An Evaluation of state-of-the-art NLP Models on Historic text in Danish

Sarah Hvid Andersen

Student no. 201910230

AU ID: AU644610

5th Semester Cognitive Science BA,

ARTS, University of Aarhus

Supervisor

Ross Deans Kristensen-McLachlan

Bachelor project in Cognitive Science

School of communication and culture, University of Aarhus

4th January 2022

0 Abstract

The capabilities of current state-of-the-art Danish NLP models have never been tested on historic Danish text. This paper therefore investigates how well the SpaCy and DaCy models perform on the named entity recognition task on historic Danish text from 1844 – 1905. The data set was annotated using the categories: person, location, organization and miscellaneous. The models were then evaluated by comparing their named entity predictions to the ground truth annotations. The results show that all models experience a drop-off in accuracy by being applied to historic text. This suggests that the orthographic and grammatical conventions of the 19th century affect the performance of the models. The result is even more revealing if only considering the oldest text from 1844 – 1871. Due to the drop-off in accuracy, which is severe for some models, it is recommended that new language models are trained on a corpus of historic text.

Keywords: Natural Language Processing · Named Entity Recognition · Historic NER · Low-resource NLP

GitHub link: https://github.com/sarah-hvid/Bachelor_ibsen

Table of Contents

1 Introduction

Natural Language Processing (NLP) is broadly defined as software-based manipulations of natural language. It is a subfield of linguistics and artificial intelligence and contains several subbranches itself. It is difficult for computers to understand natural language because language is an inherently complex phenomenon. Human language is both highly ambiguous and highly variable. A sentence may be understood in several different ways depending on the context. At the same time, humans are also quite poor at formally understanding and describing the rules that govern language (Goldberg, 2017). Additionally, there are many very different languages around the world, and they are constantly evolving. Together, these things make it challenging for computers to work with natural language, because it is impossible to account for all combinations and their meaning (Goldberg, 2017). NLP may be a powerful tool, but it is still developing and requires the correct circumstances and training. Because the best NLP models are language specific (for now), low-resource languages such as Danish generally are not as advanced as other languages. However, Danish has recently gained resources that has allowed for greater model advances. These include the Danish Gigaword corpus by Strømberg-Derczynski et al. (2021) and DaNE, a named entity resource by Hvingelby et al. (2020) that was used to train the current state-of-the-art models for Danish; DaCy and SpaCy (Enevoldsen et al., 2021; Honnibal & Montani, 2021).

Language develops over time. This means that NLP models trained on contemporary Danish might not be optimal for historic Danish. However, historic text holds relevance in much research, particularly within the digital humanities. Digitization of historic text is important in order to preserve the cultural heritage it contains and increase general accessibility (Piotrowski, 2012). In Denmark the Royal library has already digitized Danish books up to year 1600 ("*Vi Digitaliserer Samlingerne*", n.d.). They are currently working on books from 1601 – 1700. Books from 1701 – 1904 are being digitized on demand. Along with the increasing availability of historic text, there is a growing interest in applying NLP tools on historic text. Therefore, NLP models that can function at state-of-the-art levels on historic text will become more prudent, as more historic text is digitized.

With this motivation, this paper evaluated the performance of DaCy and SpaCy on a corpus of historic Danish texts. It is hypothesized that all models will experience a performance drop. However, this is nearly a given. Therefore, of greater importance is the following questions that will be investigated:

Q1: To what extent does contemporary NLP models decrease in accuracy on historic text?

Q2: Which challenges does historic text pose for NLP models?

Q3: How should future research on historic Danish text proceed?

Initially, this paper will introduce the orthographic conventions of historic Danish text from the period 1840 – 1905. It will then summarize the subtask of NLP called Named Entity Recognition (NER) and the current knowledge of historic NER. The best current Danish models, DaCy and SpaCy, will then be introduced. The specific methodology regarding the annotation procedure and annotation results will be described. Finally, the results of DaCy and SpaCy's NER predictions will be presented and evaluated. These will be discussed in the context of the orthographic conventions of the historic text. As will the limitations and future improvements of both NER and historic Danish NLP.

