%	15.40%	14.96%	14.44%	14.30%
	Retraining DA+NO	44.21%	38.26%	36.98%	35.78%	34.47%	33.70%	33.21%	32.54%	32.60%
	Retraining DA+NO+SE_full	34.96%	30.14%	27.44%	27.81%	25.18%	24.11%	22.07%	23.41%	22.42%
	Retraining DA+NO+SE_half	31.19%	25.60%	23.53%	21.58%	20.55%	19.75%	18.45%	17.73%	18.01%
	From Scratch DA+NO+SE	99.01%	32.31%	27.03%	25.12%	22.96%	21.33%	20.56%	19.76%	19.20%

Table B.7.: The Character Error Rates of the trial models and monolingual baselines on Swedish, Danish, and Norwegian alongside the average values throughout the training steps.

	Model \ Training steps	1000	2000	3000	4000	5000	6000	7000	8000	9000
Swedish	Monolingual Baseline	2.81%	2.61%	2.48%	2.41%	2.42%	2.36%	2.36%	2.27%	2.24%
	Retraining DA+NO	12.84%	13.05%	14.85%	14.66%	14.56%	15.38%	16.09%	15.73%	16.32%
	Retraining DA+NO+SE_full	3.40%	3.58%	3.93%	5.68%	5.19%	5.04%	4.09%	5.81%	5.34%
	Retraining DA+NO+SE_half	3.88%	4.30%	5.01%	5.94%	6.35%	6.33%	5.60%	5.60%	6.15%
	From Scratch DA+NO+SE	97.10%	4.30%	4.97%	4.79%	5.56%	6.27%	5.96%	6.34%	6.52%
Danish	Monolingual Baseline	18.72%	13.60%	12.69%	11.65%	11.03%	10.64%	10.47%	10.25%	10.11%
	Retraining DA+NO	22.18%	17.36%	14.94%	14.74%	13.65%	12.85%	12.09%	12.01%	11.85%
	Retraining DA+NO+SE_full	22.19%	19.57%	17.48%	16.27%	15.20%	14.62%	14.09%	13.73%	13.34%
	Retraining DA+NO+SE_half	22.25%	18.17%	16.83%	15.27%	14.29%	13.86%	12.89%	12.63%	13.04%
	From Scratch DA+NO+SE	98.18%	22.13%	19.09%	17.59%	16.27%	15.10%	14.21%	13.99%	13.53%
Norwegian	Monolingual Baseline	5.73%	2.92%	2.16%	1.77%	1.57%	1.39%	1.22%	1.11%	1.08%
	Retraining DA+NO	23.66%	16.33%	12.89%	11.54%	10.37%	8.91%	7.85%	7.07%	6.52%
	Retraining DA+NO+SE_full	6.96%	5.12%	3.92%	3.51%	3.07%	2.63%	2.48%	2.31%	2.21%
	Retraining DA+NO+SE_half	6.67%	4.42%	3.70%	3.04%	2.77%	2.50%	2.34%	2.09%	1.97%
	From Scratch DA+NO+SE	96.84%	6.81%	4.75%	4.14%	3.42%	2.85%	2.72%	2.48%	2.22%
Average	Monolingual Baseline	9.09%	6.38%	5.78%	5.28%	5.00%	4.80%	4.68%	4.54%	4.47%
	Retraining DA+NO	19.56%	15.58%	14.23%	13.64%	12.86%	12.38%	12.01%	11.60%	11.57%
	Retraining DA+NO+SE_full	10.85%	9.42%	8.44%	8.49%	7.82%	7.43%	6.89%	7.29%	6.96%
	Retraining DA+NO+SE_half	10.93%	8.96%	8.51%	8.09%	7.80%	7.57%	6.94%	6.77%	7.05%
	From Scratch DA+NO+SE	97.37%	11.08%	9.61%	8.84%	8.42%	8.07%	7.63%	7.60%	7.42%