University of Bergen

Department of Information Science and Media Studies

Master Thesis

Punctuation Prediction for Norwegian: Using Established Approaches for Under-Resourced Languages

Author:
Guro Sivertsen Prestegard

Supervisor:
Csaba Veres

June 2021

Abstract

Predominantly, the result of Automatic Speech Recognition (ASR) does not provide punctuation. In this master thesis, the focus will be punctuation prediction, with the end goal of predicting punctuation for ASR. There are at present several solutions with good results. However there are no research, to the author's knowledge, on this subject within the domain of Norwegian language. Some work has been done in the context of making multilingual models, but none which focus specifically on the Norwegian. This master thesis will therefore take a closer look at punctuation prediction for the Norwegian language, and by extension to what degree an approach designed for another language might be applicable for an under-resourced language. The project uses a bidirectional recurrent network with attention mechanism for punctuation restoration, with a dataset comprised of content from Norwegian newspapers as training data for the model. The model achieves all over high results on in-domain-data, and performs better than the reference model on almost all counts. The metrics for periods are especially good, and are consistently around 91.5 for precision, 90.7 for recall, and 91.1 for F1. The model also outperform the state of the art on overall precision, by 1.5 percent. When validating the model on out-of-domain data from transcriptions, the model performs much poorer. Some of the poor results however might be attributed to the quality of the validation transcriptions. Nevertheless, the project finds that good results can be achieved by replicating an approach initially made with different language and training data in mind, for Norwegian data. Especially if the data used for training is similar to the expected input-data for the model. This shows promise, not only for Norwegian, but also for other under-resourced languages.

Acknowledgements

First, I would like to pay my special regards to my supervisor Csaba Veres for invaluable support and encouragement, asking good questions and steering me in the right direction. A big thanks to my partner and friend Tim for being my rock throughout this thesis, and for contributing with proofreading. With his help, I have at least tried to keep the punctuation-related typing mistakes to a minimum considering the subject of the thesis. I would also like to thank my fellow students and my study hall 638, for discussions, help, and good atmosphere. Lastly, thank you to everyone who have supported me as I worked on my thesis with words of encouragement, interest, and feedback.

Acronyms

ASR Automatic Speech Recognition

BERT Bidirectional Encoder Representations from Transformers

BLSTM Bidirectional Long Short-Term Memory

BPE Byte Pair Encoding

BRNN Bidirectional Recurrent Neural Network

CNN Convolutional Neural Networks

CUI Conversational User Interface

GRU Gated Recurrent Unit

IWSLT International Conference on Spoken Language Translation

LSTM Long Short-Term Memory

MT Machine Translation

NER Named Entity Recognition

NLP Natural Language Processing

NREC Norwegian Research and Education Cloud

POS Part-of-Speech

RNN Recurrent Neural Networks

SBD Sentence Boundary Detection

VM Virtual Machine

VUI Voice User Interfaces

WER Word Error Rate

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The way we communicate continues to be constantly more digital, and textual information usually go by a computer before finding a human recipient, if it ever does. As our keyboards have shrunk by the rate of our phones, automatic speech recognition (ASR) seems a more and more viable option for transferring semantic context to our devices. A study showed that ASR was almost three times as fast for both English and Mandarin when typing on a smartphone's small touch-based keyboard [74]. The transcription technology was also less prone to errors than the human participants when typing, although the end result showed that the speech recognition software was a bit more erroneous. Furthermore, according to a report from 2018, 27 % percent of the global population uses voice search on mobile [34]. ASR services seems also to be growing in popularity amongst the younger generation, and will therefore most likely continue to rise.

ASR is found widely beyond the scope of smartphones, and is used in an abundance of applications and practices. It is used for transcribing dictation in the medical field, providing subtitles for live TV, making translations, and used in conversational interfaces, along with many other examples [24]. ASR can make humans and computers interact more seamlessly, as well as make wast amounts of data available for research and other purposes which is only available as audio [6].

Predominantly, the result of ASR processing does not provide punctuation. Punctuation can be defined as

> "the act or practice of inserting standardized marks or signs in written matter to clarify the meaning and separate structural units" [72].

In layman's terms, punctuation can tell us when a question has been asked, where a sentence should end, and where a sentence can be divided into different sub-sections. It also indicates sentiment, it tells you if something is a direct quote, and if a string of words actually is a list.

There are several ways to structure written text. One can divide sentences with new lines, often seen in free verse, one can also mark sections with a line shift, or by adding the (arguably) most important structural marker; punctuation. A non-punctuated text will prove harder to read and more difficult to understand, and will negatively impact the user experience [93]. In addition, improved structure will ameliorate ensuing processing of text and make it more exact. Examples that benefit from punctuation in text include machine translation, extracting information, named entity recognition (NER), part-of-speech (POS) tagging, and summarization [70]. In addition, keeping the punctuation data while topic modelling has been found to be highly predictive of topic carry-over [15].

In this thesis, the focus will be punctuation prediction. There are at present several solutions with good results. However there are no research, to the author's knowledge, on this subject within the domain of Norwegian language. Some work has been done in the context of making multilingual models, but none which focus specifically on the Norwegian. A surface review of research within the domain of semantic technologies indicates that there are a significantly less amount of research as well as technologies available for smaller languages in general. These are referred to as *under-resourced languages*, which is a term introduced by Krauwer[51] and Berment [11]. We can count Norwegian as one of these languages [88]. This goes for both the written styles; *bokmål* and *nynorsk* Under-resourced languages are languages which have one or more of these characteristics; Lacks a system for writing or spelling which is unique and stable, underrepresented on the internet, is in want of linguistic competence, shortfall of digital resources for language- and speech processing, is missing digital assets for language- and speech processing, for instance, bilingual digital dictionaries, text transcribed from speech, dictionaries for pronunciation, monolingual corpora, vocabulary lists, and so forth. [12]

Norway is a relatively small country, with a modest population. Per the first quarter in 2021 Norway had 5 391 369 inhabitants [10], and approximately 90% of the public where native speakers of Norwegian [79] in 2012. While Norwegian cannot be argued as the language which ticks the most boxes when it comes to be under-resourced, it does tick some. There is shortage of digital tools and corpora for language processing, and can be said to be under-reprecented on the Internet. This goes especially for the Norwegian minority written style *Nynorsk*.

Compared to languages like English, there is little work done on automatic processing of Norwegian. This is problematic when considering the possibilities for research and processing of information in Norwegian language, and might be a factor in the future development and persistence of the language. One solution to solve this issue, might be to adopt approaches designed for other language domains into under-resourced language domains.

The punctuation marks which will be attempted predicted in this thesis are shown in the table below.

Punctuation symbols	
Period	.
Comma	,
Exclamation mark	!
Question mark	?
Colon	:
Semicolon	;
Dash	-

Table 1.1: Punctuation symbols relevant for prediction in this thesis

1.1 Motivation

The motivation for this project is to create a model for punctuation prediction for the Norwegian language. The way this will be undertaken in this project is to evaluate how well a state-of-the-art system generalizes to a smaller and under-resourced language like the Norwegian language. The approach chosen for this purpose, is described in *Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration* by authors Alumäe and Tilk [83]. This approach has been chosen as it has been proven to produce good results for both Estonian and English as well as having publicly available source code.

To investigate whether the approach can generalize, the existing framework have been trained with Norwegian data, and and have resulted in a working model. These efforts will give people wishing to work with the output from ASR in Norwegian a working model which can be utilized for predicting punctuation. This can be useful for anyone working with transcribed text in Norwegian, as well as in other languages. The work as a whole can serve as a starting point for anyone else attempting to build their own model for punctuation prediction, for an under-resourced language. This thesis will show if the kind of approach discussed in this paper shows promise, or if other paths should be explored.

1.2 Research Hypothesis and Problem Statement

This thesis attempts to improve upon two issues; the improvement of punctuation restoration for Norwegian, and thereby, the lack of technologies for under-resourced languages.

The research on punctuation prediction is not perfect within the domain of any language. However, for Norwegian it is as good as non-existent. Punctuation prediction becomes more and more important with our increasing use of automatic speech recognition technologies, and the lack thereof results in poorer performance of text processing for both machines and humans. Missing punctuation can influence the readers' understanding of the text, and make machine processing of automatically transcribed text less accurate than it would have been with proper punctuation.

In general, Norwegian, together with other under-resourced languages, lack tools for language processing which are available for languages with a larger group of speakers. Smaller languages will have fewer stakeholders, and thus less motivation for developing new tools and techniques. With fever hands on deck, it stands to reason that one can not expect the same amount of research and development as for a larger language. The objective for this thesis is thus;

- Firstly, attempt to improve punctuation prediction for Norwegian and by that make a humble start to this field of research
- Secondly, shed light on whether under-resourced languages can benefit from adopting approaches used by languages with more resources available.

The fully formulated research hypothesis is thus;

In the scope of Norwegian language, a model for punctuation prediction can be created using an approach designed for another language, trained on Norwegian data sets. This study might have a positive impact on the research community and contribute to the state of Norwegian as a modern language which partakes in the digital world, as well as give insight for other under-resourced languages on how to best use existing resources.

1.3 Research Questions

- *RQ1.* Can an approach to punctuation restoration be used with other languages than it was designed for, and can this help under-resourced languages?
- *RQ2.* Can only textual training data produce desirable results for Norwegian?
- *RQ3.* How will the error calculation compare between the model trained by Tilk and Alumäe [83] trained on English data and the model trained on Norwegian data?

1.4 Research Contribution

This work aims to contribute to the field of research, by exploring the domain of punctuation prediction as well as the availability of tools for under-resourced languages. The thesis has one main contribution; A model which predicts punctuation in unsegmented text for Norwegian. Additionally, it presents an exploration of how tools for processing text can generalize between different languages, and thus make technologies available for under-resourced languages. Lastly, this thesis represents an insight into the approach described by Alumäe and Tilk.

1.5 Relevance of This Study

The author aims to produce a thesis and artefact which will be useful, beyond academia. The artefact, which will be a model for predicting punctuation for unsegmented text, can be used by organizations who wishes to make tools and products which makes use of ASR. It might also come in useful for processing data, where punctuation is not available, to ensure topic-carry over when topic modelling.

1.6 Thesis Outline

In this section follows an outline of the chapters in the thesis and their content. The present section, the *Introduction* is excluded in this outline. The thesis is organized into 5 chapters, and contains an appendix with relevant data for the thesis.

Chapter 2: Literature Overview

The current chapter aims to give the reader an introduction to earlier studies, works and experiments done in this field of study, which are relevant to the case at hand. It presents an overview over recent research in the field of punctuation prediction, as well as a brief history of the development of said field.

Chapter 3: Methods

This Chapter will give give a precise and thorough description of the way this study was conducted. Thus the approach described can be replicated, developed further and peer reviewed by researchers wishing to do so. This chapter will encompass methods, both for data collection and development, as well as result metrics and analysis tools. In addition, the overall method of the framework which will be [83] used will be described.

Chapter 4: Results

This Chapter gives an overview of the results from the different models. This includes the results from the baseline model created with English data, as well as the two Norwegian models. The results are represented as error metrics.

Chapter 5: Conclusions

The concluding chapter seeks to give a comprehensive summary of the key elements of this thesis. This includes accomplishments and achievements, as well as shortcomings and future work.

Chapter 2

Literature Overview

The current chapter aims to give the reader an introduction to earlier studies, works and experiments done in this field of study, which are relevant to the case at hand.

2.1 Automatic Speech Recognition (ASR)

This first section outlines the issues and grounds the object of the thesis in the problem area at hand, and describes the current field. A small review of available technologies have also been performed.

The objective of Automatic Speech Recognition (ASR) is to "address the problem of building a system that maps an acoustic signal into a string of words" [45]. Which in layman's terms can be said to let a machine transcribe speech.

As stated by Juang and Rabiner [43], speech is "the primary means of communication between people". Then, it is within reason to state that one of the more efficient and user-friendly methods of transferring information from humans to machines, is through speech. Designing a machine which imitates human behaviour has summoned the interest of scientists and engineers for hundreds of years. However speech analysis and synthesis can be said to start in the 1930s, on Homer Dudley of Bell Laboratories initiative [43]. It was since popularized in the 1960's by science fiction-movies by directors such as Stanley Kubrick and

George Lucas. Here we find conversational interfaces operating on screen. Take for instance HAL in 2001: A Space Odyssey, or C3PO in the original Star Wars trilogy. One can also argue R2D2 as a conversational interface, as it understands natural language.

One of the next major mainstream events in the field , was the launch of Apple's "Knowledge Navigator", which both had a Speech User Interface, as well as a Multimodal User Interface.

Ever since, the domain has been under constant development. At present day there is many uses for these kinds of technologies, and speech processing or synthesizing can be found in many forms and variations. From compiling transcripts of interviews and other oral sources, to conversational user interfaces and aids for the visually impaired. Some of the more known applications might be Apple's Siri[4], Amazon's Alexa[3] or Google Assistant[33]. These are all examples of Voice User Interfaces (VUI), and Conversational User Interfaces (CUI). IBM's Watson Assistant [89] is also a good example of this technology, and were early on a strong player in the field of CUI's. Many conversational interfaces has evolved to more than just question-answering, and is now a broader AI-service which also performs speech-to-text as well as speech synthesis based on text.

2.1.1 Current State of Automatic Speech Recognition

During the last years there have been significant advancements in the domain of ASR, where the recognition accuracy has improved significantly [55]. Systems has for instance been know to outdo the Word Error Rate (WER) of transcriptions made by humans on a set of colloquial telephone conversations [92]. ASR has also been reported as being almost three times as fast for both English and Mandarin when typing on a smartphone's small touch-based keyboard [74]. The transcription technology in the latter case was also less prone to errors than the human participants when typing, although the end result showed that the speech recognition software was a bit more erroneous.

While ASR has come a long way, there are still problem areas within the field. These include issues such as when the microphone is far from the user [75], adversarial examples

[2], child speakers [90], multiple speakers, noise [28], dialects and so forth. The output form ASR systems is also often missing textual structure which is common to find in written text. Such structures can be capitalization, punctuation or formatting of numeric data. [70].

2.1.2 Review of ASR technologies

There are currently many speech-to-text processors, or automatic speech recognition services on the market. The author has previously in this paper claimed that the status of the punctuation prediction is poor in many technologies. J. Kim, C. Liu, R. Calvo et al. conducted a review in 2019 over how the different ASR services compared to each other. This study does not however take into account the correctness of the punctuation, but focuses on the words only when calculating the score for each service. [46].

This paper has performed a small review of transcription technologies which support Norwegian, to verify the state of missing punctuation, but also the overall performance of the technologies.

The technologies tested are Google Voice Typing [86], Apple Dictation [27], SpeechNotes [80] and SpeechTexter [81]. These technologies are chosen because they are available for most people, as well as having support for Norwegian. They do not, however, support punctuation [85] [52] [27], to the authors knowledge. The services do however for the most part offer the user to dictate the punctuation, by saying codes for punctuation.

Services like, e. g. IBM's Watson and Amazon AWS speech transcription software have not made the cut of this small review as is not freely available in Norwegian.

The text used for this small review is:

Er det egentlig så farlig med den tegnsettingen? Jo, tegnsetting er like viktig å kunne som å stave ordene riktig, selv om noen velger å se på det som en ikke-sak. Noen ganger er tegnsettingen avgjørende for setningens mening. Det klassiske eksempelet som alle lærer på barneskolen, er setningen: «Heng ham ikke vent til jeg kommer!». Plasseringen av kommaet etter henholdsvis «ham» og «ikke» avgjør den dødsdømtes skjebne. Det kan altså være livsviktig med riktig tegnsetting; plutselig har man voldt noens død.

The test sentence is partly inspired by an example found on a blog for spelling. [82]. The sentence translated to English is as follows:[1]

> Is punctuation really that important? Yes, punctuation is as important as being able to spell the words correctly, even though some look at it as being of no matter. Some times punctuation is crucial for the semantic content of the sentence. The classic example which all pupils learn in primary school, is the sentence "Hang him not wait until I arrive!". The placement of the comma after "him" and "not" decides the faith of the condemned. Therefore, it might be a matter of life-and-death to have the proper punctuation; suddenly you've caused someone's death.

This sentence, contain all of the punctuation symbols this thesis is aiming to reconstruct; the period, the comma, the exclamation mark, the question mark, the colon, the semi-colon and the dash. It also contains some semantic ambiguity, which can be read differently based on the punctuation in the sentence. The Norwegian word "ikke" can in this context mean both "not" and "don't". A similar example that works better for English might be "It's time to eat, grandma!", versus " It's time to eat grandma!".

Results

The results are quite even amongst the technologies. Google Voice Typing[86] and SpeechTexter [81] have almost identical results, where SpeechTexter is a fraction better when it

[1] Translated by the author

comes to correct words. Apple Dictation [27] performs best on correct words, but worst on everything else. Apple Dictation is the only service that does not correctly identify the dash in the Norwegian compound word "ikke-sak" ("of no matter"), even though the words "ikke" and "sak" where correct. Apple might have come to the decision of excluding all punctuation, even though punctuation in compound words like the one mentioned above actually is described in the dictionary [42].

Service	Punctuation	Capital letters	Words correct
Google Voice Typing	partial	3 out of 7	68 out of 81
Apple Dictation	none	1 out of 7	76 out of 81
Speechnotes	partial	2 out of 7	72 out of 81
SpeechTexter	partial	3 out of 7	72 out of 81

Table 2.1: Overview Over Screened Transcription Services

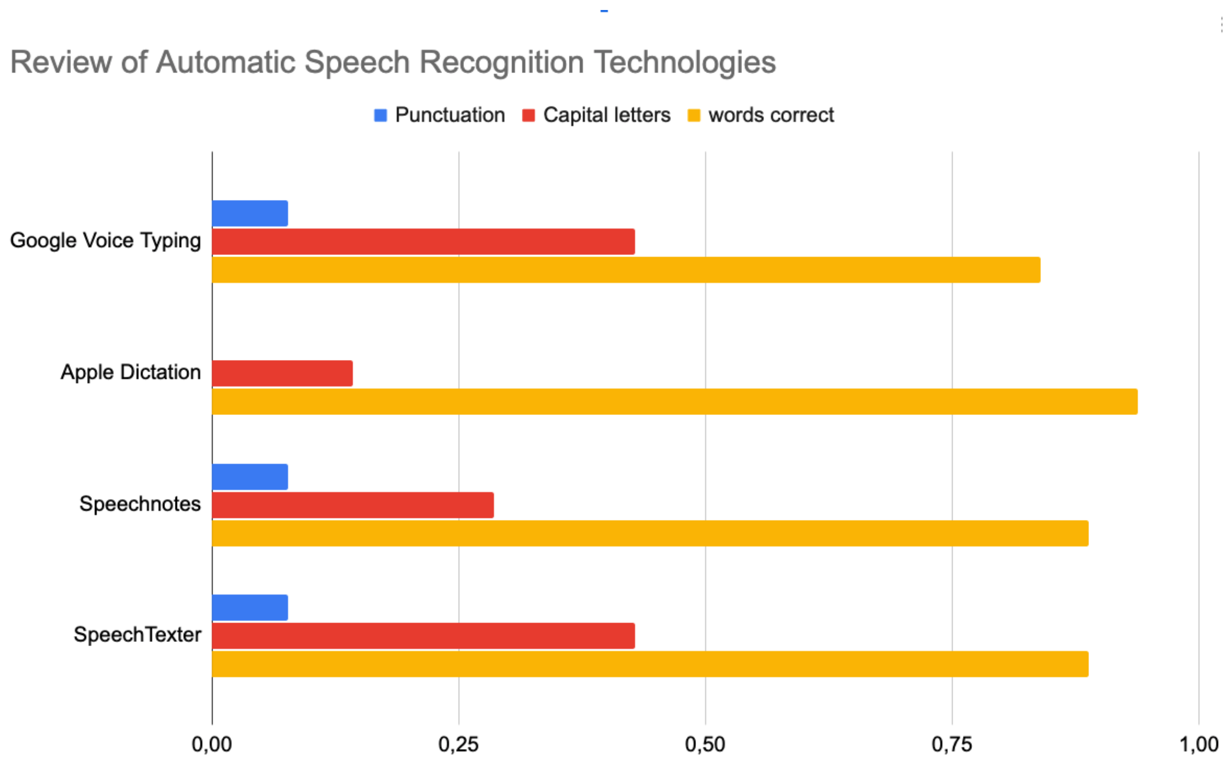


Figure 2.1: Review of Automatic Speech Recognition Technologies, results in percentage factors

2.1.3 Punctuation in ASR

The focus of this thesis, will be on the textual structure element mentioned above, known as punctuation, and it's prediction in unsegmented text. The greater part of ASR systems returns word sequences without punctuation[84]. Having the punctuation restored will considerably improve the ease of reading of transcribed text as well as increase the effectiveness of subsequent processing, like machine translation, summarizing, question answering, sentiment analysis, syntactic parsing and information extraction. [83]. One reason for taking punctuation into account, is that it has been found to be markers of topic-change and improve processing of text. Büschken and Allenby found that punctuation was a good marker for when topics changed, and found it highly diagnostic of this issue. They argue this as a point for not discarding punctuation data when processing text [16].