1.1 Named entity recognition

The task of named entity recognition and classification (NERC, generally just NER) was first presented in 1991 (Nadeau & Sekine, 2007). Research in the field accelerated in 1996 with the first event, the MUC-6 conference, dedicated to the task and the importance and relevance of NER has not decreased since. The NER task is essentially meant to extract named information from a given paragraph of text. It is a sequence labelling task, where a model given a sequence of tokens attempts to label the sequence with the appropriate classes (Ehrmann et al., 2021). This makes it useful for virtually any text mining application. Most NER models rely on either supervised machine learning, semi-supervised learning or unsupervised machine learning. However, greater advances have been achieved recently using transformers. These are created using a particular kind of deep neural network architecture. The techniques will not be explained in detail here but see Ehrmann et al. (2021) for a detailed comparison of machine-learning- and transformer based methods.

A named entity (NE) is characterized as a proper name, that is a unique identifying name. The most commonly studied and used types of NE's are person, location and organization (Nadeau & Sekine, 2007). These can be further subdivided depending on the use-case, as different domains invite different annotations. For example, monetary values or dates have often been included as NE's. A different use case entirely is within biology, where a NER model should identify proteins or drugs (Segura-Bedmar et al., 2013). The miscellaneous category was introduced in the CONLL conference in 2003 to include alternative proper names (Sang & De Meulder, 2003). In this paper, the annotation categories used are person, location, organization and miscellaneous.

1.2 Historic NER

When working with historic text, the identification of key individuals or locations are almost always of importance. NER allows for entity linking across historic databases and collections (Ehrmann et al., 2021). Overall, the tool could greatly support the exploration of historical documents. However, historical text poses a different set of challenges than contemporary text. Specifically, there is high domain heterogeneity, the input data is noisy and there is a general lack of resources (Ehrmann et al., 2021). Different texts might stem from different domains, which usually means that the language used is quite different (e.g., a poem compared to a news article). This is also a known issue for contemporarily trained NLP models. The data is also generally noisy. Words or whole passages might be missing or obscured in their raw paper format, hampering the digitization process itself. Additionally, the lack of resources is a limitation in the historic NLP field both data wise and research wise. Furthermore, the term 'historic' contains definition and delimitation issues. The problem is that the past and the present are moving targets. This makes it difficult to specify when and how a language has changed, which in turn makes it difficult to define historic text (Piotrowski, 2012). Therefore, if available data is temporally widespread, the texts might not be comparable just because they are defined as historic. Danish text from year 800 will be very different from Danish text from year 1800. The orthographic conventions are different, which is a major obstacle (Piotrowski, 2012). A clear delimitation within the historic genre should therefore generally be specified, even though this is easier said than done. Nevertheless, research on historical NER has increased over the past decade within different languages exemplifying the relevance of the field (Ehrmann et al., 2021). There has not been much research in this area for Danish, meaning that the field is currently practically unexplored.

1.3 Henrik Ibsen and Historic Danish

The historic text used in this paper is the digitized letters of Henrik Ibsen. Henrik Ibsen is considered one of the most important writers of the Modern Breakthrough. He was born in Norway but wrote both letters and poems in Danish, that were published by the Danish publisher Gyldendal (Wiingaard, 2012). From 1537 – 1814 Denmark and Norway had a commonwealth, and were collectively known as

Denmark-Norway (Vikør, 2021). During this period, it was common practice to write in Danish. In 1885 the language New Norwegian was made a nationally recognized language besides Danish-Norwegian. Therefore, Norway had two official written languages until the reform of 1907. Before this time, written Norwegian was practically identical to Danish (Vikør, 2021). The letters of Ibsen currently represent one of the best quality 19th century Danish text of which we have a sizeable quantity. Additionally, it is retained in its original orthography, which is not always the case with other digitized text from this period. The data was created by the Virtual Ibsen Center and is publicly available (University of Oslo, 2017). For these reasons, it was the best suited source for this task. The letters written by Ibsen are from the period 1844 – 1905. These were divided into four time periods: 1844 – 1871, 1871 – 1879, 1880 – 1889, 1890 – 1905.

The Danish language period ranging from 1700 to present time is called yngre nydansk (younger new Danish) (Hjorth, 2018). However, there were changes in the orthographic conventions at different times during this period. Additionally, Ibsen did not stop writing in one way or the other by the breaks the letters are divided into. The historic orthography occurs most frequently in the oldest letters from 1844 – 1871. The occurrences then gradually decrease to the most recent letters from 1890 – 1905. The most prominent differences will be described here but see appendix A for an overview of all differences found. Note that all the orthographic conventions mentioned are the ones that were found upon inspection of Ibsen's letters. Thus, it is not an exhaustive description of the changes Danish underwent in the 18th – 19th century (see Hjorth (2018) for this).

Officially, nouns and adjectives used as nouns were capitalized until 1948. However, many opposed the practice during the late 1800 and early 1900. It appears that Ibsen was one of these, as he mostly stops capitalizing nouns from 1871. One could imagine that this practice may be difficult for the models that rely on capitalization when identifying NE's (i.e., DaCy Medium).

He also often used a colon instead of a period sign, particularly in the earliest letters. This was always in relation to abbreviations, such as hr. and kgl. that instead was written as Hr: and kgl:. It was not possible to find a source suggesting whether or when this was common practice. However, it occurs often enough that it is highly unlikely to be an error. This may also cause the models problems, as colon today is used to separate sentences, and not to signify an abbreviated word. Another similar convention was altered rules for conjugation of words. Generally, altered use of word and sentence separation markers is a critical issue for NLP, because it affects the tokenization process. This is typically the first step, which means that the results influences all further processing steps (Piotrowski, 2012).

Double vowels were also used to signify vowel length. This was predominantly the case for ‘e’, ‘i’ and sometimes ‘u’. Sometimes vowel length was marked by a supporting ‘e’ attached to the vowel, or a silent ‘e’ at the end of a word. From 1855 double vowels and the silent ‘e’ should instead be signified by an accent sign. All these spelling variations were found in Ibsen’s letters, particularly the oldest from 1844 – 1871. Spelling variations cause NLP models problems because the model’s understanding of the sentence is hampered. If the alterations affect key words of a sentence or of the NE itself, the models accuracy decreases (Ehrmann et al., 2021). Summarized, the language of historic Danish used in Ibsen’s letters differs from modern Danish in several ways. These include altered spelling, capitalization, symbol use, historic words, and sentence structure.

1.4 SpaCy and DaCy

SpaCy is an open-source software library that contains NLP models for many different languages. It is developed and maintained by the company Explosion (Montani et al., 2021). It is a user-friendly and open-source framework that is very popular. SpaCy contains four Danish models: SpaCy large;- SpaCy medium;- SpaCy small;- and SpaCy trf, based on a Danish BERT. The SpaCy trf model is the only model that is transformer based. This model achieves the best performance with an F1-score of 0.83 (Honnibal & Montani, 2021). The three other models are CPU-optimized pipelines, which are generally less accurate but faster to run. They are based on machine learning methods. However, the finer details of their models are difficult to tease out, as they are created by a private company and not dedicated researchers. For the available details see Montani et al. (2021). All their Danish models are created and evaluated using the UD Danish DDT v2.5 and the DaNE dataset (Johannsen et al., 2015; Hvingelby et al., 2020).

DaCy is an end-to-end framework for Danish NLP analysis, and is built on SpaCy v.3 (Enevoldsen et al., 2021). DaCy includes NER, part-of-speech tagging and dependency parsing. It contains three fine-tuned models: DaCy small, based on a Danish Electra;- DaCy medium;- based on the Danish BERT;- and DaCy large, based on the multilingual XLM-Roberta. All DaCy models are transformer based. The models were both fine-tuned and evaluated on the DaNE dataset by Hvingelby et al. (2020). Currently, DaCy large achieves the best performance on Danish text with an F1-score of 0.86.