Tilk and Alumäe [84] states that while unsegmented text might be adequate for some uses, like retrieval and indexing, dictation, and so forth, most other uses that involves speech transcripts would improve when punctuation is included. For one, it would, according to Tilk and Alumäe, make the text easier to read. Another point they make is that certain language processing tools expect their input to be already punctuated text, and would thereby make the use of these tools and techniques possible. Examples include sentiment analyzers, syntactic parsers and translation systems. The simple process of sentence tokenization is also easier if the text contains full stops [41].

P. Hlubík, M. Španěl, M. Boháč et al. claims something similar. They state that speech recognition technologies output a "continuous stream of words, that has to be post-processed in various ways, out of which punctuation insertion is an essential step" [38]. They also report that text that is punctuated is easier to understand by a reader, and it also has many uses; it can be used as subtitles, and it is also a necessity for continued Natural Language Processing, e. g. machine translation [38].

2.2 Punctuation Restoration

This section will look at the most used approaches and methods used within punctuation prediction.

To get a overview over the field, a systematic literature review into the field of punctuation prediction has been undertaken. The field emerges as rather small, and does not yield an abundance of results (see table 2.2). It can also be noted that many of the results for each of the search phrases were overlapping, so the total where not as high as the calculated result. In addition, a portion of the search results are patents, which this study will not be including. If one is to search for "machine learning" on Google Scholar, the amount of hits will at present count about three and a half million. By comparison, the field of punctuation seems quite limited, where the result for the 7 different queries done clocks in just shy of 2000 hits. Although the field is not as large as other fields, there is quite a bit of work undertaken on the topic of punctuation restoration. The rest of this section will largely concern the different approaches found in the literature review, and a short introduction to the different methods and terms, as well as papers which uses the approaches mentioned.

Query	Results
"Insert punctuation"	522
"Punctuation prediction"	370
"Punctuation annotation"	282
"Punctuation restoration"	267
"Punctuation insertion"	196
"Improved punctuation"	150
"Restoring punctuation"	133
Total	1920

Table 2.2: Search results for *Google Scholar* for punctuation restoration related search terms (Search performed May 2021)

There are several different approaches for punctuation prediction. Some studies use textual features only, some use only prosodic features, and others use a combination of both. The studies which use a combination of both, usually either have an approach which utilize one model, or an approach which combines two or more models [83]. In the next sections an overview over some approaches and papers which make use of them will be presented.

2.2.1 Prosodic Features

A prosodic feature is "a speech feature such as stress, tone, or word juncture that accompanies or is added over consonants and vowels; these features are not limited to single sounds but often extend over syllables, words, or phrases"[91]. These features can say something about the segmentation of a sentence, e. g. how some languages uses a higher pitch in the end of a sentence when asking a question, or a pause which a speaker is making use of to indicate a comma or a full stop. The pause duration is the prosodic feature which is most commonly used.

Authors O.Tilk and Alumäe[84][83] uses features from the discourse to enhance the results. A similar approach is taken by P. Żelasko, P. Szymański, J. Mizgajski et al. They make use of "conversation side indicators and word time information" [94]. J. Huang and G. Zweig [40] and F. Batista, D. Caseiro, N. Mamede et al. [9] also makes use of prosodic features, and included pause durations into their model. F. Batista, H. Moniz, I. Trancoso et al. [8] makes use of energy as well as the pitch of the voice from the speech. T. Levy, V. Silber-Varod and A. Moyal presented a paper where the authors attempted to only use prosodic features [53].

2.2.2 Conditional Random Fields (CRF)

CRFs are a framework based on a conditional approach. They define a "conditional probability $p(Y|x)$ over label sequences given a particular observation sequence x". This framework is used for calculating probability for providing labels and segmentation for sequential data [13]. CRFs have been shown to work well as a method for predicting punctuation. W. Lu and H. Ng [59] is an example of an approach that uses CRFs with transcribed text, without prosodic features. They are are using the IWSLT-data set as their evaluation-data. Approaches utilizing neural networks have more recently been proven to outdo the CRF based models.

2.2.3 Machine Translation

Machine translation uses machine learning to translate natural language. The model is trained on samples of translations prepared by humans, and is thereafter trained in how to translate. [58]

An approach described by E. Cho, J. Niehues, A. Waibel et al. [19] makes use of models for translation which are based on phrases. This system looks at the punctuation restoration as a machine translation task.

2.2.4 N-gram

A disadvantage in word-processing is according to Müller and Guido, that the order of words is not always taken into account. That means that two sentences containing the same words would be interpreted as being the same, no matter the positioning of the words. This leaves the sentences; "Summer is warm, winter is not", and "Winter is warm, summer is not" as being interpreted in the same way even thought they have different semantic content. Context needs to be included to steer clear of the above mentioned example. One can do this by considering several tokens, instead of one. A pair of tokens is called a bigram, while three tokens, or a triplet, is called trigrams. Sequences of tokens are called n-grams. The longer sequences that is used, the more features there will be. However the risk of overfitting might rise in accordance with how specialized and feature-rich the data gets [62].

One example is "Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts" [8] which is using machine learning. Another example is "Restoring punctuation and capitalization in transcribed speech" [35].

2.2.5 Maximum Entropy Model

Maximum entropy models present a system for joining varied parts of contextual proof for calculating the likelihood of a specific linguistic class appearing with a specific linguistic context [73].

J. Huang and G. Zweig uses a maximum entropy model for punctuation prediction for spontaneous text. They state that this approach "provides a easy and natural framework to incorporate both textual and prosodic information"[40], and they treat the issue as a tagging problem.

2.2.6 Neural Networks

Neural networks is the most popular approach at present. S. Kim presents a Recurrent Neural Network(RNN) approach, which uses multi-head attention mechanism for relevant contexts at each time step [47]. T. Levy, V. Silber-Varod, and A. Moyal uses a feed-forward neural network with "sigmoid hidden and output neurons" [53]. X. Che, C. Wang, H. Yang et al. propose a CNN with fully connected layers [17].

Long Short-Term Memory (LSTM)

LSTM, or long short-term memory, is an algorithm that was developed in 1997. It was developed by Hochreiter and Schmidhuber[39], and was the result of their research into the vanishing gradient-problem. This often represents a layer in a deep learning model, and is a variant for a simple Recurrent Neural Network (RNN) layer with the addition that it can carry information across steps in time. LSTM stores data so it can be used at a later time, which assures that older data is not lost int the process. [21]

Alumäe and Tilk[84], presents a recurrent network in two stages; the first stage trains on a large textual dataset, while the second trains on textual features with pause duration which makes the model more adaptive to a speech domain [84] Tilk and Alumäe states that their approach reduces errors in punctuation with 16.9%. This approach focus on commas and periods. Exclamation marks, question marks, colons and semicolons are mapped to periods, and all others are discarded. The model they use is a LSTM RNN "with forget gates and peephole connections in LSTM layers"[84].

Bidirectional Recurrent Neural Network (BRNN)

Bidirectional recurrent neural networks seems one of the most popular approaches to the issue of restoring punctuation recently (per 2020). This type of network was introduced in 1997 by M. Schuster and K. K. Paliwal [76]. Here they propose to take a regular recurrent neural network, and extend it into a bi-directional recurrent neural network. "The BRNN can be trained without the limitation of using input information just up to a preset future frame. This is accomplished by training it simultaneously in positive and negative time direction." [76]. They write that the network for each time direction focus on minimizing the objective function, and that there for this reason is no issue about merging the two different outputs. To find the optimal delay in relation to minimizing the objective function is also unproblematic, as all information, past and future, are surrounding the evaluated time point and is available. V. Pahuja, A. Laha, S. Mirkin et. al. describes an approach using bidirectional recurrent neural networks, as well as using multiple sequence labelling. C. Juin, R. Wei, L. D'Haro uses a bi-directional recurrent neural network, with Part-of-Speech tags attention mechanism [44]. S. Kim [47] makes use of a bidirectional recurrent neural network, amongst other techniques.

Transformers

This is one of the newer additions to the field. The current state of the art is held by a paper using a bidirectional transformer architecture[87]. The previous state of the art also used this approach [26]. M. Courtland, A. Faulkner and G. McElvain also makes use of bidirectional transformers [24]. A fourth study making use for Transformers is B. Nguyen, V. Nguyen, et al. They make use of chunk merging. [64].

2.3 State of the Art

Authors J. Devlin, M. Chang, K. Lee et al. have introduced a model called Bidirectional Encoder Representations from Transformers (BERT). This model pre-trains "deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers."[26]. This is a bit step forward for unsupervised pre-training. Further other models have been made available; RoBERTa [57] which is based on, and ELECTRA [22].

M. Courtland, A. Faulkner and G. McElvain [24] describes a novel approach to punctuation restoration in their paper from 2020. They use bidirectional transformers, and make use of a model trained on the above mentioned BERT for making use of unsupervised pre-training, and the architecture of transformer networks which are parallel. They achieve a overall F1 score of 83.9 for their best model, as well as overall Precision at 84.0. The above mentioned approach, gets just 1.5 points higher F1 score than the previous best approach by V. Pahuja, A. Laha, S. Mirkin et al. [70]. Their model has an overall F1 score of 82.4 and a Slot Error Rate (SER) of 33,0.

2.4 Norwegian Language

This section goes into the subject of punctuation prediction in Norwegian, and the state of the current research field.

2.4.1 Related works for under-resourced Languages

Q. Chen, W. Wan,g M. Chen et al. [18] propose a "discriminative Self-Training approach with weighted loss and discriminative label smoothing to exploit unlabeled speech transcripts". They do this as to lower the threshold for punctuation prediction, as compiling the labelled data sets which is often used is laborious and expensive.

2.4.2 Related Works for Norwegian Language

The research into the field of language processing for Norwegian language is rather limited. Before the summer of 2020, there was hardly any search results concerning Norwegian and punctuation prediction. After the EU launched the ELITR project, which will be described in more detail below, the field has been enriched with more research. However, there are still no studies concerning punctuation prediction for Norwegian, which is not a multilingual approach.

Query	Results	Relevant hits
"insert punctuation" norwegian	25	2
"punctuation prediction" norwegian	12	3
"punctuation annotation" norwegian	11	2
"punctuation restoration" norwegian	7	3
"punctuation insertion" norwegian	9	2
"improved punctuation" norwegian	8	0
"restoring punctuation" norwegian	7	1
Total	79	13

Table 2.3: Search results for *Google Scholar* for punctuation restoration related search terms for Norwegian (Many of the listed relevant articles per search are the same articles, about 6 are unique

2.4.3 The ELITR Project

ELITR stand for "European Live Translator", and is funded by the European Union as a part of the Horizon 2020 program [30]. The overall purpose of the project is to tear down the barriers between EU- and European countries, with a focus on language. The project aims to create live automatic subtitling systems, translation systems for speech, systems for summarizing, as well as improve the state of the art for machine translation. They have EU-languages specified as target languages, as well as the official languages for the EUROSAI member countries, which count 19 languages apart from the EU-languages. This project aligns with the values of this thesis; to make Natural Language Processing (NLP) tools available for smaller and/or under-resourced languages. There are no submissions into this field with Norwegian language as the main focus, to the authors knowledge.

The ELITR project is relevant in the context of this thesis, by two counts: The ELITR project has it's main focus on other languages than English, and to have the best starting point for machine translation, as many of the entries into the project are utilizing, punctuation is necessary. And as earlier stated, ASR usually outputs an unsegmented stream of words.

ELITR Non-Native Speech Translation at IWSLT 2020 [60] is a study which has some of the same objectives as this paper; It is using available methods to create and end-to-end ASR system for other languages than English. This paper is especially interesting, as it uses the punctuation framework developed by Tilk and Alumäe [83]. However, this is where D. Macháček, J. Kratochvíl, S. Sagar et al. and this thesis part ways, as the study trained their model on English data, from CzEng [25]. The model is also used for output from ASR for speech other than the language it was trained for; which is English.

A second study which makes use of the approach developed by Tilk and Alumäe is ELITR Multilingual Live Subtitling: Demo and Strategy by O. Bojar, D. Macháček, S. Sagar [14]. Another study, which also has mentioned Norwegian (as a language of a EUROSAI-country), is Removing European Language Barriers with Innovative Machine Translation Technology [31].

2.4.4 Multilingual Approaches

Authors X. Li and E. Lin has created multilingual approach which specifically mentions Norwegian [54]. They claim to have support for 43 languages, distributed across 69 countries. These languages are part of many different language families. The technologies they use are Long Short-Term Memory (LSTM) and Byte Pair Encoding (BPE). They claim to achieve good results, by benefiting from shared information across different languages [54].This paper utilizes a data set from the news-domain. It also aims to provide tools to under-resourced languages by making a multilingual approach. The model, on average, with test data in this domain achieves an F1 score of 80.2%. This paper does not present results which are directly comparable with results from a Norwegian model, as this is multilingual and has not been specifically tested on Norwegian data.

There is also done some research on the subject of Sentence Boundary Detection (SBD), with a multilingual approach. This paper is older, however it is mentioned for good measure. T. Kiss and J. Strunk [48] have trained on a newspaper corpora, and claim a mean accuracy of 98.74% for eleven tested languages.

2.5 Data Sets

The subsequent section discusses data sets in general as well as the data chosen for this study.

2.5.1 Common Datasets

Europarl

Europarl is a corpus expanded by researchers led by P. Koehn [50]. The dataset contain transcribed speech from proceedings from the European Parliament. This data exists for 11 languages, which are the 11 official languages of the European Union. These languages are Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish. The corpus has seen much use within the domain of NLP.

IWSLT

The IWSLT dataset is a MT dataset that contains ASR transcriptions of TED-talks, as well as translations. This seems to be the most popular dataset for punctuation prediction. These talks covers many different topics, and are held by many different speakers. This is not considered one of the biggest datasets, and it only contains about 130 000 sentences, and has vocabulary size of 17188. One of the advantages of smaller datasets is that the models and results can be calculated faster.

Pre-training for BERT

Wiki Extractor [5] and BookCorpus [49] can be used for scraping sources. They are often used for pre-training datasets for BERT. Datasets produced with these tools include Wikipedia raw text and data from available book [26].

Spontal-N

A corpus which might contain data suitable for this study, is the *Spontal-N* corpus. This is a corpus of spoken and unscripted Norwegian, which is manually transcribed. However, this corpus has focused on Norwegian speakers who have lived a large portion of their lives in Sweden, and might in that way not be applicable. On the other hand, it might result in a model which works both for Swedish and Norwegian.[78]

2.5.2 Nasjonalbiblioteket

Nasjonalbiblioteket (The National Library of Norway) has recently done an incredible job when it comes to making data sets for language processing available. And new data sets are in constant development.

Norsk Aviskorpus

The choice for dataset for this thesis, fell on *Norsk Aviskorpus* [65]. This dataset is divided into two parts, *Norsk Aviskorpus Nynorskdelen*[66] and *Norsk Aviskorpus Bokmål*[65] . While Nynorsk is considerably more under-resourced than Norwegian Bokmål, the size of available data made the choice for the author in this case. The dataset for Bokmål is more than 120 times as big as the one for Nynorsk, and has a size of 1 961 933 309 tokens, while the corpus for Nynorsk counts 16 070 002 tokens.

Title	Value
Newspapers in the corpus	Adresseavisen
	Aftenposten
	Bergens Tidende
	Dagsavisen
	Dagbladet
	Dagens Næringsliv
	Fædrelandsvennen
	Nordlys
	Odin
	Stavanger Aftenblad
	Verdens Gang
Tokens Nynorsk	16 070 002
Tokens Bokmål	1 961 933 309
Genre	Newspaper and Magazines
Time Coverage	October 1998 - September 2015

Table 2.4: Norsk Aviskorpus [65]

Stortingskorpuset

While this thesis has been in progress, new data sets have been made available. One resource, that the author considered compiling as a part of this thesis but could not be done for the reason of time and resources available, is *Stortingskorpuset* (The parliament corpus). It was made available at the end of May 2021. This is a corpus compiled of orthographically transcribed debates in the Norwegian parliament as well as the audio recordings. The transcripts are made with ASR, but they have been reviewed by trained linguists that have made the necessary edits. The transcripts also contain data such as speaker, and have preambles with meeting data [63].

Chapter 3

Methods

The purpose of this chapter is to give a precise and thorough description of the way this study was conducted. Thus the approach described can be replicated, developed further and peer reviewed by researchers wishing to do so. Therefore, this chapter will encompass methods, both for data collection and development, as well as result metrics and analysis tools. In addition, the overall method of the research project will be described.

3.1 Development Methods

This paper will make use of several different methods and approaches. In this section an overview will be presented.

3.1.1 Design Science Research

Hevner et al. suggests seven guidelines for design science research. [37] They say that design science is a domain of solving problems. "knowledge and understanding of a design problem and its solution are acquired in the building and application of an artifact." [37]. What they mean by that is that one needs to understand more than one problem space to be able to develop a viable artefact, and that it in itself should be novel, and useful. The process of developing an artifact using design science is precisely that, a process. Through iterations, the artifact is constantly being improved. The next sections entails a walk-through of the design guidelines[29], and how this paper fills these criteria. All the criteria used below are from Hevner et al[37]

Title	Principle
Design as artifact	Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.
Problem Relevance	The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
Design Evaluation	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
Research Contributions	Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
Research Rigor	Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
Design as a Search Process	The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
Communication of the research	Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences.

Table 3.1: Table of Design Science guidelines [37]

Design as Artifact

> *Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.*

The project in this paper has resulted in two models which has been created through an iterative process; one for each of the Norwegian written styles. The models can be useful for people who wishes to continue researching the field, or wishes to use the model for punctuating text.

Problem Relevance

> *The objective of design-science research is to develop technology-based solutions to important and relevant business problems.*

This thesis attempts to solve the issue of punctuation in the field of automatically transcribed text in the domain of Norwegian language. This has not been done previously with focus on Norwegian language. The development of this model can make processing of information in the form of audio more accurate and in less need of manual correction. Punctuation prediction can be added to ASR services to improve the overall output. This can in turn be used for many different organizations and businesses who needs ASR for their business cases; such as live captioning, companies performing data analysis, dictation services, voice search, and so forth.

Design Evaluation

> *The utility, quality, and efficiency of a design artifact must be rigorously demonstrated via well-executed evaluation methods.*

The model has been tested and trained with Norwegian data, which was largely chosen based on availability. Other factors where, number of samples and generalization. The performance of the model was measured in metrics such as precision, recall, and F1, SER and ERR. This makes the model easy to compare to other instances, as these are the most common metrics used for punctuation prediction. These metrics give a good indicator towards how well the proposed punctuation symbols are placed in the text.

Research Contribution

> *Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.*

The research contribution of this thesis first and foremost lies in the fact that the field has up until now had no research for Norwegian. This thesis rectifies this, and is the first attempt at creating punctuation prediction for Norwegian.

Research Rigor

> *Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.*

In the process of developing the model, a scientific method has been applied. The progress has been carefully documented. This chapter describes exactly that, and will go into details in the following sections.

Design as a Search Process

> *The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.*

The means required for this project have been within reach. Some of the means, like data, have taken more time to compile than others. The data available is not the ideal data for this task, and it would be preferable if there was transcribed data available.

Communication of The Research

> *Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences.*

This thesis is mostly directed towards the scientific community. It is however written in such a way, that it hopefully is not deterrent for audiences beyond the scope of academia. The author is of the opinion that the interests of management-oriented audiences who are dealing in this problem area might have their interest peaked by this project.

3.1.2 System Development Techniques

As this project is of a certain size and scope, a method is needed for structuring the work with the project. This project will make use of *Agile methods* to structure the work on this thesis.

Agile Development

Shore and Warden writes in *The Art of Agile Development* that Agile development is often described as a method.[77] However, to *be agile* is more of a philosophy than a fixed recipe to follow according to Shore and Warden. Agile methods, on the other hand, are practices which follows the agile way of thinking. Examples of these are Scrum, Kanban and XP, amongst others. These agile methods, consists of parts which are named *practices.* In the core of the agile way of thinking, there are four values.

- Individuals and interactions over processes and tools
- Working software over comprehensive documentation
- Customer collaboration over contract negotiation
- Responding to change over following a plan

[77] In addition to these, there are twelve principles which are based on these values, which will not be expanded on further in this paper.

According to Shore and Warden, Agile methods might make productivity higher, without necessarily working faster; Just differently [77]. Agile methods are inherently iterative. They make space for iterations and re-iterations, and focus on getting things to marked in shorter periods instead of keeping the focus on the end goal. This way issues and unreasonable deadlines are uncovered earlier, and it will therefore be easier to set the project back on course should it start to derail.

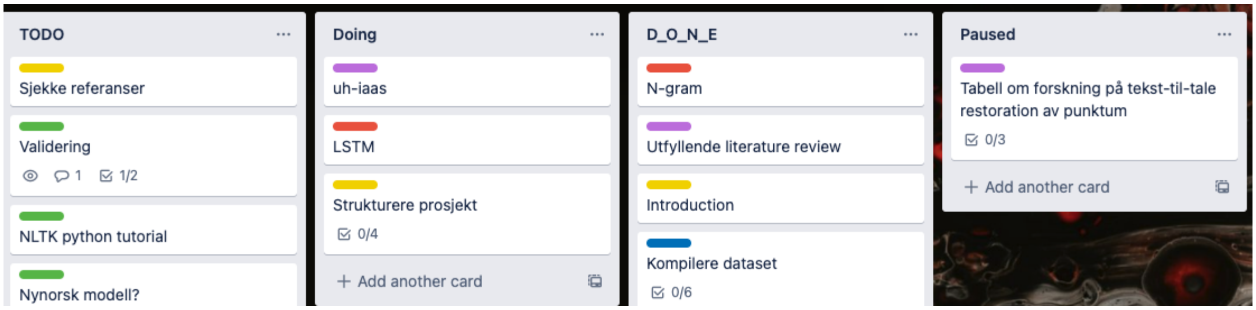
As this project moved along, and the day-to-day tasks became apparent, the project where divided up into smaller parts, so as to keep control over the whole as well as making it possible to deliver smaller pieces continuously. In that way small parts of the project will always be delivered and evaluated, instead of keeping deliverance for large completed parts of the project that then will be harder to change should they prove to not hold up in the evaluation.

The project have not adhered strictly to any Agile methodology. It has rather picked those methods which fit the project, unrelated to what school of agile development it belongs to. As the philosophy it teaches, the employment of the agile ways of development need also be agile, and possible to switch out if proven to not fit the project after all. Two agile practices ended up been selected, and stayed the course of the project; Kanban-boards and Scrum sprints.

Kanban Board

A Kanban board is a graphic representation of the work in progress. It has columns going vertically, and cards representing the tasks which are to be done. It can also have horizontal columns, if there is the need for representing multiple projects. There can only be a limited amount of cards in under each column, to limit how much work are in progress at the same time. [23] This is not the only configuration for a Kanban board, and they can be represented in many ways. In this paper the columns will represent in what state the work is - if it has not started, is in progress, is done, verified, in re-iteration, and so forth. This project has also operated with a 5 card maximum for the doing-column. This is more cards than is usually allowed per person on Kanban boards. This was chosen for the reason of the many different types of tasks that were needed from this project which overlapped. Training of the model would take several days, the writing of the thesis often coincided with need for research and knowledge building, at the same time as as collecting and processing data were needed.

The Kanban board has been chosen as it a way of having a visual overview over the process and progress. This might help keep the project on track as it highlights what work is left to do and if any of the tasks are taking more time than foreseen so the project schedule can be restructured accordingly.



TODO

Sjekke referanser

Validering

1 ☑ 1/2

NLTK python tutorial

Nynorsk modell?

Doing

uh-iaas

LSTM

Strukturere prosjekt

☑ 0/4

+ Add another card

D_O_N_E

N-gram

Utfyllende literature review

Introduction

Kompilere dataset

☑ 0/6

Paused

Tabell om forskning på tekst-til-tale restoration av punktum

☑ 0/3

+ Add another card

Figure 3.1: Kanban board used for the project

Scrum Sprints

The name Scrum comes from an event during a rugby match, where players congregate around the ball, and the teams co-operates to send the ball forward. According to Popli and Chauhan, Scrum facilitates incremental and iterative development of products, effectiveness, punctuality, and growth in revenue.[71]

Scrum focuses on continuous delivery, at set intervals which often has the duration of a month called sprints. The goal of a sprint is usually to deploy and develop a piece of software. A sprint has phases. The sprint planning, the sprint, and the sprint review. These phases are usually done with several people, or a team.[71] This project will be undertaken by a single person. For that reason the sprint planning and reviews will be a bit shorter than with a project which involves a full team.

The advantages of using Scrum Sprints in this project, has resulted in a better overview of the state of the project. It has also been more motivating to work with the project, when the tasks have been broken down into smaller pieces as it might be hard to keep the motivation up when working on a project spanning well over a year. During the course of this project, the proposed weekly planning sessions and the monthly sprints became too rigid for the development speed. The sprints where instead organized around blocks of tasks that made sense to complete as one, which then where estimated.

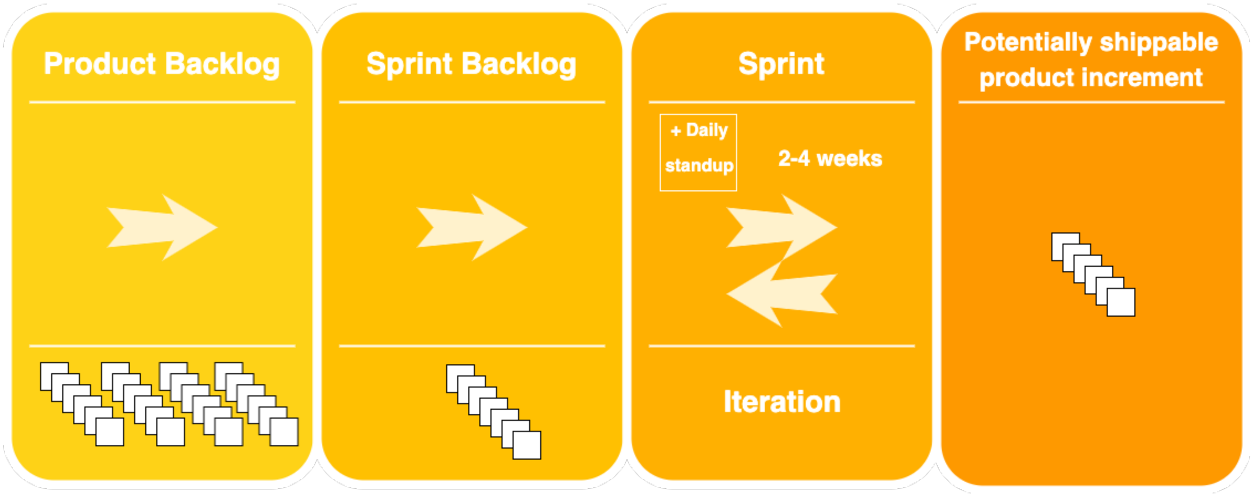


Figure 3.2: Scrum sprint

3.1.3 Design Science Research

Three Cycle View

This project will be following an iterative approach. This is in line with agile development, as mentioned above, as well as *The Three Cycle View*, which A. Hevner presents in the paper *A Three Cycle View of Design Science Research* [36]. This article describes an iterative approach to artefact development, by looking at what he describe as three closely related cycles of activities; *the Relevance Cycle*, *the Rigor Cycle* and *the Central Design Cycle*.

The relevance Cycle, connects the environment around the problem area, and the artefact being produced. It takes requirements from the environment around the artefact, and exposes the research to field testing. The thought behind this is that the artefact should improve the world, as mentioned above. If the artefact can be said to do so in a satisfactory manner, there might not be need for further iterations. But if it does not, the requirements are not satisfied and the artefact needs to be improved further with feedback from the problem space.

The Rigor Cycle, has to do with knowledge. This knowledge consists of research and practical methods, and it makes the basis of "rigorous design science research"[36]. In addition to this, there is also knowledge on what defines the state of the art in the environment specified, as well as what kind of artefacts and ways of doing things that already exists. This cycle implements the above mentioned information into the research project, so as to make sure it is inventive and original. This ensures that the research is something new and interesting to the field. It is key that the research remains relevant and a contribution, and builds on the research that already can be found in that field's knowledge base. Hevner writes that, "Additions to the knowledge base as results of design science research will include any extensions to the original theories and methods made during the research, the new meta-artifacts (design products and processes), and all experiences gained from performing the research and field testing the artifact in the application environment."[36].

The last cycle, namely the design cycle, is the main part of a project within the methodology of design science research. The two previous cycles can be said to provide design options. And these cycles needs to be repeated until a satisfactory design is achieved, which satisfies both the previous cycles as well as the last one. However, even thought the design cycle depends on the other relevance cycle and the rigor cycle, it is also independent. [36]

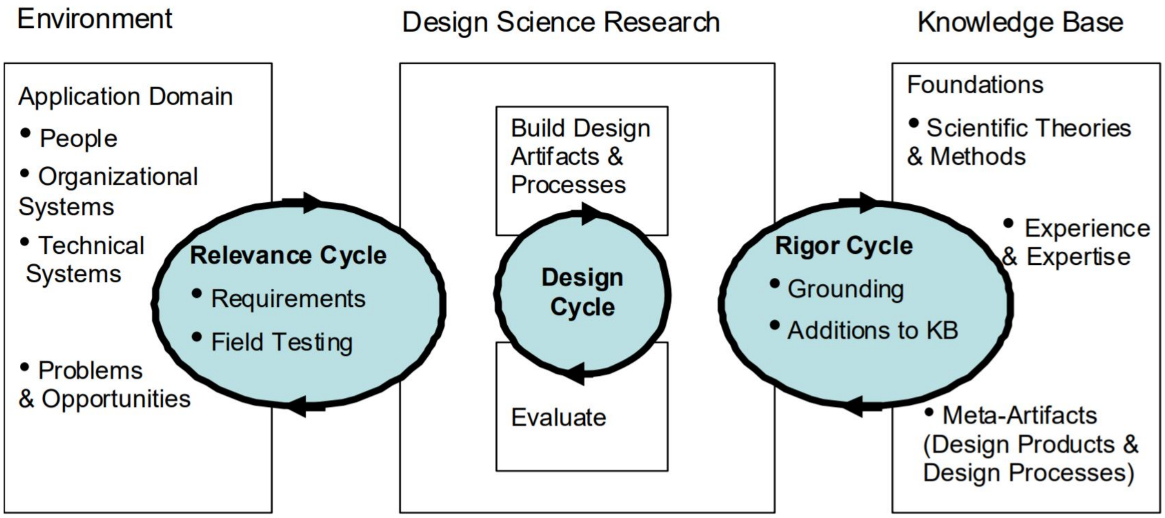


Figure 3.3: Design Science Research Cycles [36]

This approach makes sense to use in any project which produces an artefact. While the artefact presented in this project will not be of the tangible kind, or the kind that one can explicitly interact with, it is still a real concept. What following these principles of Hevner's means for the project at hand, is that it is always assured to be solving a real issue and contributing to science. The evaluation guides the design process which guides the artefact which is evaluated, and on and on it goes until one arrives at an artefact, a research contribution, and a project which has some leverage in the real world. These are factors this paper is aiming to achieve. This is to make sure this papers resulting artefact is contributing to research, contributes to the real world as well as has a design that makes the artefact the best it can be.

Throughout the work with this paper and the model which accompanies it, the work being done have been evaluated, rethought and redone as needed. The first approaches taken to training the model needed tweaks to the pre-processing of the data. In total four models have been trained thorough out the iterations with different variations of data, depending on the feedback gained from the previous model.

3.2 Data

There are some different data sources available for this project. The closer the data is the validation data, the better, but however it seems that the closer the data is in structure the more time consuming it would be to accumulate. Below follows an overview over some considered sources.

3.2.1 Collection of data

Data from Written Sources

Written sources are such as newspaper articles, written speeches, or other publicly available material would be gathered and then stripped of punctuation. The original would be test-data, and the stripped the train data. However, this model might not be able to generalize to text transcribed from audio, as this often is more colloquial in structure.

Data from Oral Source

The goal of this project is to predict punctuation. Punctuation prediction is, for obvious reasons, most useful in cases where there is no punctuation and it comes from natural speech. However, this makes training the model hard, as there is no truth to feed the model. A solution to this issue is to have the dataset manually annotated with punctuation. This is a very time-consuming task, and might be seen as a whole thesis just in and of itself. Therefore, the next best thing will be to find data, which has a high probability of having similarity to spontaneous speech which is the input and output for ASR technologies. A way of collecting data similar to ASR output from spontaneous speech, which is both oral and annotated with punctuation, is to transcribe for instance audio books, where a textual reference already exists. This is however not spontaneous, but the format would be that of ASR. The same can be said for subtitles and the audio they are based on. In this approach, the subtitle or the book would be the validation data, and the audio book or the sound for the video would have to be processed through a transcriber as to get the train-data.

Chosen data

The choice for training data fell on Norsk aviskorpus, which is described in the literature review. This data was chosen for the reason that it was available, but it is also written by many different people as it is comprised of newspaper articles. However, the writers have important shared factors which might make their style of writing similar, such as education, having an editor, and writing text within the same domain. This possibly makes the chance of high variety in the textual samples smaller than the number of writer would indicate. Nevertheless, articles from newspapers tend to contain interviews. This does not represent natural speech, however it might still present enough data so as to provide variety which will allow the model to generalize. One point which is quite peculiar for Norwegian and Swedish Journalism, is the use of the *en-dash(–)*. This is used to paraphrase the semantic content of quote, while still presenting it as a quote [56]. That means that also the interviews, where there might be hope of natural speech, probably will contain at least fewer direct quotes than a similar dataset for another language. Be that as it may, the chosen dataset seems the best choice considering the limitations and availability of datasets in Norwegian.

3.3 Evaluation

The approach generated by this work can be tested in several ways. The most relevant however is by metrics. Other solutions include crowd sourcing to check whether the predictions make sense and get someone who has expertise on the subject to go over and check that the punctuation is grammatically correct. The evaluation of the results needs to be tailored to the different aims outlined in this thesis. Automating the evaluation of the improved punctuation makes sense in this case, and it will be evaluated by metrics. This is also the preferred approach of the research field.

3.3.1 Metrics for Evaluation

In Tilk and Alumäe's *Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation restoration* [83] they evaluate the models per punctuation and overall precision, as well as F1-score and recall. In addition to this, they also use the Slot Error Rate (SER). The ERR metric is used for the model trained with the Europarl dataset [1]. This will also be the way this thesis will be undertaking evaluation of the model discussed in this thesis.

Positives and Negatives

When calculating precision, recall and F1, we use

- True Positive (TP)
- False Positive (FP)
- True Negative (TN)
- False Negative (FN)

Let's say that a model is going to predict whether something is True, or False, amongst 4 items. The goal the model is trying to predict is as follows: True, False, False, True. Let's say that the model predicts: True, False, True, False. This makes the first item a True Positive, the next item a True Negative, the third item a False Positive and the fourth a False Negative. The table below expands the example, and will be used for demonstrating the error metrics.

Truth	Attempt	TP	FP	TN	FN
True	True	1	0	0	0
False	False	0	0	1	0
False	True	0	1	0	0
True	False	0	0	0	1
True	True	1	0	0	0
True	False	0	0	0	1
False	True	0	1	0	0
True	False	0	0	0	1
True	False	0	0	0	1
True	True	1	0	0	0
Total:		3	2	1	4

Table 3.2: TP, FP, TN FN for error calculation

Precision

Precision is a metric which says something about how correct the predictions was. If a model has the task of predicting full stops, and predicts 5 out of 10 possibilities to be full stops, the precision metric is going to give us a metric saying something about how many of those 5 where actual full stops. The metric does not take into account the other 5 which were predicted to not be full stops.

$$Precision = \frac{TP}{TP + FP} \tag{3.1}$$

With data from the table over, the calculation would be as follows:

$$Precision = \frac{3}{3 + 2} = 0.6 \tag{3.2}$$

Recall

Recall rectifies an issue with precision; it takes some of the predictions which were not predicted to be full stops into account. When calculating recall we take the True Positives, and divide by the sum of True Positives and False Negatives. That is to say, we take the correct guesses, and divide them by the correct guesses as well as the times the model should have predicted a full stop, but did not.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

With data from the table over, the calculation would be as follows:

$$Recall = \frac{3}{3 + 4} = 0.43 \tag{3.4}$$

F1-score

The F1-score combines precision and recall into one score. It divides the product of precision and recall by the sum of precision and recall times two.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.5}$$

With data from the table over, the calculation would be as follows:

$$F1 = 2 * \frac{0.6 * 0.43}{0.6 + 0.43} = 0,5009 \tag{3.6}$$

We end with an normalized F1-score of about 50. This reflects the predictions well. There where three correct guesses, out of seven, which is a bit less than half. The model correctly identified about 43% of the correct cases of full stops (recall). It made five guesses that an item was a full stop, and three of these were correct which makes 60% of the punctuation guesses correct (precision). Combining these features, we land at about 50, which seems, given our results, a reasonable number.

SER

The F1-measure can be described as the weighted harmonic mean of recall and precision. J. Makhoul, F. Kubala, R. Schwartz et al. claim that this measure is problematic in some aspects, and the SER is proposed; it is equivalent to the aggregation of the different kinds of errors combined divided by the number of slots in the reference [61].

ERR

ERR can be used for saying something about whether or not the top result is relevant. This metric will never return a score which is perfect, which is 1.0 or normalized 100. The metric will allways continue to say that the result could be better. [61].

3.4 Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration

This section will address the way the study have been performed, so as to be easily reproduced. This thesis is in itself a reproduction of another study. It is taking another study and both reproducing their results, as well as trying out the approach on a new domain. This stresses the importance of a good and detailed recipe for how to approach the issue.

The first step of this project, have been to establish the problem space, and from here research questions and a hypothesis. Secondly, a literature study where performed, as well as a study into what kind of methods would work for the problem space at hand. From there, defining what kind of data is needed was done. The next steps where be to collect the data, clean it and pre-process it in a way that can be used by the model. Here the same approach as Alumäe and Tilk where adopted [1]. Next task was to develop the model, train, test, and validate it, and compute how well it is performing, using metrics. This is a computational expensive process, and this project has made use of an online, cloud-based virtual machine hosted by NREC [67], which can be used for processing, co-hosted by the *University of Bergen* and the *University of Oslo*.

The next step after initial training, testing and validating, was to improve the model. Building, training and testing the model is an iterative process, where each part might trigger a re-iteration of any of the other parts depending on how the model performs, if hypotheses are dis-proven or new research comes to light. This has happened during the course of this study, mostly because of the training data used were not properly cleaned and was containing to much contextual data and formatting.

In conclusion, the study has largely had a structure based on design science research, by its iterative nature. It has also used Agile principles.

3.4.1 O. Tilk and T. Alumäe's approach

In the end of 2016 authors Tilk and Alumäe published the paper *Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration* in context of the *Annual Conference of the International Speech Communication Association, INTERSPEECH*. They write that their approach can "utilize long contexts in both directions and direct attention where necessary enabling it to outperform previous state-of-the-art on English (IWSLT2011) and Estonian datasets by a large margin" [83].

Tilk and Alumäe are using two methods in their paper. They have one approach where they are using text only, and one approach where they use both text and prosodic features. The first approach is for English language, and they are using the IWSLT-dataset. They state that the reason for not using prosodic features, is that they do not have those available for English. The other approach is for Estonian language. By using the same base approach for two different languages, it shows good promise for generalizing further to Norwegian. For the Estonian approach, they use both textual, as well as a prosodic feature. The prosodic feature they are using is the pause duration between spoken words. They way they incorporate both the prosodic features as well as the textual features, is by training the same model in two stages.

Detailed Description

The approach in this paper, is a BRNN which is used in combination with an "attention mechanism for punctuation restoration in unsegmented text"[83]. The attention mechanism is fused together with the model state by a late fusion approach which they have adapted to GRU from LSTM. This makes it possible for the result of the attention model to interrelate with the state of the model without affecting the state memory.