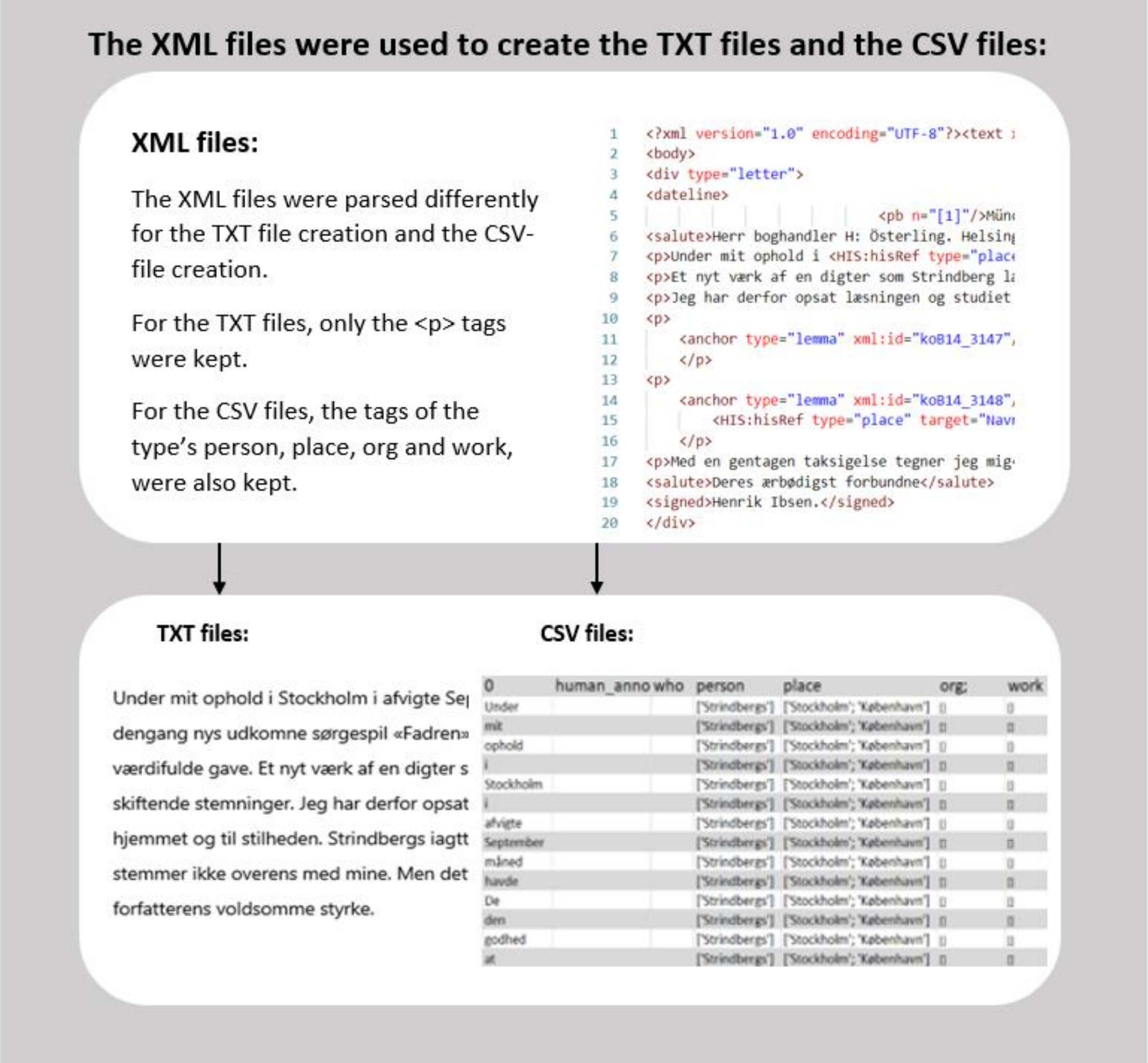
2 Materials and Methods

The following section presents the raw data and all preprocessing steps applied. It then gives an overview of the annotation process, including examples of the annotation categories. The annotation results will then be illustrated along with metadata of the letters. The next part will describe the metrics used to assess the performance of the seven models.
The code used to conduct preprocessing, wrangling and model fitting was written in Python v. 3 (van Rossum & L. Drake Jr., 2009). Additionally, Visual Studio Code was used to inspect and manipulate the files (Microsoft, 2021). The metadata of the software and package versions used are available in appendix B.

2.1 Data

The data used to evaluate the performance of Danish NLP models was the letters written by Henrik Ibsen from 1844 – 1905 in XML format. There were 2449 letters. Each letter included tags created by the Ibsen Center, intended to easily link the reader to additional information. These tags were an identification of persons, places, organizations and works. Initially, Beautifulsoup4 was used to parse the letters (Richardson, 2012). Two different files were created from each XML file. Firstly, a text file including only the body of the letter, thus, excluding the address and signature lines of each. Secondly, a CSV file with each word in its own row, that also contained the XML tags. An empty column was also created, to mark whether the human annotation was backed by the Ibsen Center. See Figure 1 for an example of the files.