One-hot encoded sequences of input words $X = (x_1, ..., x_T)$ are processed first by two recurrent layers, which are preceded by the same embedding layer with weights W_e, with GRU units making up a bidirectional layer. The two recurrent layers process the encoded sequence both forwards and backwards. State $\overrightarrow{h_t}$ at the step for time t of the first recurrent layer which moves forward is the following:

$$\overrightarrow{h_t} = GRU(x_t W_e, h_{t-1}) \tag{3.7}$$

GRU here corresponds to the approach described in K. Cho, B. van Merrienboer, C. Gulcehre et al. [20], however Tilk and Alumäe does not use added biases like shown in the referenced study. Gated Recurrent Unit (GRU), is at type of Recurrent Neural Network, which is similar to LSTM, but they take less computational power and are thereby less expensive to run. GRUs "saves" from an earlier time, and uses this for coming calculations.

Tilk and Alumäe uses *tanh* as the "new hidden state nonlinearity ϕ"[83]. *tanh* is often used to get around the vanishing gradient problem. *tanh* is useful here as the second derivative of the function before going to zero, can handle a range of considerable length. In the equation below, the reverse step is similarly processed, however it is done the other way around. The equation depicts both direction of the recurrent neural network, and by uniting them we see the bidirectional state h_t.

$$h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t] \tag{3.8}$$

For each of the input words, the layer described will learn representations. These will both rely on the previous and the following context. Tilk and Alumäe expects this to assist the model to better identify both when a question is asked, and when a sentence starts or ends as the model is given more information. Whether or not a sentence is a question relies very much on where words are placed in sentences. One example being ***What*** *do you think?* and *This is* ***what*** *I think.* A model being able to see both backwards and forwards might be well placed to predict questions.

After the above described layer, Tilk and Alumäe introduces a layer with attention mechanism; a GRU layer which moves in only one direction as opposed to the bidirectional layer. As this layer only goes forward, is used for going through the states from the bidirectional layers and monitor the current position as the attention mechanism moves back and forth looking for "relevant bidirectional context aware word representations"[83]. State s_t of the layer

$$s_t = GRU(h_t, s_{t-1}) \tag{3.9}$$

"is late fused with the attention model output α_t which is computed based on the previous state s_{t-1} and bidirectional layer states $H = (h_1, ..., h_T)$ [83]. Now it is time to combine the result from the attention mechanism and the state of the model, as described by [7]. We then have the late fused state f_t. This can be represented as:

$$f_t = a_t W_{fa} \circ \sigma(a_t W_{fa} W_{ff} + h_t W_{fh} + b_f) + h_t \tag{3.10}$$

Finally, the model will produce the probability for punctuation. This happens when the late fused state, described above is taken as input to the output layer. This produces the predicted punctuation y_t, at the step t time. y_t is represented as:

$$y_t = Softmax(f_t W_y + b_y) \tag{3.11}$$

A figure of the model described [83] can be seen in Figure 3.4[83]

The approach described above, represents the first stage of the training, in the approach proposed by Alumäe and Tilk. As mentioned, they also have proposed a second step to their model, which takes prosodic features into account. They use this approach for the Estonian data, but not for the English data. The reason they did not use the compound approach for both languages, is the lack of prosodic features available to the authors for English. For this thesis, the same issue is present for this thesis and the data collected for building the model. The author has not been able to find or compile data which contains prosodic features that can be used for training.

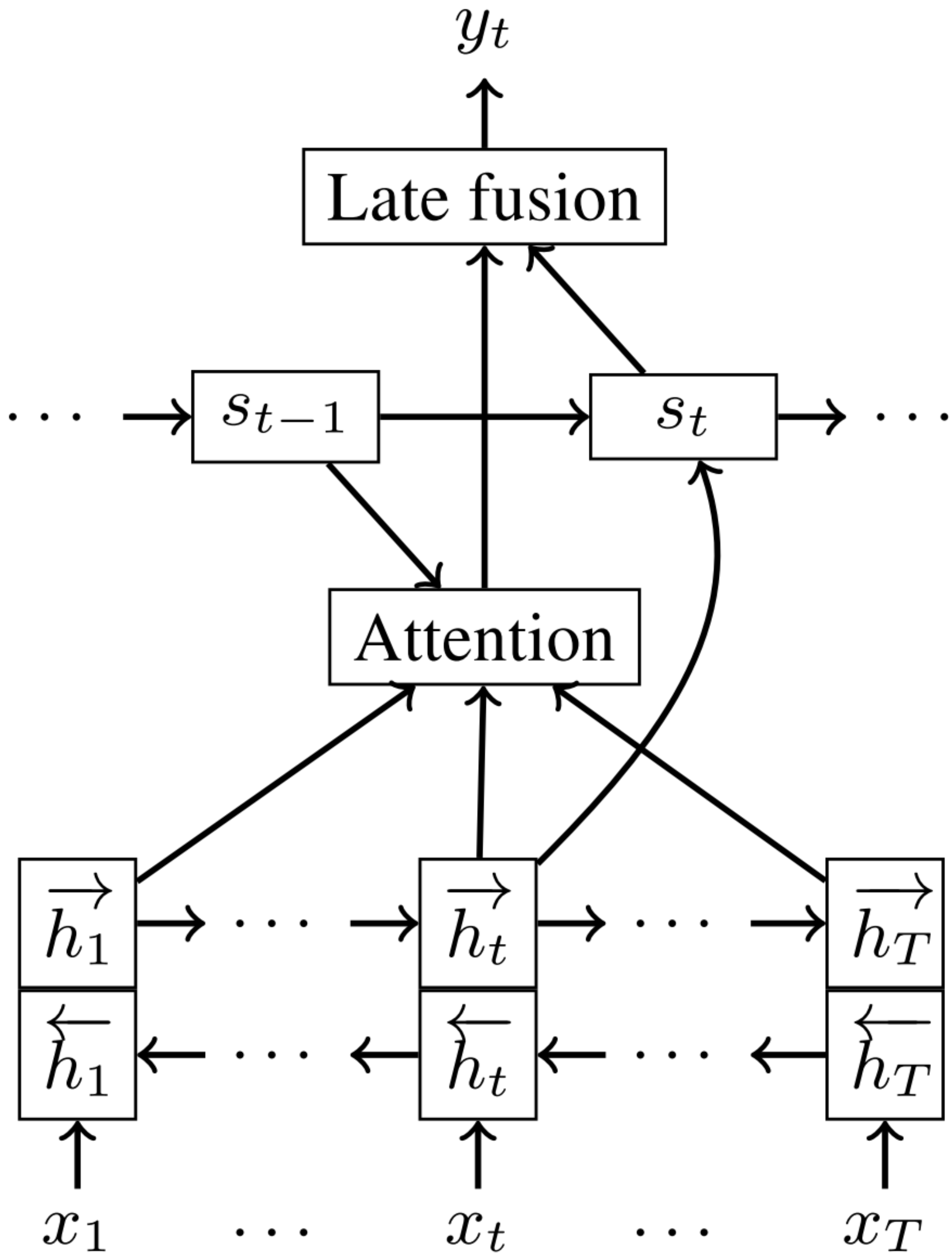


Figure 3.4: Description of Tilk and Alumäe's model predicting punctuation y_t at time step t for the slot before the current input word x_t. [83]

Training

Tilk and Alumäe have been training their models with a learning rate of 0.02. The weights in their approach is updated with AdaGrad. It might very well be chosen for the fact that there is no need to manually adjust the learning rate. During the training, the negative log-likelihood is minimized. The training is completed when the perplexity for validation worsens for the first time when training the first model. The input to the model is divided up into slices, where each contain 200 words. If the slices causes a sentence to break, the sentence in question will be moved to the next slice. A tactic Tilk and Alumäe have been using to reduce training time, is to shuffle the slices before each of the epochs, and organized into mini-batches counting 128 each.

The vocabulary for the English dataset, is 27 244 words, and in additions comes two special tokens (*end-of-sequence* and *out-of-vocabulary*). The dataset used for the English model is the IWSLT-dataset, which is described in the introduction of this thesis. Here, the training and development datasets are respectively ca 2 100 000 and ca 296 000 words. The ASR set that is used for validation as well as their reference dataset contains about 13 000 words each.

Estonian and English showed improvements for all punctuation types compared to the state-of-the-art at the time when Alumäe and Tilk's paper was published. The overall F1- score was improved by 1.810.5% and SER was reduced by 2.6 15.5%. The biggest improvements were achieved when comparing text-only models to the compound-approach. [83]

For the English model, the overall F_1-score has been bettered by 8.9%, on the reference text, as well as being 10.5% better when comparing the output from the ASR punctuated data with their baseline. The baseline for this model is the DNN model from [17]. The model shows improvement for all metrics, when compared to the baseline results. A difference that the authors state is that the prediction for question marks is significantly improved, where the baseline results seemed unable to predict them due to "the limited fixed size context". They further states that factors point towards that the following context is important for predicting sentence boundary, and one needs to look further than one word ahead.

Puncutator2

The authors of this paper has made their source code available on GitHub [32]. The repository contains most of the code needed for reproducing their approach.

3.5 Approach

Literature Review

No research projects exists in a vacuum, and we are dependent on earlier research to move forward. During the work on this thesis, *Google Scholar* has been the search engine most frequently used to find articles and research on the subject of this project. In the field of punctuation restoration, there is to the authors knowledge, no research done in the domain of Norwegian language. There is however recent work on punctuation for European languages, and amongst those Norwegian is mentioned. There is also some multilingual approaches.

3.5.1 The Model

This thesis aims to explore whether an approach made for different languages can be applied to Norwegian. To do this, one need access to mainly two entities; the code for making the model or a very detailed description of the approach, as well as similar training data.

Baseline

What makes the most sense for this project, is to make the baseline to compare the results produced from this thesis, the results gained from O. Tilk and T. Alumäe in their work. The results i use as baseline is the model that was produced for English, without the prosodic feature.

PUNCTUATION	PRECISION	RECALL	F-SCORE
?QUESTIONMARK	77.7	73.2	75.4
!EXCLAMATIONMARK	50.0	0.1	0.1
,COMMA	68.9	72.0	70.4
-DASH	55.9	8.8	15.2
:COLON	60.9	23.8	34.2
;SEMICOLON	44.7	1.1	2.2
.PERIOD	84.7	84.1	84.4
Overall	75.7	73.9	74.8

Table 3.3: Table depicting the results from Alumäe and Tilk using the English Europarl v7 with Training set size: 40M words

Validation

Validation of the model will be done both with in-domain data, as well as out-of-domain transcribed data. The transcribed-dataset used for validation has been extracted from The Norwegian National Library's collection of parliament speeches, and is still in a beta-stage. It has very recently become available, and has therefore not been used for training-data for the model. However, a small part of the dataset has been used for validating the model on, and will contribute to give a better view of how the model generalizes.

3.6 Architecture and Technical Prerequisites

3.6.1 Norwegian Research and Education Cloud (NREC)

Machine learning and deep learning require an abundance of data. Often, the more the better. Therefore training models is a very computational expensive process. If the machine used does not have the possibility of utilizing GPU, it will be a very slow process. For this reason, the project have been hosted on NREC-architecture. However, the training time for the main model has still taken between 2-4 days to complete. NREC is formely known as UH-IaaS, and is a collaboration between the University of Oslo, and the University of

Bergen, Nordic e-Infrastructure Collaboration (NeIC) as well as Uninett. The service have been available since 2016, and provide infrastructure to several academic projects of size. They e. g. facilitate the infrastructure of CERN'S ALICE and ATLAS experiments. Their software is also based on Open Source (Open Stack), and is thereby a very transparent actor in the sphere of cloud computing [67].

The setup used for this project, is the one with the most computational power possible within the NREC architecture available to the author. The instances are created with Ubuntu 19.04. Instances are in this context virtual machines (VM) which are cloud-based. This means that they can be accessed remotely, and processing done on the instances does not affect the machine you access the instances from. Instances can be create from several sources, the most common might be snapshots and images [68]. In this project, snapshots where used to be able to run several model trainings at the same time.

After the code and training data was set up, the VM was snapshotted, and a new instance was created from it. This allowed for faster computation and freedom to test several approaches in a shorter amount of time.

Before arriving at the preferred setup for the instances, there was some trial and error involved. However, the configuration which ended up being final is described below. They are set up with the flavour *m1.xlarge*. A flavor in OpenStack says something about the storage capacity, the memory and the computing power of a virtual server [69]. The instances run on the operating system Ubuntu 19.04 (disco). The RAM available for the instances is 16GB, and they have 4 VCPUs. The hard drive is 20GB.

Instance configuration NREC

Operating System	Ubuntu 19.04 (disco)
Flavour	m1.xlarge
RAM	16 GB
VCPU	4
Disk space	20 GB

Table 3.4: Configuration for NREC instances used in this study

Chapter 4

Results

During the work on this thesis, three different models have been trained. One is trained on the IWSLT dataset, another is trained on a Norwegian Bokmål dataset, and the third is trained on a Norwegian Nynorsk dataset. The purpose for training the English model, is to create a baseline. This model is trained under the same circumstances as a Model trained by Alumäe and Tilk, and can therefore be used as a confirmation towards the correct use of the approach described in the Methods-chapter. The second model, is the model which is the artifact of this thesis. It is a model trained with Norwegian data, based on an approach created for other languages. The third and last model is thought as an experiment - to see whether there is a difference in training models for Norwegian Bokmål, and Norwegian Nynorsk, as well as how the approach responds to data sparsity.

Excerpts of the model-punctuated data can be found in the appendix.

4.1 Reference model trained with English Europarl v7

Authors Tilk and Alumäe trained their English model with the IWSLT-dataset, and achieved good results. They have also made a model with the English Europarl v7 dataset, which outperforms the old English model. This model is also trained only on second stage data. To validate that this thesis is achieving correct results according to the model that Tilk and Alumäe built, a model has also been trained for this paper with the English Europarl

Data. This is to verify that the results achieved in this study are in accordance with the results presented by Alumäe and Tilk. If such a baseline is not established, it will be hard to know whether the approach being tested actually generalizes to other languages or not. The Europarl-model is chosen for two reasons; the results achieved with the Europarl dataset for first stage data is better than the results achieved with the IWSLT dataset. Secondly, the dataset is made available by Tilk and Alumäe, and is therefore a good option for baseline. The model is trained with a hidden layer size of 256 and a learning rate of 0.02. The exact hidden layer size and learning rate for the model built with the Europarl dataset is not described, however the learning rate and hidden layer size chosen is the ones described in their article [1]. The results achieved in this study when training their model with the same approach were similar to the results Tilk and Alumäe achieved, however there are some discrepancies.

PUNCTUATION	PRECISION	RECALL	F-SCORE
?QUESTIONMARK	78.6	69.7	73.9
!EXCLAMATIONMARK	50.0	0.1	0.1
,COMMA	66.1	68.5	67.3
-DASH	51.1	5.9	10.5
:COLON	49.8	24	32.4
;SEMICOLON	69.6	2.8	5.5
.PERIOD	82.7	82.8	82.7
Overall	73.1	71.0	72.0
Err	4.24%		
SER	43.7%		

Table 4.1: Results from training baseline with Europarl English dataset

PUNCTUATION	PRECISION	RECALL	F-SCORE
?QUESTIONMARK	77.7	73.2	75.4
!EXCLAMATIONMARK	50.0	0.1	0.1
,COMMA	68.9	72.0	70.4
-DASH	55.9	8.8	15.2
:COLON	60.9	23.8	34.2
;SEMICOLON	44.7	1.1	2.2
.PERIOD	84.7	84.1	84.4
Overall	75.7	73.9	74.8

Table 4.2: Table depicting the results from Alumäe and Tilk using the English Europarl v7 with Training set size: 40M words

As seen by the two tables above, the results are quite similar, however the model trained for the work with this thesis, Table 4.1, is routinely beaten by for the most part small amounts by the model trained by Alumäe and Tilk, Table 4.2, or scores the same. There are a couple of exceptions however. For precision for question marks, where the baseline model has outperformed the reference model with just shy of a point. Also the metrics for semicolons shows that the baseline-model performs quite a bit better than the reference model when it comes to precision. The baseline-model is outperforming the reference model with 24,9 points. It also slightly outperforms the reference model when it comes to F-score for semi-colons. The overall metrics show however that the metrics produced by the two models are quite similar, and deviates with about three points in general in the favor of the reference model. This seems therefore to be an acceptable baseline, even thought the result from Alumäe and Tilk has not been exactly reproduced. As can be seen by the Figure 4.1 which combines all the different punctuation symbols and metrics for the two different models. The figure shows that the results correspond well with each other, and have similar structure even thought they are some points apart..

It is hard to say where the differences originates from, but one good guess is that the part of the datasets used for training, test and development are different. When the model is trained, the data is also shuffled. The slices will differ as they are shuffled at random. There is also a chance the pre-processing of the data have differed, however the exact approach described [1][1] where used, to the best of the authors knowledge. An additional possibility is that the learning rate and hidden layer size might differ. Nevertheless, it is common and makes sense that results with the same training data differs in training models based on neural networks where training data is shuffled at random.

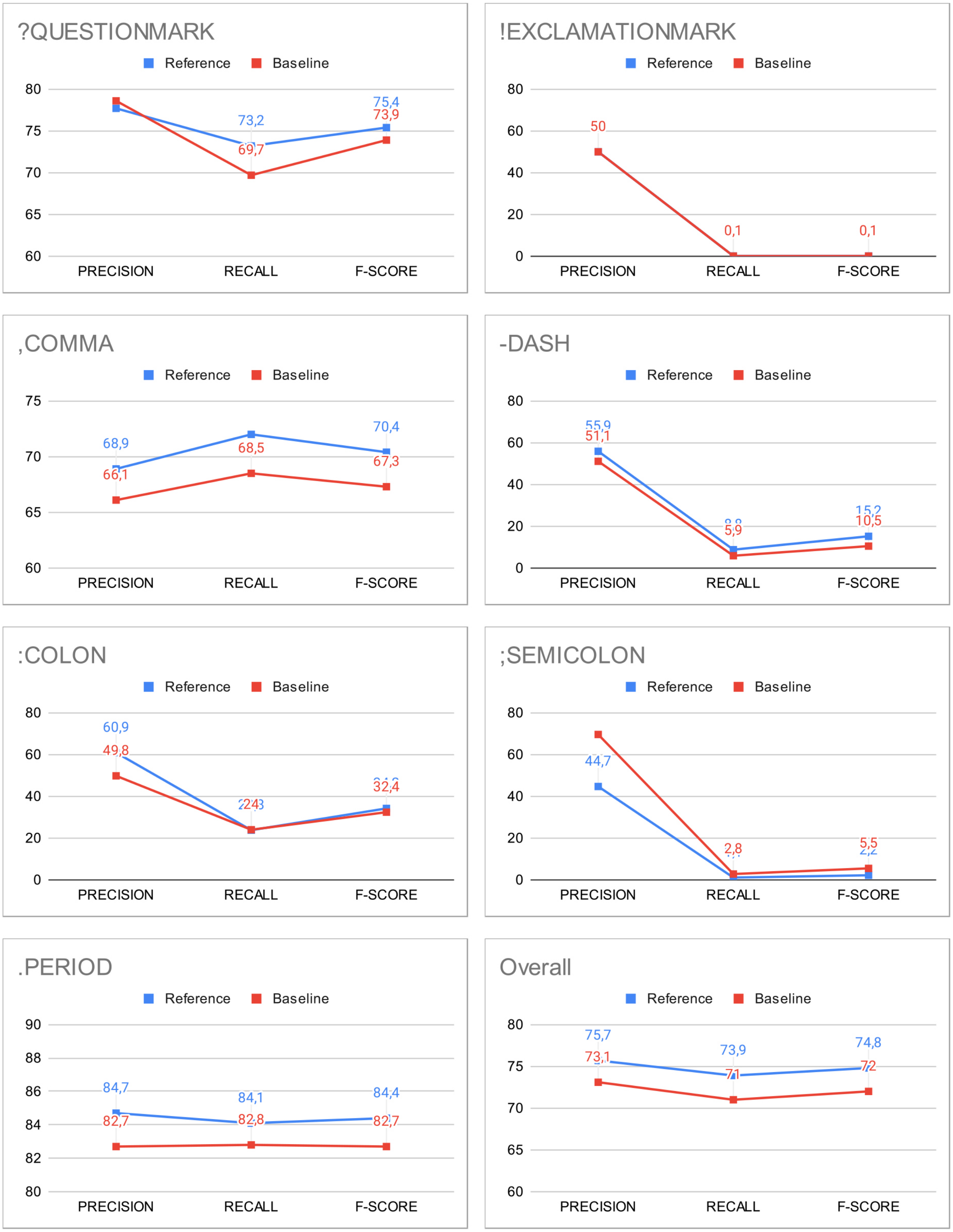


Figure 4.1: Comparison between baseline and reference model for English

4.2 Norwegian model

After the baseline model was trained and evaluated, the work of establishing the Norwegian model started. The Norwegian model was trained with data from *Norsk Aviskorpus* which is described in the literature review. The model had to be trained more than once, and the training data tweaked and cleaned. A part from necessary cleaning caused by the structure of the training data, the Norwegian data was pre-processed the same way as the data for Alumäe and Tilks model.

When choosing the configuration for training, the author have here opted for choosing the same hidden layer size and learning rate as the baseline model. The hidden layer size is 256 and a learning rate of 0.02. This is also the same configuration as Alumäe and Tilks models [83].

When training the model, the best validation perplexity was 1.1134. Perplexity, which is used for finding out when to stop training the model, is an intrinsic method for evaluation. A definition can be that it is the probability in inverse for the test set, which is then normalized by the word count. The vocabulary for the trained model for Bokmål is 100002.

Training of Norwegian model	
Total number of training labels	46412800
Total number of validation labels:	7087360
Best validation perplexity	1.1134

Table 4.3: Training labels, validation labels and validation perplexity for Norwegian model

PUNCTUATION	PRECISION	RECALL	F-SCORE
?QUESTIONMARK	77.7	73.2	75.4
!EXCLAMATIONMARK	50.0	0.1	0.1
,COMMA	68.9	72.0	70.4
-DASH	55.9	8.8	15.2
:COLON	60.9	23.8	34.2
;SEMICOLON	44.7	1.1	2.2
.PERIOD	84.7	84.1	84.4
Overall	75.7	73.9	74.8

Table 4.4: Table depicting the results from Alumäe and Tilk using the English Europarl v7 with Training set size: 40M words

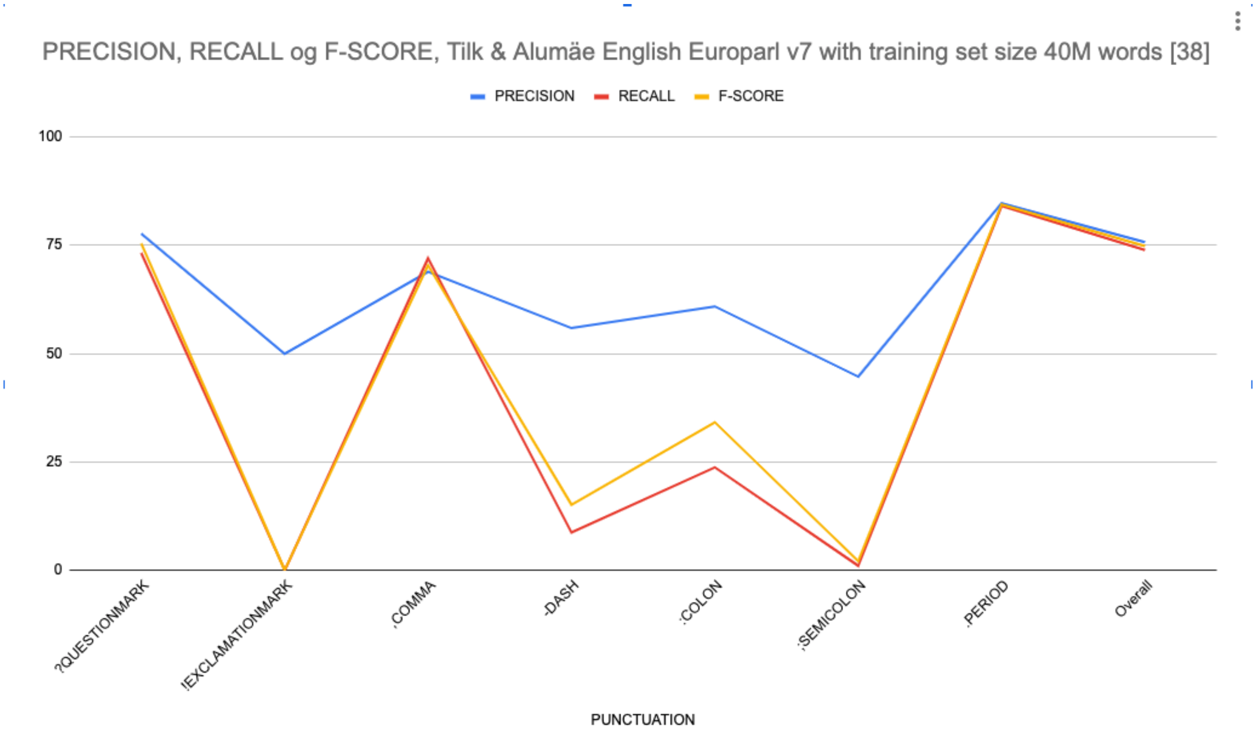


Figure 4.2: PRECISION, RECALL and F-SCORE, Tilk and Alumäe English Europarl v7 with training set size 40M words [38]

PUNCTUATION	PRECISION	RECALL	F-SCORE
?QUESTIONMARK	70.3	16.8	27.1
!EXCLAMATIONMARK	37.2	1.6	3.1
,COMMA	73.6	62.2	67.4
-DASH	75.7	34.6	47.5
:COLON	63.5	40.1	49.2
;SEMICOLON	79.3	6.1	11.3
.PERIOD	91.5	90.7	91.1
Overall	85.5	79.1	82.2
Err	3.68%		
SER	30.4%		

Table 4.5: Results Norwegian model trained with Norsk Aviskorpus

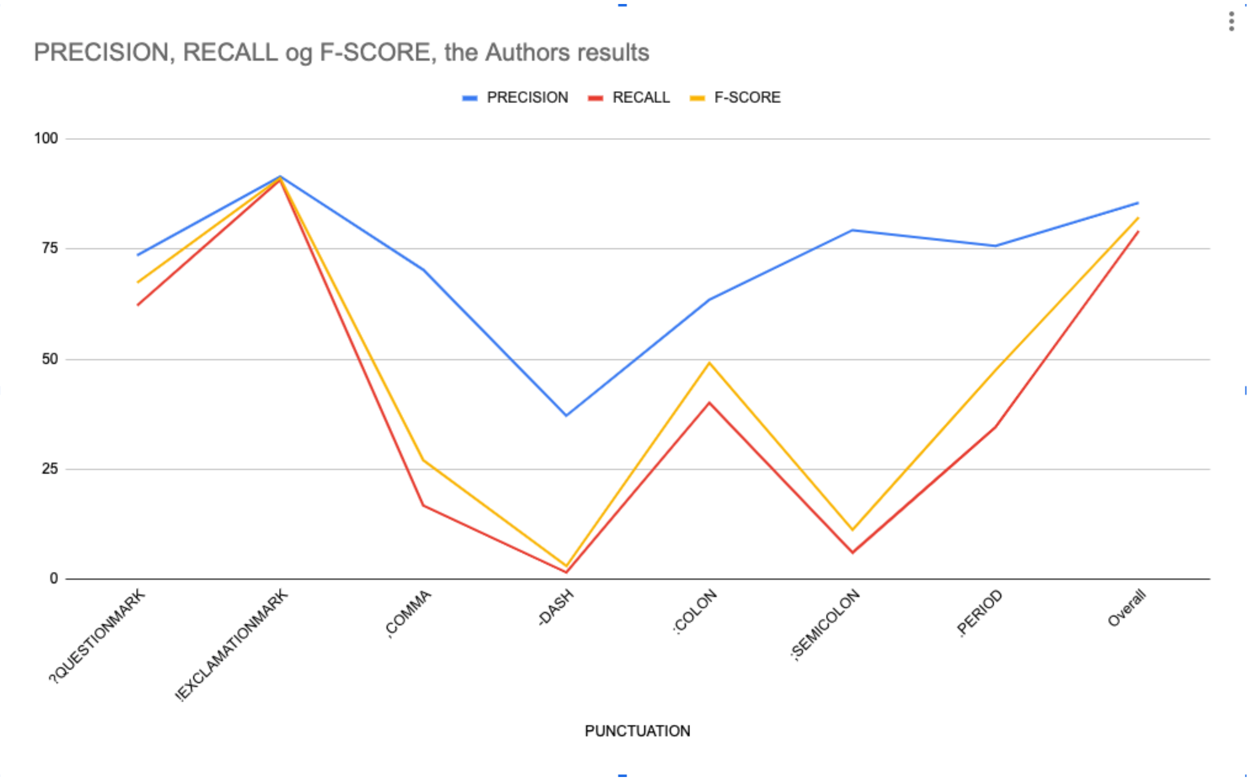


Figure 4.3: Results for Norwegian model

The error metrics are calculated on in-domain-data. When examining the results in Table 4.4, it becomes obvious that the metrics are quite good compared with the results from Alumäe and Tilks model trained on Europarl, which can be seen in Table 4.2.

For question marks, the precision metric is bit better in the English model, and it keeps the metric both for recall and F-score. The Norwegian model, achieves good precision for question marks, but low recall and thereby a quite low F-score. This tells us that while a lot of the question marks predicted where correct, there were a lot of question marks which where not targeted.

When it comes to exclamation mark, we see the same trend for both models as we saw for question marks for the Norwegian model; a big difference in the precision and recall score. Also here the English model is better. Exclamation marks, as well as question marks are however tricky. They can both be replaced by a full stop, depending on the conditions.

For predicting comma, the Norwegian model shows better precision, but is worse when it comes to recall and the combined F-score.

Dash however, the Norwegian model does well. It is better than the English model both when it comes to precision and recall, as well as F1. The values for the dash is the second highest discrepancy between the Norwegian and English model there is, with almost 20 points in difference for precision. One reason for t his, as was mentioned in the methods chapter, might have something to do with the domain the text is from. As this is texts from newspapers, and Norwegian journalists are fond of using the dash as a way of quoting people, this might be why the model has such good grip on it: It usually comes before a sentence. If this is the case, it can be considered bad news for how well the model will handle out-of-domain text.

Second to last, we have the semi-colon. Here we have similar metrics for both models. However, the semi-colon score for precision for the Norwegian model is about 35 points higher than for the English model. For both models we see the same trend as for the exclamation mark, with a drop from precision to recall.

Lastly, we have the full stop, or the period symbol. Here, both models perform very consistently, with minimal difference between precision, recall and F-score. The Norwegian model has precision of 91.5 for punctuation, together with a recall at 90.7 and an F-score of 91.1. These results are very good, and to the authors knowledge, the best metrics for full stops achieved by any study. The state of the art described in the literature review, have their best results for precision for periods at 86.1, recall at 89.3 and f-score at 87.7. A side-by-side comparison can be seen in Figure 4.3. Alumäe and Tilk's model trained on the IWSLT dataset only contain data for comma, period, question marks and overall score. Here, the Norwegian model beats all the scores available, a part from the ones for question marks, where Alumäe and Tilk's model perform better for all metrics. Table showing results from [83] can be found in the appendix, Appendix A.7. One of the reasons Norwegian might beat the English models when it comes to punctuation, might be for the reason that Norwegian often consists of short sentences. Which again should provide more punctuation per word, and therefore more opportunities for the model to develop a pattern. Another reason to why the Norwegian model might be outperforming the other models, is that it is mostly comes from structured textual data which has never been spoken. There also some interviews, but the samples taken from the datasets shows that it consists mostly of articles.

Also, the over-all metrics for precision, is better than the approaches described in state of the art which desribe an overall precision is 84, while the result from this study offers a slight improvement by 1.5, and ends on 85.5. The overall F1-score is however better in M. Courtland, A.Faulkner and G. McElvains's approach [24] where they achieve overall F1 of 83,9 and overall recall of 85.9, while the Norwegian model gets an overall recall score of 79.1.

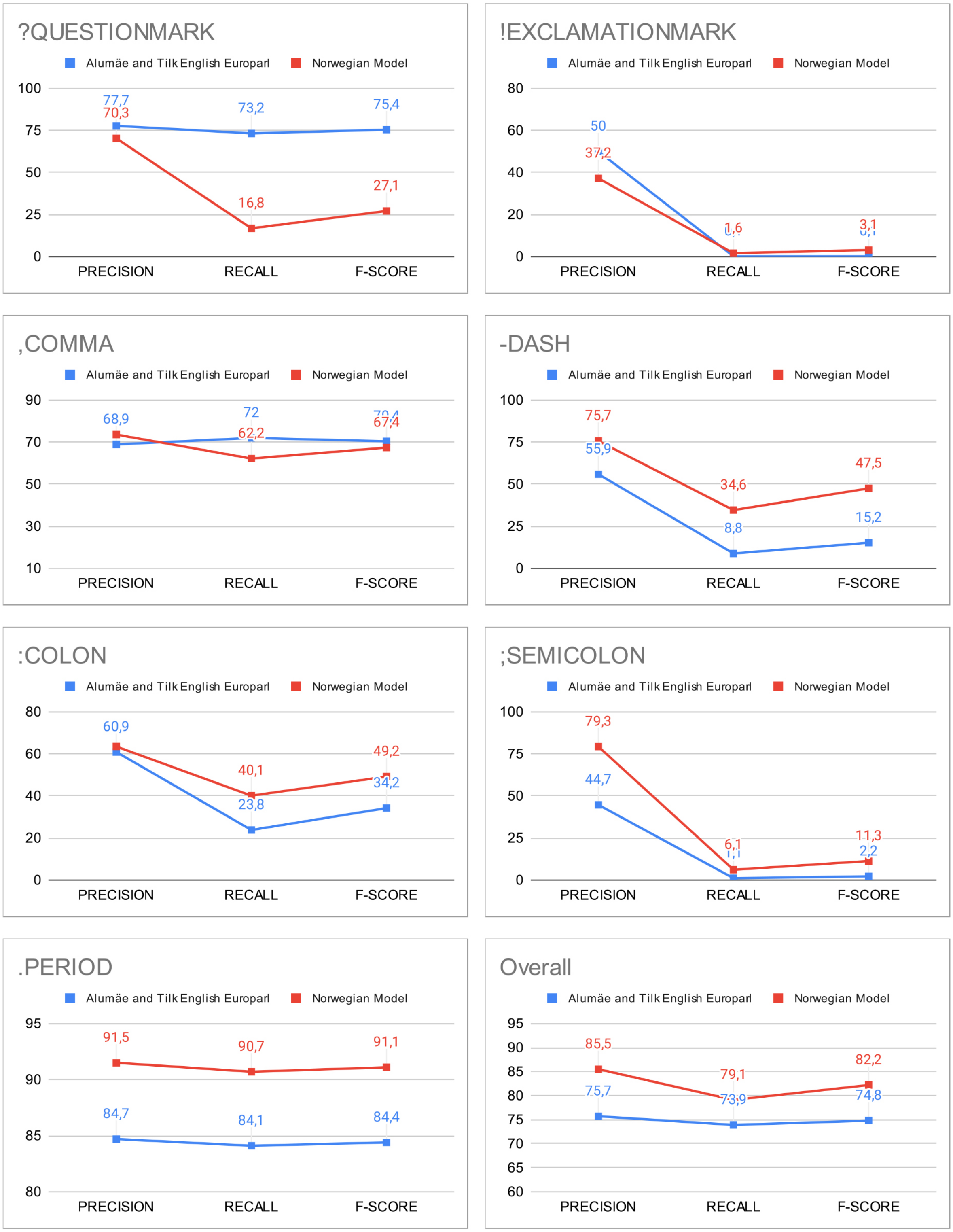


Figure 4.4: Result comparison between the Norwegian model and Alumäe and Tilks Europarl model

4.2.1 Validation

To validate the model, out-of-domain transcribed data has been used. The data in question is described in the literature overview. This dataset is in many ways similar to the English Europarl dataset, as it is transcriptions from political proceedings. The whole set available is not used for validation. This have to do with the late publishing date of the data, considering the deadline for this thesis. The data is predominately made for ASR, and is therefore not organized in a way which makes it easy to make large, cleaned files. There was not time to organize and pre-process a larger amount of the data within the bounds of the time limits for this thesis. The files contain, as mentioned in the literature review, contextual data and are only available in several .ref files. After cleaning and compiling the data, the data was pre-processed same way as the data for the training of the models. However, as the dataset is rather small, only periods, question marks and commas seems to have had enough samples for the error calculation in the text. The number of words left in the validation set after the pre-processing was just shy of 9000 words. This dataset where then punctuated by the Norwegian model. After punctuating, error calculation was performed. The only relevant data to take away from the validation, is the metrics for comma and periods. The rest of the metrics have not had enough samples to be calculated.

The model does not perform particularly well when it comes to predicting commas, when looking at the reference-model (Alumäe and Tilk's Europarl model). It does a lot better when it comes to periods, however, the results are not exceptional.

As the model does well on the in-domain-dataset, it stands to reason that the model built is overfitted, or overly specialized on a field or domain. The training data might be too specific. That might make sense, when one considers that all the text mostly is written by individuals sharing the same profession. The producers of the texts are journalists. The one's that are educated as such share similar training in writing, and a common denominator of newspapers is that it they have an editor, and a format which articles are produced in. The validation dataset is comprised of more or less spoken natural language, with experts going over the data and annotating it with punctuation. The other thing that might explain the discrepancy in results, is the sparsity of the validation data. There might not be enough data to calculate useful metrics. The data might also be of bad quality. The data have been cleaned, but due to the time constraint of the dataset being released in the end of may, there might have been an oversight.

PUNCTUATION	PRECISION	RECALL	F-SCORE
?QUESTIONMARK	100.0	2.0	4.0
!EXCLAMATIONMARK	n/a	0.0	n/a
,COMMA	37.2	33.2	35.1
-DASH	n/a	n/a	n/a
:COLON	0.0	0.0	n/a
;SEMICOLON	n/a	n/a	n/a
.PERIOD	69.9	24.2	35.1
Overall	58.9	23.8	33.9
Err	14.01%		
SER	83,4.7%		

Table 4.6: Error metrics for validation of Norwegian model using excerpt from the Stortingskorpus

Small Qualitative Validation

This is the textual example which was used in the literature review, to test different ASR technologies, to verify that they did not provide punctuation, and to assess how well they perform in general. It has been stripped of punctuation and capital letters.

> er det egentlig så farlig med den tegnsettingen jo tegnsetting er like viktig å kunne somå stave ordene riktig selv om noen velger å se på det som en ikke sak noen ganger er tegnsettingen avgjørende for setningens mening det klassiske eksempelet som alle lærer på barneskolen er setningen heng ham ikke vent til jeg kommer plasseringen av kommaet etter henholdsvis ham og ikke avgjør den

> dødsdømtes skjebne det kan altså være livsviktig med riktig tegnsetting plutselig har man voldt noens død

Here is the same text, but it has been punctuated by the Norwegian model.

> er det egentlig så farlig med den tegnsettingen jo tegnsetting er like viktig å kunne som å stave ordene riktig ,COMMA selv om noen velger å se på det som en ikke sak noen ganger er tegnsettingen avgjørende for setningens mening .PERIOD det klassiske eksempelet som alle lærer på barneskolen er setningen heng ham ikke vent til jeg kommer plasseringen av kommaet etter henholdsvis ham og ikke avgjør den dødsdømtes skjebne .PERIOD det kan altså være livsviktig med riktig tegnsetting .PERIOD plutselig har man voldt noens død

It performs rather well, and all the punctuation marks that have been placed are correct, except the last one, which should have been a semi-colon. The model is fairly confident when it actually places the punctuation marks, but it misses out on some. This explains some of the low recall scores. If one looks at the punctuated text in appendix A.2, one can see that the punctuated text seems to have a good flow. The authors sees few punctuation marks that is in need of changing.

4.3 Experiments

As a way of researching the generalizability of the model, two experiments where performed. One was to punctuate Norwegian data from *Norsk Aviskorpus Bokmål* with the English baseline model to investigate whether trained models could be used on more than one language. The other experiment performed, was to try out the much smaller dataset *Norsk Aviskorpus Nynorsk*, to see if there are any discernible differences between Norwegian Bokmål and Norwegian Nynorsk. The second reason for trying the Nynorsk dataset was to see how well the approach works with smaller datasets, which can further shed light on the issue of under-resourced languages which might not have large datasets available.

4.3.1 Norwegian data punctuated with the English model

To check whether there might be a possibility for achieving good results by simply feeding Norwegian text to a model trained with English data, the same data used for testing the Norwegian data with in-domain-data, where fed to the English model. The results can be seen in the table below:

PUNCTUATION	PRECISION	RECALL	F-SCORE
?QUESTIONMARK	33.1	5.1	8.8
!EXCLAMATIONMARK	0.0	0.0	n/a
,COMMA	5.7	33.0	9.7
-DASH	0.0	0.0	n/a
:COLON	0.7	0.8	0.7
;SEMICOLON	0.2	0.0	0.0
.PERIOD	12.7	2.4	4.1
Overall	6.1	13.1	8.3
Err	32.38%		
SER	267.4%		

Table 4.7: Results English tested with Norwegian Bokmål newspaper data

As expected, the attempt went rather poorly. If one examines the excerpt from the punctuated data in the appendix A.6, there is an abundance of commas. This corresponds with the metrics we see for commas. Very few of the predicted commas are accurate, however the model actually managed to hit a fair bit of the true positives. One theory towards whey the model spews out commas, might also have to do with the fact that Norwegian language uses rather short sentences. At least a fair bit shorter than English grammar allows for. When it comes to question marks, the metrics are more or less flipped; while a third of the predicted question marks are correct, the recall-score of 5.1 tells us that the model was to tentative. With a SER of 167.4 percent, the conclusion must be that this is not an approach which can be adopted for Norwegian, or any other under-resourced languages.

4.3.2 Norwegian *Nynorsk* model

The pre-processed training data for the model for Nynorsk, only have a vocabulary of 9579 words. That is 1/10 of the vocabulary when comparing to the Bokmål model. The number of token for Nynorsk is 16 070 002. This is 0.8% of the tokens used for the Bokmål-model.

PUNCTUATION	PRECISION	RECALL	F-SCORE
?QUESTIONMARK	0.0	0.0	0.0
!EXCLAMATIONMARK	0.0	0.0	0.0
,COMMA	64.5	17.1	27.0
-DASH	42.9	0.4	0.8
:COLON	12.7	0.8	1.4
;SEMICOLON	0.0	0.0	0.0
.PERIOD	75.3	56.4	64.5
Overall	71.1	31.6	43.8
Err	8.43%		
SER	70.4%		

Table 4.8: Results Norwegian Nynorsk model trained on *Norsk Aviskorpus Nynorskdelen*

Training of Norwegian model	
Total number of training labels	263424
Total number of validation labels:	7306880
Best validation perplexity	1.2916

Table 4.9: Training labels, validation labels and validation perplexity for Norwegian *Nynorsk* model

The model trained for Nynorsk, performs surprisingly well. It performs best on periods, as is expected, but preforms surprisingly well on dashes. This might have to do with the previous mentioned curiosity in Norwegian and Swedish journalism and the habit of writing dashes in front of loose quotes.

One reason this model might perform worse than the one trained on data written in Bokmål, a part from the size, is that Norwegian Nynorsk has many allowed styles for writing. One can choose freely between an abundance of words and suffixes. This comes from the history of the language, as one that should fit most of rural Norway. Bokmål is fairly more standardized and probably easier to create patterns from.

Chapter 5

Conclusions

This final and last chapter of the thesis, seeks to give a comprehensive summary of the key elements of this thesis. This includes the accomplishments and achievements of the research, as well as shortcomings of the approach. This section will also discuss how this research can be expanded, improved or replicated.

This thesis has been focused on Punctuation Prediction for Norwegian, with focal point on on utilizing pre-existing approaches across domains and languages, as a possibility for under-resourced languages to get access to tools for language processing. The main focal point have been to reproduce the approach outlined by Alumäe and Tilk, and explore whether it can be trained with training data based on other languages. There have also been undertaken some experiments, which have tried to shed light over how much data is needed to make a functional model, and if there are differences in training models for the Norwegian written styles Nynorsk and Bokmål. The last experiment performed concerned the performance of punctuating text with a model which was trained on another language. Throughout this thesis, it has been made clear that approaches can be used across domains and languages, but one can expect different results for different languages, as they have different rules and vocabularies. As a result of this thesis, the field of punctuation prediction for Norwegian is no longer a blank slate, but contains some light scribbles.

This thesis has show how the bidirectional recurrent neural network with attention mechanisms for punctuation restoration proposed by Tilk and Alumäe can be trained to good results for Norwegian, even passing the level of the state of the art.

5.1 Research Hypothesis and Research Questions

This thesis started out with a research hypothesis, and some research questions which the author, throughout this thesis have been attempting to answer. The Research hypothesis presented in the start of this thesis was as follows:

> In the scope of Norwegian language, a model for punctuation prediction can be created using an approach designed for another language, trained on Norwegian data sets. This study might have a positive impact on the research community and contribute to the state of Norwegian as a modern language which partakes in the digital world, as well as give insight for other under-resourced languages on how to best use existing resources.

And Research Questions for the thesis have been:

- *RQ1.* Can an approach to punctuation restoration be used with other languages than it was designed for, and can this help under-resourced languages?

Throughout this paper, there have been multiple examples towards the possibility of approaches which work well across languages. The approach reproduced in this study [83], used the same approach for two languages; English and Estonian. In association with the ELITR project, an end-to-end ASR system, which also made use of the same approach as in this paper [60]. Lastly, the findings of this study also supports the validity of the research question. The main Norwegian model performs very well on data collected from Norwegian Newspapers. However, the Model trained on Norwegian Nynorsk, did not perform as well. This probably partly for the reason that there where not enough data to train the model satisfactorily, as well as Nynorsk being a complex written style with many allowed words and forms. This finding still support the fact that approaches can be adopted between languages, however it does not bode so well for under-resourced languages, as they one of the reasons they are called under-resourced is that they lack resources, such as data sets for training of a certain size, as well as data sets in general.

- *RQ2.* Can only textual training data produce desirable results for Norwegian? As discussed in the results for the Norwegian model, the results where very much desirable, and outperformed both the reference model as well as the state of the art on certain metrics. The overall precision for the state of the art [24], 84, saw a slight improvement by 1.5, by the overall precision from the Norwegian model, 85.5. The Norwegian model also did exceptionally well when it came to periods, an also here improved the state of the art with up to about 5 points. The model also did well on most other metrics as well, a part from question marks. However, the model did not generalize well to transcribed data. And was beat by Alumäe and Tilks model by bout 16,8 points in overall precision, 50 points in overall recall and 40,9 for overall F-score. The reason for this might be a combination of very structured training data, and validation data put together under time pressure. The results with only textual training data have shown that it produces a very accurate model, however it does not generalize well to the domain it is intended for. Training on a dataset transcribed from speech would be preferable, and so would probably prosodic features. But for punctuating regular text, the model works well. The point being made here, in favor of transcribed training data and prosodic features is also unfortunately not good news for the domain of under-resourced languages.

- *RQ3.* How will the error calculation compare between the model trained by Tilk and Alumäe [83] trained on English data and the model trained on Norwegian data?

The two models both perform well. Both of the models does certain things well, and have poorer performance when it comes to other things. The English model performs overall well on question marks for instance, and has a predicts more of the possible commas correctly, even thought it performs slightly worse on hitting correctly every time. The Norwegian model performs noticeably better when it comes to periods, as well as dashes. The Norwegian model performs better overall, but it has also had more structured data. To do a true comparison, a dataset which is closer in structure needs to be utilized.

5.2 Limitations

This study, has like most studies, some limitations. The most pronounced, from the authors' point of view, is the lack of a good amount of validation data. The data used for validation in this thesis, was too sparse, and could therefore not produce error calculations for all punctuation symbols. Transcribed data with punctuation is hard to come by, and when the dataset which was used was made available during the last days of May, the time to process it was limited.

Another limitation is the training data. While the training data showed great promise and excellent metrics on the in-domain validation, it did not generalize well to the transcribed text (probably for several reasons, but the difference in data and text was most likely prominent). It give the model an unfair edge, when comparing metrics with other models. There should preferably have been an additional model trained on ASR-output, to make comparing to other models easier as well as probably improving the metrics when punctuating real data.

5.3 Future work

This sections describes what could be done to further investigate this domain, beyond the scope of this thesis.

5.3.1 Prosodic Features

An obvious next step is to train the model with second stage data. This involves a dataset with prosodic features. The approach proposed by Alumäe and Tilk supports text annotated with pause durations, but their approach can be expanded to accept other prosodic features as well quite easily.

5.3.2 Validation Based on Out-of-Domain Data and ASR Data

This thesis has only performed validation on a relatively small dataset of transcribed data. The transcribed data is in addition not true to a common ASR-output, as the transcript is corrected by professionals. This opens for two possibilities for future work; namely validation and calculation of error metrics on a larger portion of the dataset used, *Stortingskorpuset*, and validation on a set of pure ASR data.

In the second case, an expert assessment might also be needed. This is because spontaneous speech does not have segmentation in the form of punctuation. Therefore the results needs to be interpreted by a person to assess if the punctuation symbols are correctly placed.

An expanded of evaluating the validation round, is to take a selection of testers, and give them the same text with and without the inserted punctuation. This can be done in addition to the expert assessment, where the difference in reading time and understanding between the expert-punctuated text and the machine punctuated text can be compared.

5.4 Replication

All the tools for reproducing the research described in this paper are available as open source software of publicly available datasets. All datasets used for this thesis in Norwegian can be found in *Språkbanken*(the Language Bank) at the web pages of the National Library Of Norway [1]. The code used to reproduce the results from Alumäe and Tilks results can be found on O. Tilks github-account [2]. The Europarl dataset used for the English model can be dowloaded from Kaggle [3]. Students at the University of Bergen can get resources for computing without cost at NREC[4].

[1] https://www.nb.no/sprakbanken/

[2] https://github.com/ottokart/punctuator2/

[3] https://www.kaggle.com/nltkdata/europarl

[4] https://www.nrec.no/

Bibliography

[1] Tanel Alumäe and Ottokar Tilk. *Punctuator2*. 2016. URL: https://github.com/ottokart/punctuator2.

[2] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. *Did you hear that? Adversarial Examples Against Automatic Speech Recognition*. Tech. rep. 2018.

[3] *Amazon Alexa – Learn what Alexa can do | Amazon.com*. URL: https://www.amazon.com/b?ie=UTF8&node=21576558011.

[4] Apple Inc. *Siri - Apple*. 2011. URL: https://www.apple.com/siri/.

[5] Giusepppe Attardi. *WikiExtractor*. \url{https://github.com/attardi/wikiextractor}. 2015.

[6] Lukasz Augustyniak et al. *Punctuation prediction in spontaneous conversations: Can we mitigate ASR errors with retrofitted word embeddings?* Tech. rep. 2020, pp. 4906–4910.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate”. In: (2014).

[8] Fernando Batista et al. “Bilingual Experiments on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts”. In: *IEEE Transactions on Audio, Speech and Language Processing* 20.2 (Feb. 2012), pp. 474–485. ISSN: 15587924.

[9] Fernando Batista et al. “Recovering punctuation marks for automatic speech recognition”. In: *International Speech Communication Association - 8th Annual Conference of the International Speech Communication Association, Interspeech 2007*. Vol. 3. 10. Antwerp, Belgium: Interspeech 2007, Aug. 2007, pp. 1977–1980. ISBN: 9781605603162.

[10] *Befolkning - kvartalvis - SSB*. URL: https://www.ssb.no/folkemengde.

[11] Vincent Berment. “Méthodes pour informatiser les langues et les groupes de langues « peu dotées »”. PhD thesis. J. Fourier University – Grenoble I, 2004.

[12] Laurent Besacier et al. “Automatic speech recognition for under-resourced languages: A survey”. In: *Speech Communication* 56.1 (Jan. 2014), pp. 85–100. ISSN: 01676393.

[13] David M Blei et al. *Conditional random fields: An introduction*. Tech. rep. 4-5. 2004, pp. 1–9.

[14] Ond Bojar et al. *{ELITR} Multilingual Live Subtitling: Demo and Strategy*. Tech. rep. 2021, pp. 271–277.

[15] Joachim Büschken and Greg M. Allenby. *Improving text analysis using sentence conjunctions and punctuation*. Tech. rep. 4. 2020, pp. 727–742.

[16] Joachim Büschken and Greg M. Allenby. “Improving text analysis using sentence conjunctions and punctuation”. In: *Marketing Science* 39.4 (Jan. 2020), pp. 727–742. ISSN: 1526548X.

[17] Xiaoyin Che et al. *Punctuation prediction for unsegmented transcript based on Word Vector*. Tech. rep. 2016, pp. 654–658.

[18] Qian Chen et al. “Discriminative Self-training for Punctuation Prediction”. In: (2021).

[19] Eunah Cho, Jan Niehues, and Alex Waibel. *Segmentation and punctuation prediction in speech language translation using a monolingual translation system*. Tech. rep. December. 2012, pp. 252–259.

[20] Kyunghyun Cho et al. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”. In: (2014).

[21] François Chollet. *Deep Learning with Python*. New York, NY: Manning Publications., 2017, p. 384. ISBN: 9781617294433.

[22] Kevin Clark et al. “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators”. In: (2020).

[23] Erika Corona and Filippo Eros Pani. *A review of Lean-Kanban approaches in the software development*. Tech. rep. 1. 2013, pp. 1–13.

[24] Maury Courtland, Adam Faulkner, and Gayle McElvain. "Efficient Automatic Punctuation Restoration Using Bidirectional Transformers with Robust Inference". In: Online: Association for Computational Linguistics, 2020, pp. 272–279.

[25] *CzEng | ÚFAL*. URL: https://ufal.mff.cuni.cz/czeng.

[26] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Vol. 1. Association for Computational Linguistics (ACL), Oct. 2019, pp. 4171–4186. ISBN: 9781950737130.

[27] *Dictate messages and documents on Mac - Apple Support*. URL: https://support.apple.com/en-gb/guide/mac-help/mh40584/mac.

[28] Yu Dong and Li Deng. *Signals and Communication Technology Automatic Speech Recognition A Deep Learning Approach*. Tech. rep. 2015, p. 329.

[29] Aline Dresch, Daniel Pacheco Lacerda, and José Antônio Valle Antunes. *Design science research: A method for science and technology advancement*. 2015, pp. 1–161. ISBN: 9783319073743.

[30] ELITR. *elitr.eu – European Live Translator*. 2021. URL: https://elitr.eu/.

[31] Dario Franceschini et al. *Removing {E}uropean Language Barriers with Innovative Machine Translation Technology*. Tech. rep. 2020, pp. 44–49.

[32] *GitHub - ottokart/punctuator2: A bidirectional recurrent neural network model with attention mechanism for restoring missing punctuation in unsegmented text*. URL: https://github.com/ottokart/punctuator2.

[33] Google. *Google Assistant | Your own personal Google*. 2019. URL: https://assistant.google.com/.

[34] Google. *Voice Search Insight Report*. Tech. rep. 2018.

[35] Agustín Gravano, Martin Jansche, and Michiel Bacchiani. "Restoring punctuation and capitalization in transcribed speech". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE, Apr. 2009, pp. 4741–4744. ISBN: 9781424423545.

[36] R Hevner Alan. *A Three Cycle View of Design Science Research*. Tech. rep. 2. 2007, pp. 87–92.

[37] Alan Hevner et al. "Design Science in Information Systems Research". In: *Management Information Systems Quarterly* 28 (2004), pp. 75–.

[38] Pavel Hlubík et al. "Inserting punctuation to asr output in a real-time production environment". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Ed. by Petr Sojka et al. Vol. 12284 LNAI. Brno: Springer Science and Business Media Deutschland GmbH, Sept. 2020, pp. 418–425. ISBN: 9783030583224.

[39] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 08997667.

[40] Jing Huang and Geoffrey Zweig. *Maximum entropy model for punctuation annotation from speech*. Tech. rep. Yorktown Heights, NY: IBM T.J. Watson Research Center, Sept. 2002, pp. 917–920.

[41] Emil Hvitfeldt and Julia Silge. "Tokenization". In: *Supervised Machine Learning for Text analysis in R*. CRC Press, 2021. Chap. Tokenizati. ISBN: 9780367554187.

[42] *ikke-sak - Det Norske Akademis ordbok*. URL: https://naob.no/ordbok/ikke-sak.

[43] B.-H. Juang and L.R. Rabiner. "Automatic Speech Recognition – A Brief History of the Technology Development". In: *Encyclopedia of Language & Linguistics* (2004), pp. 806–819.

[44] Chin Char Juin et al. "Punctuation prediction using a bidirectional recurrent neural network with part-of-speech tagging". In: *IEEE Region 10 Annual International Conference, Proceedings/TENCON*. Vol. 2017-Decem. Institute of Electrical and Electronics Engineers Inc., Dec. 2017, pp. 1806–1811. ISBN: 9781509011339.

[45] Veton Kepuska. “Wake-Up-Word Speech Recognition”. In: *Speech Technologies*. InTech, June 2011.

[46] Joshua Y. Kim et al. “A Comparison of Online Automatic Speech Recognition Systems and the Nonverbal Responses to Unintelligible Speech”. In: (2019), pp. 1–13.

[47] Seokhwan Kim. “Deep Recurrent Neural Networks with Layer-wise Multi-head Attentions for Punctuation Restoration”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 2019-May. Institute of Electrical and Electronics Engineers Inc., May 2019, pp. 7280–7284. ISBN: 9781479981311.

[48] Tibor Kiss and Jan Strunk. *Unsupervised multilingual sentence boundary detection*. Tech. rep. 4. 2006, pp. 485–525.

[49] Sosuke Kobayashi. *Homemade BookCorpus*. \url{https://github.com/BIGBALLON/cifar-10-cnn}. 2018.

[50] Philipp Koehn. “EuroParl: A parallel corpus for statistical machine translation”. In: 5 (2004).

[51] Steven Krauwer. “The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap”. In: *International Workshop Speech and Computer SPECOM*. Moscow, Russia, 2003, pp. 8–17.

[52] *Language support | Cloud Speech-to-Text Documentation | Google Cloud*. URL: https://cloud.google.com/speech-to-text/docs/languages.

[53] Tal Levy, Vered Silber-Varod, and Ami Moyal. “The effect of pitch, intensity and pause duration in punctuation detection”. In: *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*. 2012, pp. 1–4.

[54] Xinxing Li and Edward Lin. “A 43 language multilingual punctuation prediction neural network model”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2020-Octob (2020), pp. 1067–1071. ISSN: 19909772.

[55] Junwei Liao et al. *Improving Readability for Automatic Speech Recognition Transcription*. Tech. rep. 2020.

[56] Kristine Lindeboe. *Ikke alle forstår norske journalisters favorittegn*. Jan. 2020. URL: https://journalisten.no/kildeutvalget-lesere-oslo-met/ikke-alle-forstar-norske-journalisters-favorittegn/393414.

[57] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: (2019).

[58] Adam Lopez. "Statistical Machine Translation". In: *ACM Comput. Surv.* 40.3 (Aug. 2008). ISSN: 0360-0300.

[59] Wei Lu and Hwee Tou Ng. *Better punctuation prediction with dynamic conditional random fields*. Tech. rep. 11. 2010, pp. 177–186.

[60] Dominik Macháček et al. "ELITR Non-Native Speech Translation at IWSLT 2020". In: *IWSLT 2020*. 2020, pp. 200–208.

[61] J Makhoul et al. *Performance Measures For Information Extraction*. Tech. rep. 1999, pp. 249–252.

[62] Andreas C. Müller and Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. 1st. Vol. 53. 9. Sebastpol, CA.: O'Reilly Media, Inc, 2018. ISBN: 978-1-449-36941-5.

[63] Nasjonalbiblioteket. *Stortingskorpuset - Språkbanken*. May 2021. URL: https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-58/.

[64] Binh Nguyen et al. *Fast and Accurate Capitalization and Punctuation for Automatic Speech Recognition Using Transformer and Chunk Merging*. Tech. rep. 2019.

[65] *Norsk aviskorpus bokmål - Språkbanken*. URL: https://www.nb.no/sprakbanken/ressurskatalog/oai-clarino-uib-no-avis-plain/.

[66] *Norsk aviskorpus nynorskdelen - Språkbanken*. URL: https://www.nb.no/sprakbanken/ressurskatalog/oai-clarino-uib-no-avis-nno/.

[67] *NREC*. URL: https://www.nrec.no/#about.

[68] OpenStack. *Launch and manage instances — horizon 19.2.0.dev61 documentation*. Jan. 2018. URL: https://docs.openstack.org/horizon/latest/user/launch-instances.html.

[69] OpenStack. *Manage flavors — horizon 19.2.0.dev61 documentation*. Aug. 2018. URL: https://docs.openstack.org/horizon/latest/admin/manage-flavors.html.

[70] Vardaan Pahuja et al. *Joint learning of correlated sequence labeling tasks using bidirectional recurrent neural networks*. Tech. rep. 2017, pp. 548–552.

[71] Rashmi Popli and Naresh Chauhan. *Scrum: an Agile Framework*. Tech. rep. 1. 2011, pp. 147–149.

[72] *Punctuation*. 2020. URL: https://www.merriam-webster.com/dictionary/punctuation..

[73] Adwait Ratnaparkhi. *A Simple Introduction to Maximum Entropy Models for Natural Language Processing*. Tech. rep. May. 1997, p. 81.

[74] Sherry Ruan et al. “Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.4 (Aug. 2018), pp. 1–23. ISSN: 2474-9567.

[75] Tara N. Sainath et al. “Multichannel Signal Processing with Deep Neural Networks for Automatic Speech Recognition”. In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 25.5 (2017), pp. 965–979. ISSN: 23299290.

[76] Mike Schuster and Kuldip K. Paliwal. *Bidirectional recurrent neural networks*. Tech. rep. 11. 1997, pp. 2673–2681.

[77] James Shore and Shane Warden. *The Art of Agile Development*. First. O’Reilly, 2007. ISBN: 9780596527679.

[78] Rein Ove Sikveland et al. *Spontal-N: A corpus of interactional spoken norwegian*. Tech. rep. 2010, pp. 2986–2991.

[79] Koenraad De Smedt, Gunn Inger Lyse, and Gyri S Losnegaard. *the Norwegian Language in the Digital Age*. Ed. by Georg Rehm and Hans Uszkoreit. 1st. Berlin Heidelberg: Springer, 2012, p. 81. ISBN: 9783642313882.

[80] Speechlogger. *Speechnotes - Speech to Text Online Notepad*. 2018. URL: https://speechnotes.co/.

[81] *SpeechTexter | Type with your voice!* URL: https://www.speechtexter.com/.

[82] *Tegnsetting: bruk tegnene riktig - Korrekturavdelingen.* URL: https://www.korrekturavdelingen.no/tegnsetting-intro.htm.

[83] Ottokar Tilk and Tanel Alumäe. "Bidirectional recurrent neural network with attention mechanism for punctuation restoration". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 08-12-Sept (2016), pp. 3047–3051. ISSN: 19909772.

[84] Ottokar Tilk and Tanel Alumäe. "LSTM for punctuation restoration in speech transcripts". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2015-Janua.June (2015), pp. 683–687. ISSN: 19909772.

[85] *Transkriber intervjuer ved hjelp av maskinlæring.* URL: https://nrkbeta.no/2018/06/06/transkribering-av-intervjuer-ved-hjelp-av-maskinlaering/.

[86] *Type with your voice - Docs Editors Help.* URL: https://support.google.com/docs/answer/4492226?hl=en&ref_topic=6039805.

[87] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems.* Vol. 2017-Decem. Neural information processing systems foundation, June 2017, pp. 5999–6009.

[88] Pidong Wang. *A Text Rewriting Decoder with Application to Machine Translation.* Tech. rep. 2013.

[89] *Watson Assistant - Intelligent virtual agent | IBM.* URL: https://www.ibm.com/cloud/watson-assistant?cm_sp=Scheduler-_-CopyChng2-_-C.

[90] Fei Wu et al. "Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network". In: (2019), pp. 1–5. ISSN: 19909772.

[91] Www.britannica.com. *suprasegmental | Definition, Features, Examples, & Facts | Britannica.* URL: https://www.britannica.com/topic/suprasegmental.

[92] W. Xiong et al. “The Microsoft 2017 Conversational Speech Recognition System”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 2018-April. Institute of Electrical and Electronics Engineers Inc., Sept. 2018, pp. 5934–5938. ISBN: 9781538646588.

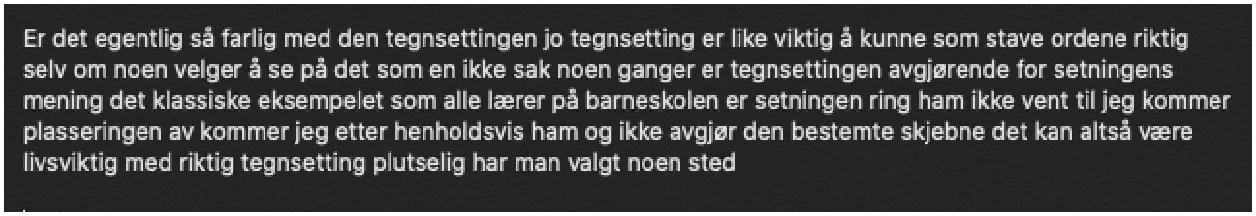
[93] Jiangyan Yi et al. *Adversarial Transfer Learning for Punctuation Restoration*. Tech. rep. 2020.

[94] Piotr Zelasko et al. “Punctuation prediction model for conversational speech”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2018-Septe (July 2018), pp. 2633–2637. ISSN: 19909772.

Appendix A

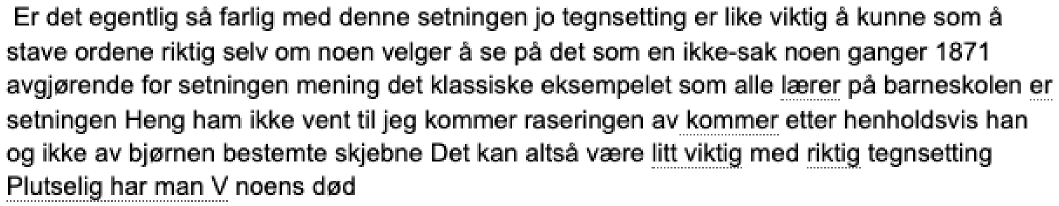
Appendix

A.1 Results from preliminary review of speech recognition softwares



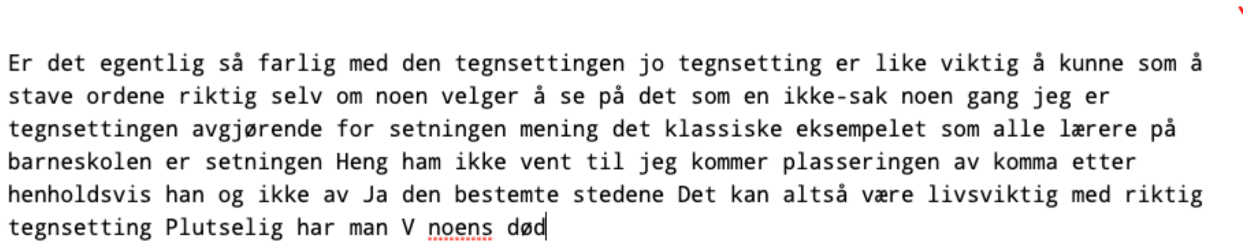
Er det egentlig så farlig med den tegnsettingen jo tegnsetting er like viktig å kunne som stave ordene riktig selv om noen velger å se på det som en ikke sak noen ganger er tegnsettingen avgjørende for setningens mening det klassiske eksempelet som alle lærer på barneskolen er setningen ring ham ikke vent til jeg kommer plasseringen av kommer jeg etter henholdsvis ham og ikke avgjør den bestemte skjebne det kan altså være livsviktig med riktig tegnsetting plutselig har man valgt noen sted

Figure A.1: Simple test of Apples' dictation service



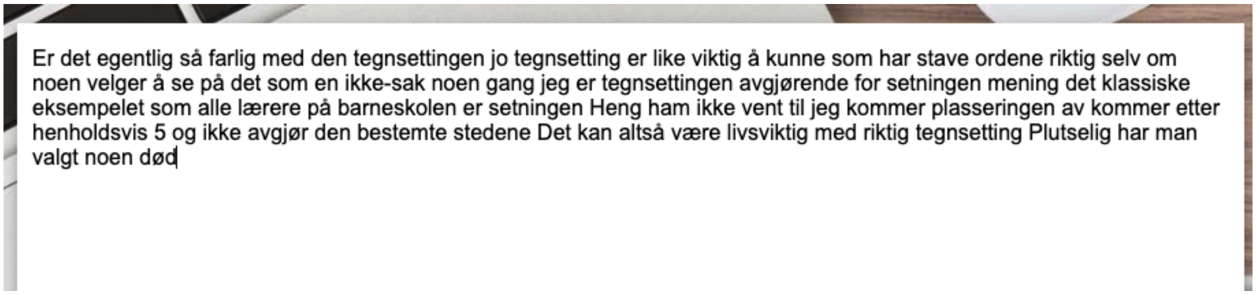
Er det egentlig så farlig med denne setningen jo tegnsetting er like viktig å kunne som å stave ordene riktig selv om noen velger å se på det som en ikke-sak noen ganger 1871 avgjørende for setningen mening det klassiske eksempelet som alle lærer på barneskolen er setningen Heng ham ikke vent til jeg kommer raseringen av kommer etter henholdsvis han og ikke av bjørnen bestemte skjebne Det kan altså være litt viktig med riktig tegnsetting Plutselig har man V noens død

Figure A.2: Simple test of Google's dictation service



Er det egentlig så farlig med den tegnsettingen jo tegnsetting er like viktig å kunne som å stave ordene riktig selv om noen velger å se på det som en ikke-sak noen gang jeg er tegnsettingen avgjørende for setningen mening det klassiske eksempelet som alle lærere på barneskolen er setningen Heng ham ikke vent til jeg kommer plasseringen av komma etter henholdsvis han og ikke av Ja den bestemte stedene Det kan altså være livsviktig med riktig tegnsetting Plutselig har man V noens død

Figure A.3: Simple test of SpeechNotes' transcription service



Er det egentlig så farlig med den tegnsettingen jo tegnsetting er like viktig å kunne som har stave ordene riktig selv om noen velger å se på det som en ikke-sak noen gang jeg er tegnsettingen avgjørende for setningen mening det klassiske eksempelet som alle lærere på barneskolen er setningen Heng ham ikke vent til jeg kommer plasseringen av kommer etter henholdsvis 5 og ikke avgjør den bestemte stedene Det kan altså være livsviktig med riktig tegnsetting Plutselig har man valgt noen død

Figure A.4: Simple test of SpeechTexter'sd transcription service

A.2 Excerpt from punctuated test data from Norsk Aviskorpus Bokmål with the Norwegian model

.PERIOD en av hans ansatte ,COMMA den norskættede leonhard seppala ,COMMA var uten tvil den eneste tenkelige hundekjører for den <NUM> <NUM> km lange rundturen fra nome til nulato og tilbake .PERIOD han hadde tidligere gjort denne turen på rekordtid fire dager og vunnet all alaska sweepstakes tre ganger ,COMMA og var blitt en slags legende for sine sportslige prestasjoner og håndteringen av sine siberian huskies .PERIOD hans lederhund « togo » var tilsvarende kjent for sine lederegenskaper .PERIOD u.s. public health service hadde et lager på <NUM> million enheter i sykehus på vestkysten som kunne sendes til seattle og deretter til alaska .PERIOD « alameda » var første båt nordover ,COMMA og hun ville ikke komme til seattle før <NUM> januar ,COMMA og så ville det ta ytterligere <NUM> – <NUM> dager frem til seward .PERIOD men den <NUM> januar fant man <NUM> <NUM> enheter med vaksine i anchorage railroad hospital .PERIOD da sjefskirurgen john beeson hørte om behovet på ordre fra guvernør bone ,COMMA ble de pakket og overlevert til konduktøren frank knigh på toget til fairbanks som ankom nenana den <NUM> januar uten noe annet å bekjempe epidemien .PERIOD med så ville disse <NUM> <NUM> enhetene holde den i sjakk inntil det store partiet kunne komme opp fra vestkysten .PERIOD mens den første forsendelsen med vaksine var på vei til nenana ,COMMA gav guvernør scott bone sin autorisasjon til en hundestafett med postkjørere og gav ordre til edward wetzler inspektøren for u.s. post office om å arrangere en stafett med de beste hundekjørere og de beste hunder som fantes i alaska .PERIOD spannene skulle reise dag og natt til de kunne overlevere pakken til seppala i nulato .PERIOD postruten fra nenana til nome krysset tundraen i alaskas innland .PERIOD de første <NUM> km av strekningen gikk langs tanana-elva til tettstedet tanana ,COMMA der tanana-elva renner inn i yukon-elva .PERIOD de neste <NUM> km gikk langs yukon fram til kaltag .PERIOD deretter gikk ruten vestover gjennom skog og over fjellpass .PERIOD de neste <NUM> km til den nådde unalakleet ute ved kysten .PERIOD fra unalakleet gikk de neste <NUM> km av postruten nordvestover rundt sewardhalvøya .PERIOD på denne delen var det svært vanskelig å søke ly for storm eller snøstorm ,COMMA spesielt på en <NUM> km strekning som krysset isen over beringsjøen .PERIOD den totale strekningen på postruten var <NUM> <NUM> km .PERIOD wetzler kontaktet tom parson ,COMMA en forretningsfører i northern commercial company som leverte post mellom fairbanks og unalakleet .PERIOD telefon og telegram fikk kjørerne til sine anviste vertshus .PERIOD postbudene hadde høy status i dette området ,COMMA og var de beste hundekjørerne .PERIOD flertallet av stafettkjørerne over innlandet var innfødte athabaskere direkte etterkommere av de opprinnelige hundekjørerne .PERIOD den første kjøreren i stafetten var « wild bill » ,COMMA shannon ,COMMA som fikk den ni kilo tunge pakken ved jernbanestasjonen i nenana klokken ni om kvelden den <NUM> januar .PERIOD selv om temperaturen da var <NUM> minusgrader celsius ,COMMA forlot shannon togstasjonen umiddelbart med kurs mot minto .PERIOD spannet hans bestod av ni uerfarne hunder som ble ledet av « blackie » .PERIOD temperaturen falt i løpet av natten ,COMMA og spannet ble også tvunget ut på den kalde elveisen fordi løypen var blitt ødelagt av hester .PERIOD selv om shannon jogget ved siden av sleden for å holde varmen ,COMMA fikk han hjerteproblemer .PERIOD han nådde minto klokken tre om morgenen ,COMMA og temperaturen var da på <NUM> minusgrader celsius .PERIOD det iskalde været hadde sørget for at deler av ansiktet hans var svart på grunn av frostskader .PERIOD etter å ha varmet opp serumet ved peisen og hvilt i fire timer ,COMMA satte shannon fra seg tre hunder før han fortsatte med de seks andre edgar kallands som var halvt athabasker hadde ankommet minto kvelden før han dro derfra til tolovana ,COMMA der han skulle overta serumet .PERIOD i svært dårlig forfatning ankom shannon og hundespannet hans vertshuset hvor kallands ventet .PERIOD klokken var elleve om formiddagen da kallands overtok serumet .PERIOD

A.3 Excerpt from punctuated test data from the *Norsk Aviskorpus Nynorsk*

skilnaden mellom desse og dei nye ungdomslaga gjekk på føremålet der dei gamle var utelukkande politiske i verksemda .PERIOD dei første ungdomslaga .PERIOD ein veit om vart skipa i <NUM> på hindsholm nørre aaby og taasinge på kvar si side av fyn .PERIOD i følgje ungdomslaga sjølv kjente ingen av dei til kvarandre før skipinga i åra etter vart det skipa fleire ungdomslag på fyn på skjælland og jylland i <NUM> gjekk dei fynske ungdomslaga saman og skipa de sammensluttede fynske ungdomsforeninger og i <NUM> vart de sydøstjyske ungdomsforeninger skipa .PERIOD det var desse to samanslutningane som gjekk saman for å skipe de danske ungdomsforeninger i <NUM> med sine omlag <NUM> ,COMMA lokallag og <NUM> medlemmar .PERIOD allereie i <NUM> var medlemstalet auka til <NUM> .PERIOD <NUM> og samskipnaden fekk sitt høgdepunkt i <NUM> med <NUM> <NUM> medlemmar fleire og fleire av dei regionale samanslutningane vart med i samskipnaden med åra sjølv om somme heile tida valte å stå utanfor under andre .PERIOD verdskrig var ddu ein av skiparane av dansk ungdomssamvirke ein samanslutning av organisasjonar og einskildpersonar som ønskte å motverke samtidas ,COMMA nazistiske og fascistiske tendensar .PERIOD føremålet med organisasjonen var å opplyse engasjere og lære opp ungdom i demokrati .PERIOD etter krigen var det fleire og fleire innan de danske ungdomsforeninger og de danske gymnastikforeninger som såg at det ideologiske grunnlaget i organisasjonane var meir eller mindre likt og at det einaste som skilde .PERIOD dei var kva aktivitetar dei dreiv på med i <NUM> gjekk dei saman og vart heitande de danske gymnastik- og ungdomsforeninger .PERIOD i <NUM> var ddu med å skipe nordisk samorganisasjon for ungdomsarbeid .PERIOD i lag med søsterorganisasjonane sine ,COMMA samt 4h-forbunda og bondelagas ungdomsforeininger i dei ulike nordiske landa fabrikkane som gjekk saman var ålgård ullvarefabrikker nydalens fabrikker i nord-trøndelag hjula veveri i oslo si ullvareavdeling .PERIOD grorud textilfabriker med skauger fabrikker ved drammen .PERIOD i tillegg til fredfoss uldvarefabrik ved vestfossen nydalens fabrikker var allereie før denne samanslutninga vorte ein del av ålgård ullvarefabrikker dfu etablerte i løpet 1970-talet .PERIOD ei landsdekkjande butikkjede .PERIOD utover 1990-talet vart det slutt på tekstilproduksjon i noreg ,COMMA boghandlernes gyldne ,COMMA laurbær eller berre de gyldne laurbær er ein dansk litterær pris .PERIOD prisen vert utdelt av den danske boghandlerklubben de haas–van .PERIOD alphen-effekt er eit fenomen som skjer hos enkelte metall .PERIOD ved låge temperaturar består i at den magnetiske susceptibiliteten varierer periodisk ved ei kontinuerleg forandring av eit ytre magnetfelt .PERIOD effekten kan forklarast som ein verknad av kvantisering av energinivåa til elektrona i magnetfeltet .PERIOD fenomenet vart først observert for vismut av dei nederlandske fysikarane wander johannes de haas og studenten hans p. m. van alphen i <NUM> de havilland .PERIOD comet vart sett inn i trafikk frå <NUM> og var verda si første jetdrevne passasjerfly i regulær trafikk .PERIOD dei tidlege modellane var utsett for metalltretthet på grunn av for store passasjervindauge og det var først med 4-serien at flyet var fullt sikkert.raf til ca .PERIOD <NUM> dvs om lag <NUM> år etter jomfruturen .PERIOD modellnamnet vert då nimrod mra <NUM> ronald bishop mannen bak mosquito-jagerbomberen vart sett på oppdraget etter ein del radikale tankar som to-kroppsfly og flygande vingar landet ein på eit meir konvensjonelt utsjånad som vart presentert som comet i desember .PERIOD <NUM> første levering var forventa <NUM> prototypen <NUM> comet flaug <NUM> minutt <NUM> juli <NUM> i tilknyting til flyshowet på farnborough .PERIOD eitt år seinare var neste prototyp på vingane .PERIOD <NUM> april <NUM> vart dette flyet overlevert boac med registrering g-alzk .PERIOD det gjennomførde <NUM> flytimer med ruteprøving og pilottrening .PERIOD straumlinjeprofilen inneheldt fleire innovative element som bakoverstrøken vingeprofil med integrerte drivstofftankar og to firehjuls .PERIOD landingsboggiar utvikla av fabrikken .PERIOD trykkabinen var òg tidleg men propellflya boeing <NUM> og <NUM> var like tidlege redningsflåtar var plssert i vingerota livbelte under alle seta to par .PERIOD turbojet de havilland ghost .PERIOD <NUM> mk.1- motorar sat i vingane .PERIOD tett inntil kroppen .PERIOD designen var vald for å unngå vanskane med ytre motoroppheng som fuglar i inntaket og asymmetrisk trykk .PERIOD ein kunne òg gjera rorflatene mindre til gjengjeld vart konstruksjonen i seg sjølv tyngre og meir kompleks metallhuden bestod av nye legeringar som var både lima og nagla på plass ,COMMA noko som hindra spreiing av sprekkdanningar frå naglehola og som spara vekt .PERIOD dei flya som vart overlate boac i mai <NUM> var dei til då mest gjennomtesta fly .PERIOD nokosinne flykroppane var senka under vatn og sett under trykk og utsett for vakuum .PERIOD <NUM> <NUM> gonger tilsvarande ca .PERIOD <NUM> <NUM> fytimar kabinane tolde det meste opp til noko over forventa .PERIOD g-alyp var det fyrste flyet i serien og kom i lufta i januar .PERIOD <NUM> .PERIOD g-alys fekk sit sertifikat for sivil passasjertrafikk <NUM> mai <NUM> som det fyrste jetflyet i verda .PERIOD g-alyp fekk æra av å ta første passasjerlast til johannesburg .PERIOD det siste i serien på ti maskiner vart levert september <NUM> og flaug først fraktgods som utprøving av framtidige ruteplanar .PERIOD både dronning elizabeth og dronningmoren var tidleg ute med dei fyrste turar med jetfly i ein spesialarrangert .PERIOD flight cometen var ca <NUM> % raskare enn propellflya og klatra raskare ,COMMA slik at reisetida på mange distansar vart halvert .PERIOD motoren var sterk og ukomplisert hadde låge vedlikehadskostnader og fekk flyene raskt over vêr som andre måtte fly rundt eller gjennom fyrste arbeidsåret vart <NUM> <NUM> ,COMMA passasjerar transportert og <NUM> nye comet vart bestilt ein boac-flyvning frå ciampino lufthamn nær roma kom seg ikkje i lufta og fleire passasjerar fekk småskader i mars .PERIOD året etter var det ein canadian pacific-maskin som fekk for lite løft og traff ei bru i karachi i pakistan med elleve dødsoffer av mannskap og passasjerar begge hendingane vart karakterisert som pilotfeil .PERIOD seinare viste det seg at vingeprofilene ikkje gav løft samstundes med at motorane på eine .PERIOD sidan fekk for lågt trykk fabrikken monterte nye venger og ein forkantspoiler ,COMMA men forheldt seg taus når det gjaldt kva det skulle tena til den første døyelege .PERIOD ulukka var med g-alyv som møtte ein tropestorm seks minutt etter avgang frå kolkata .PERIOD india comet <NUM> vart kritisert for låg « stikkefølsomheit » ,COMMA noko som kan ha bidrege til hendinga den <NUM> april <NUM> krasja g-alyy nær napoli på veg til kairo .PERIOD

A.4 Excerpt from punctuated validation data from Stortingskorpuset with the Norwegian model

hensynene vi er på jakt etter her ledelsen skal jeg ikke gi noen dom over vår teknologiske kompetanse av embetsverket .PERIOD feltet åpnes for oppfølgingsspørsmål – først til sverre myrli sikkerhetstiltak og risiko- og sårbarhetsanalyser ikke er gjennomført .PERIOD store virksomhetsoverdragelsen politiske føringer overfor helse sør-øst knyttet til valg av løsning ,COMMA utvikling og å lage et grunnlag for god behandling .PERIOD videre og eventuelt overføring av ansvar ,COMMA v. bollestad – til neste oppfølgingsspørsmål kommer på avveier .PERIOD det er det som nrk hevder at de kan dokumentere når ulike tjenester skal utføres fra utlandet ,COMMA bl.a .PERIOD kryptering .PERIOD det er norsk lov som skal gjelde toppe – til oppfølgingsspørsmål har skjedd no kan eg ikkje forstå ei fullstendig utgreiing om saka lukket går utover dette kjenseth – til oppfølgingsspørsmål om både personvern og datasikkerhet ,COMMA men det handler også om objektsikring betviler den informasjonen er vi nødt til å få til statsråden er om han vil vurdere helsedata som en del av sikkerhetsloven skal gjøres på dette området .PERIOD den lokale lovgivningen og dermed ikke falle inn under sikkerhetsloven slik den er i dag ,COMMA gjør denne type vurderinger ,COMMA men da med bakgrunn i dagens lov .PERIOD knag fylkesnes – til oppfølgingsspørsmål av noregs befolkning har hamna på avvegar ,COMMA lidingar ,COMMA sjølvmord osv .PERIOD dette har hamna på avvegar ut av landet som statsråden har stått her i salen og garantert sikkerheita for korleis kan noregs befolkning ha tillit til ministeren etter dette på før han trekker bastante konklusjoner går videre til neste hovedspørsmål står her for mitt spørsmål går til ham som er større enn vestfold fylke til tider ikke kan lande .PERIOD det er umulig med både bil- og båttransport ,COMMA <NUM> pst til traumesenteret på haukeland eller til sykehuset i haugesund i flekkefjord og i odda når statsråden allikevel griper inn ,COMMA men at en skal bevare og videreutvikle også de mindre sykehusene i tråd med det som er føringene i nasjonal helse- og sykehusplan .PERIOD det skal også utarbeides en videre plan for hvordan dette skal gjennomføres som helse vest nå skal jobbe med og som gjelder fram mot <NUM> for befolkningen lokalt til lokalt anbefale sin stortingsgruppe ,COMMA faktisk å være imot å bevare odda sykehus gjennom disse beslutningene .PERIOD i det enkelte sykehus blir oppfølgingsspørsmål – først hans fredrik grøvan rundt flekkefjord sykehus ,COMMA og at det ikke gis rom for flere tolkninger etter at vedtaket i helse sør-øst ble avvist .PERIOD nå kan skrinlegges av dagens akuttkirurgi og traumefunksjonene skal ha videre – og det er den rollen de har i dag karoline knutsen – til oppfølgingsspørsmål slik at stortinget skal slippe å sitte og detaljstyre sykehusene beredskap på gravdal sykehus .PERIOD lofoten har man gjort nettopp det når man har hatt dialog med odda ,COMMA føle trygghet for at de skal videreføres som akuttsykehus toppe – til oppfølgingsspørsmål er det mellom to og tre timar til eit anna sjukehus .PERIOD lokalsjukehus med traumefunksjon i nærleiken forståelse av hvilken kompetanse som finnes på de ulike sykehusene befolkningens trygghet .PERIOD den starter lenge før sykehuset hardanger breivik – til oppfølgingsspørsmål i <NUM> skulle ta tilbake den overordna styringa av spesialisthelsetenesta fjerna – kjempebra frå regelen nemleg odda .PERIOD med vedtaket om nedlegging av akuttkirurgien ministeren forståing for dette argumentet og kva er svaret hans et akuttsykehus rundt sin videre drift ,COMMA noe som også har gitt trygghet for rekruttering ,COMMA god rekruttering ,COMMA lysbakken – til oppfølgingsspørsmål som fra de andre i det <NUM> mai-toget det opp og sørge for at det blir satt ut i livet .PERIOD stortingsflertallet bestemmer seg for ikke å følge den i denne saken konsekvensene av hvis en bryter det prinsippet det alltid til å være går videre til neste hovedspørsmål høie nasjonal helse- og sjukehusplan forsvann omgrepet « lokalsjukehus » .PERIOD dette betyr må ha av .PERIOD akuttberedskap er nøyaktig definert i den nasjonale traumeplanen med traumekompetanse med nasjonal helse- og sjukehusplan .PERIOD dette kan sjølvsagt ikkje fortsetja .PERIOD det skal ikkje ha kirurgar i døgnvakt .PERIOD korleis er dette fagleg .PERIOD mogleg sjukehusa skal ha kirurgar i vakt – kanskje bare ett – i norge som kan ta imot alle akuttpasienter .PERIOD

A.5 Excerpt from punctuated test data from the Europarl dataset with the English Model

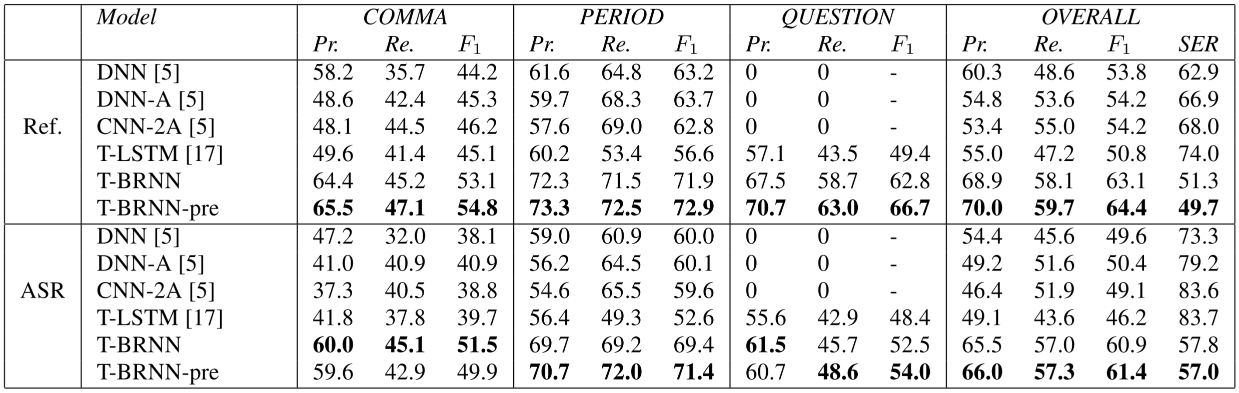
that must be our objective during <NUM> .PERIOD i hope the commission and the council will come forward with proposals on this matter .PERIOD mr president ,COMMA like so many others ,COMMA i want to congratulate the irish presidency on the success of its term of office and to say that ,COMMA because smaller countries have fewer resources ,COMMA the success which they achieve therefore deserves greater commendation .PERIOD i want to compliment mr bruton ,COMMA the taoiseach ,COMMA mr spring ,COMMA the tánaiste and mr mitchell ,COMMA all of whom worked extremely hard and contributed immensely to that success .PERIOD i want also to acknowledge today that it was not just while .PERIOD mr bruton was president-in-office ,COMMA that he was dedicated to the european ideal .PERIOD all through the years he has promoted the european ideal and been a hard-working and well-informed advocate of european union .PERIOD it is particularly important to say that ,COMMA at this time ,COMMA in his new year message speaking to his own people in his own state ,COMMA he set out as one of his three objectives to work for the creation of a federal europe .PERIOD it is easier for us in the institutions to advocate these things .PERIOD we are expected to do so in the years immediately ahead .PERIOD we have three important tasks to accomplish .PERIOD i refer to the intergovernmental conference ,COMMA the single currency and enlargement .PERIOD some of these things in particular enlargement ,COMMA will not be achieved without sacrifices by the people who are in the european union at the present .PERIOD if we are to achieve success in this area ,COMMA we cannot do it by making agreements between our own institutions ,COMMA the council ,COMMA parliament and the commission .PERIOD we must win over national public opinion .PERIOD national and regional parliaments have a very big role to play .PERIOD that is why i appreciate so much the long-standing service to the european ideal in the irish context ,COMMA which the present prime minister of ireland ,COMMA mr bruton ,COMMA has given and in which he has been well supported by mr mitchell .PERIOD i want to refer briefly to the statement by ms mckenna that european monetary union is a bad thing and that ireland should not be involved .PERIOD she based this on our trade flows with the united kingdom and the rest of europe .PERIOD her figures are out of date .PERIOD today we export <NUM> % of our goods to the hard-core countries of the european union and <NUM> % to the british market .PERIOD i would remind her that in the <NUM> years of monetary union ,COMMA with britain we made less progress than in the <NUM> years of monetary agreement with the countries of the european union .PERIOD mr president ,COMMA i hope you will forgive me ,COMMA but i see the results of the irish presidency and the dublin summit above all in terms of their value as a starting-point for the dutch presidency ,COMMA especially when it comes to employment .PERIOD without wishing to criticize the irish presidency .PERIOD in this respect ,COMMA i would nevertheless point out that there is so much fog surrounding these issues that i cannot say that things have now been made so clear that nothing can go wrong with the dutch presidency .PERIOD our aim with the creation of the chapter on employment was that a balance should at least be established between monetary policy on the one hand ,COMMA and the policy on growth ,COMMA employment and cohesion in the community .PERIOD on the other .PERIOD when we look at the results of your presidency ,COMMA i have to say in any event that such a balance is not present in the draft treaty .PERIOD according to this proposal ,COMMA employment policy has to be in keeping with the economic guidelines .PERIOD that is the only thing which is said on the matter ,COMMA but the reverse must also be the case .PERIOD that is now precisely the point at issue .PERIOD one cannot simply continue with an economic and monetary policy without ,COMMA at the same time ,COMMA taking account of the effects on employment and not only on the people who are already in work ,COMMA but on the many millions who are outside the employment process and who ,COMMA despite the financial measures taken in recent years ,COMMA have still not found a job .PERIOD i hope that this fact will be brought home clearly once again in the summit discussions .PERIOD in short ,COMMA there is a great deal more to be done here ,COMMA not least because the proposal from the dutch presidency does indeed suffer from the same fault .PERIOD the inclusion of the essen procedure in the treaty ,COMMA which we now seem to have in mind ,COMMA is totally inadequate .PERIOD

A.6 Excerpt from punctuated test data from the *Nors Aviskorpus Bokmål* dataset with the English Model

i <NUM> makten gled ,COMMA deretter gjennom forfatningen av <NUM> tilbake til kongen ,COMMA i <NUM> fikk sverige ,COMMA ny forfatning den bestemte at makten skulle deles mellom kongen og riksdagen ,COMMA som fortsatt besto av de fire ,COMMA stendene domstoler og myndigheter fikk en selvstendig stilling og ordningen ,COMMA med justisombudsmann ,COMMA og ,COMMA konstitusjonskomité ,COMMA ble innført ,COMMA læren ,COMMA om maktfordeling ,COMMA som skiller ,COMMA mellom ,COMMA lovgivende ,COMMA dømmende ,COMMA og ,COMMA utøvende ,COMMA makt hadde ,COMMA stor innflytelse på den nye forfatningen med en del viktige endringer var forfatningen av <NUM> ,COMMA i kraft ,COMMA til <NUM> riksdagsordningen ble gjennomført .PERIOD i <NUM> .PERIOD i årene <NUM> ble det gjort en rekke endringer i forfatningen for at de fremvoksende klassene skulle bli representert .PERIOD i riksdagen .PERIOD i <NUM> ble rekrutteringen til riksdagen etter stender avskaffet og erstattet ,COMMA med et tokammersystem det første kammeret ble valgt direkte gjennom landstingene og de største byenes kommunale ,COMMA forsamlinger det ,COMMA ble ,COMMA ansett ,COMMA å ,COMMA representere ,COMMA « ,COMMA bildningen ,COMMA och ,COMMA förmögenheten ,COMMA » ,COMMA hver ,COMMA representant ,COMMA var valgbar ,COMMA på ,COMMA grunnlag av alder ,COMMA inntekt ,COMMA og ,COMMA formue ,COMMA ved ,COMMA valget ,COMMA til ,COMMA andrekammeret ,COMMA på ,COMMA 1800-tallet ,COMMA var stemmeretten begrenset ,COMMA til menn og for å kunne stemme måtte man ,COMMA ha fast ,COMMA eiendom eller ,COMMA betalt skatt på en årlig inntekt valgbare ,COMMA var de som hadde ,COMMA stemmerett ,COMMA og ,COMMA var fylt ,COMMA <NUM> år hvilket innebar at kun ,COMMA <NUM> prosent av alle svenske menn over <NUM> år hadde stemmerett til andrekammeret spørsmålet ,COMMA om stemmerett ,COMMA ble debattert ,COMMA ivrig fra 1860-tallet da et krav om noe som .PERIOD i praksis betydde allmenn stemmerett ble lagt frem allmenn stemmerett for menn ved valg til andrekammeret ble innført .PERIOD i <NUM> .PERIOD i

<NUM> ble allmenn og lik stemmerett for menn innført og ,COMMA i <NUM> fikk også kvinner stemmerett først da kan en si at riksdagen fullt ut representerte hele folket etterhvert ble også stemmerettsalderen senket til valget av andrekammer .PERIOD i <NUM> var stemmerettsalderen for eksempel <NUM> år samtidig med at stemmeretten ble utvidet ble også parlamentarismen praksis at regjeringens eksistens er avhengig av riksdagens tillit ,COMMA i <NUM> avviklet en tokammersystemet og en samlet riksdag med <NUM> representanter ble innført samtidig endret en på komitesystemet ordningen ,COMMA med forskjellige komiteer for lover :COLON og budsjettspørsmål ,COMMA ble ,COMMA fjernet og ,COMMA i stedet fikk en totalt <NUM> komiteer for ulike fagområder tre år senere ,COMMA i <NUM> fikk sverige en ny forfatning parlamentarismens prinsipper ble formelt nedtegnet og talmannen ble gitt en sentral rolle ved dannelsen av regjering ,COMMA det viste seg snart at det ,COMMA var lite heldig ,COMMA med et likt antall representanter til riksdagen ,COMMA ved riksdagsvalget ,COMMA i <NUM> fikk de sosialistiske og de borgerlige ,COMMA <NUM> representanter hver seg det førte til at flere avstemninger i riksdagen måtte avgjøres ved loddtrekning ,COMMA i <NUM> bestemte riksdagen at valgperioden skulle forlenges fra ,COMMA tre til fire år og at budsjettarbeidet ,COMMA skulle effektiviseres det siste innebærer at budsjettåret ,COMMA følger kalenderåret og at budsjettproposisjonene ,COMMA legges frem og behandles ,COMMA om høsten ,COMMA axel vennersten .PERIOD i full uniform som sveriges riksmarskalk ,COMMA i <NUM> sveriges riksmarskalk er den høyeste embedsmannen ved det kongelige hoff .PERIOD i sverige riksmarskalken utnevnes av den svenske monarken riksmarskalken er ansvarlig for hoffets virksomhet for riksdagen og regjeringen riksmarskalken tituleres eksellense ved høytidelige anledninger bærer riksmarskalken hoffuniform og en lang ,COMMA stav ,COMMA med kongelig ,COMMA krone staven støtes ,COMMA i gulvet ved riksdagens høytidelige åpning for å signalisere at monarkens trontale skulle begynne riksmarskalken leder riksmarskalksembedet som omfatter en ekspedisjonssjef og et kanselli som håndterer ,COMMA juridiske og konstitusjonelle spørsmål ,COMMA i riksmarskalksembedet inngår også hoffets personalavdeling økonomiavdeling samt presse- og informasjonsavdeling under riksmarskalken sorterer dessuten läkarstaten og kleresistaten samt hoffauditøren og overhoffmesterinnen nære medarbeidere til riksmarskalken er første hoffmarskalk ,COMMA som leder

A.7 Alumäe and Tilk results from Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration[83]



	Model	*COMMA*			*PERIOD*			*QUESTION*			*OVERALL*			
		Pr.	*Re.*	F_1	*Pr.*	*Re.*	F_1	*Pr.*	*Re.*	F_1	*Pr.*	*Re.*	F_1	*SER*
Ref.	DNN [5]	58.2	35.7	44.2	61.6	64.8	63.2	0	0	-	60.3	48.6	53.8	62.9
	DNN-A [5]	48.6	42.4	45.3	59.7	68.3	63.7	0	0	-	54.8	53.6	54.2	66.9
	CNN-2A [5]	48.1	44.5	46.2	57.6	69.0	62.8	0	0	-	53.4	55.0	54.2	68.0
	T-LSTM [17]	49.6	41.4	45.1	60.2	53.4	56.6	57.1	43.5	49.4	55.0	47.2	50.8	74.0
	T-BRNN	64.4	45.2	53.1	72.3	71.5	71.9	67.5	58.7	62.8	68.9	58.1	63.1	51.3
	T-BRNN-pre	**65.5**	**47.1**	**54.8**	**73.3**	**72.5**	**72.9**	**70.7**	**63.0**	**66.7**	**70.0**	**59.7**	**64.4**	**49.7**
ASR	DNN [5]	47.2	32.0	38.1	59.0	60.9	60.0	0	0	-	54.4	45.6	49.6	73.3
	DNN-A [5]	41.0	40.9	40.9	56.2	64.5	60.1	0	0	-	49.2	51.6	50.4	79.2
	CNN-2A [5]	37.3	40.5	38.8	54.6	65.5	59.6	0	0	-	46.4	51.9	49.1	83.6
	T-LSTM [17]	41.8	37.8	39.7	56.4	49.3	52.6	55.6	42.9	48.4	49.1	43.6	46.2	83.7
	T-BRNN	**60.0**	**45.1**	**51.5**	69.7	69.2	69.4	**61.5**	45.7	52.5	65.5	57.0	60.9	57.8
	T-BRNN-pre	59.6	42.9	49.9	**70.7**	**72.0**	**71.4**	60.7	**48.6**	**54.0**	**66.0**	**57.3**	**61.4**	**57.0**

Figure A.5: Results achieved by Tilk and Alumäe[83]