The CSV files were randomly sampled by year to get 30 letters from each period. The ones not written in Danish were manually replaced. Additionally, the time-period from 1890 – 1905 contained markedly shorter letters than the other time periods, resulting in too few annotations compared to the other time periods. Larger letters were therefore chosen to replace some of the smaller letters to have approximately the same number of annotations across periods.



The XML files were used to create the TXT files and the CSV files:

XML files:

The XML files were parsed differently for the TXT file creation and the CSV-file creation.

For the TXT files, only the <p> tags were kept.

For the CSV files, the tags of the type's person, place, org and work, were also kept.

```
<?xml version="1.0" encoding="UTF-8"?><text :
<body>
<div type="letter">
<dateline>
                        <pb n="[1]"/>Münc
<salute>Herr boghandler H: Österling. Helsin
<p>Under mit ophold i <HIS:hisRef type="plac
<p>Et nyt værk af en digter som Strindberg l
<p>Jeg har derfor opsat læsningen og studiet
<p>
    <anchor type="lemma" xml:id="koB14_3147",
    </p>
<p>
    <anchor type="lemma" xml:id="koB14_3148",
        <HIS:hisRef type="place" target="Nav
    </p>
<p>Med en gentagen taksigelse tegner jeg mig
<salute>Deres ærbødigst forbundne</salute>
<signed>Henrik Ibsen.</signed>
</div>
```

TXT files:

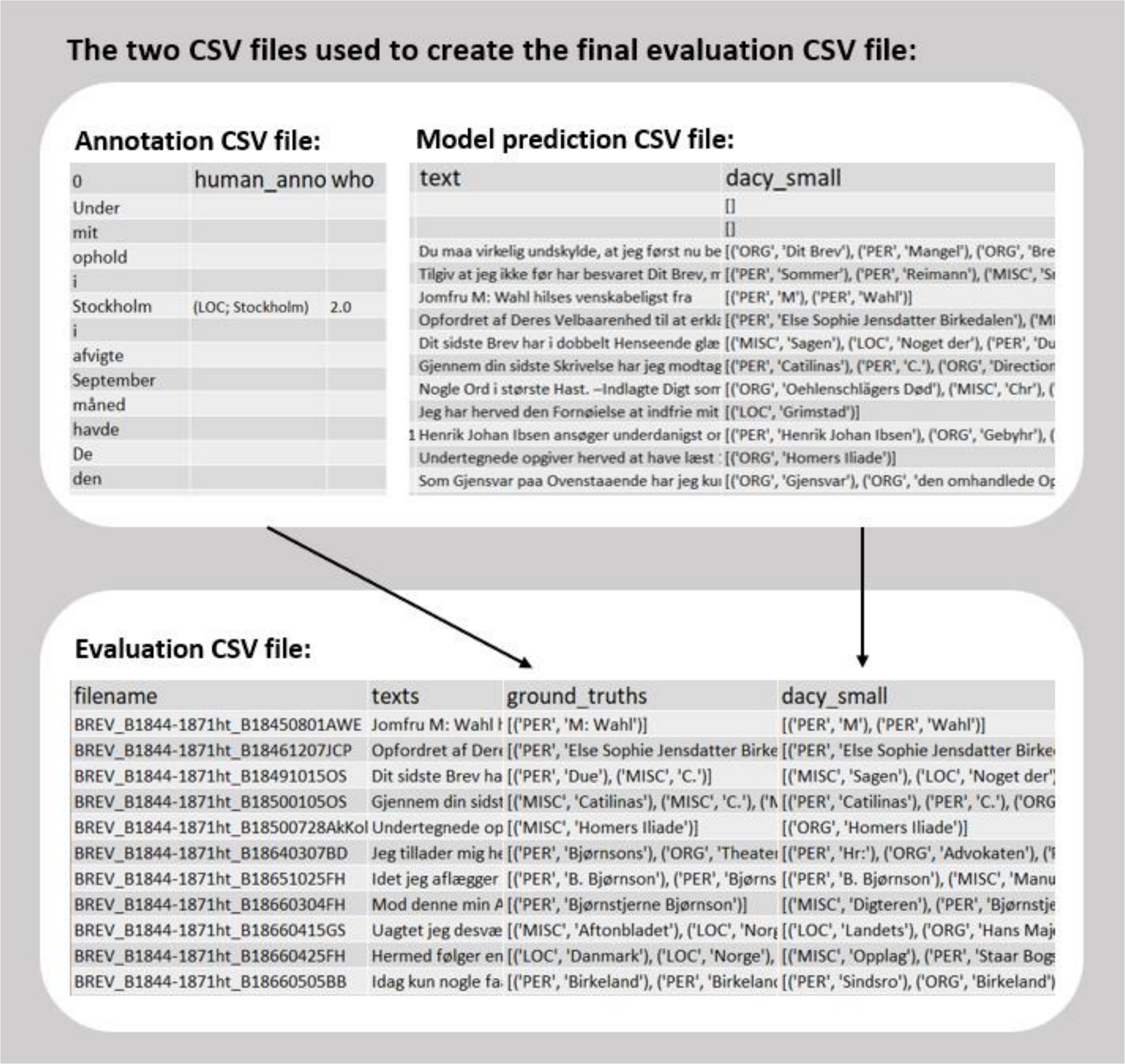
Under mit ophold i Stockholm i afvigte Se
dengang nys udkomne sørgespil «Fadren»
værdifulde gave. Et nyt værk af en digter s
skiftende stemninger. Jeg har derfor opsat
hjemmet og til stilheden. Strindbergs iagtt
stemmer ikke overens med mine. Men det
forfatterens voldsomme styrke.

CSV files:

0	human_anno	who	person	place	org:	work
Under			['Strindbergs']	['Stockholm'; 'København']	[]	[]
mit			['Strindbergs']	['Stockholm'; 'København']	[]	[]
ophold			['Strindbergs']	['Stockholm'; 'København']	[]	[]
i			['Strindbergs']	['Stockholm'; 'København']	[]	[]
Stockholm			['Strindbergs']	['Stockholm'; 'København']	[]	[]
i			['Strindbergs']	['Stockholm'; 'København']	[]	[]
afvigte			['Strindbergs']	['Stockholm'; 'København']	[]	[]
September			['Strindbergs']	['Stockholm'; 'København']	[]	[]
måned			['Strindbergs']	['Stockholm'; 'København']	[]	[]
havde			['Strindbergs']	['Stockholm'; 'København']	[]	[]
De			['Strindbergs']	['Stockholm'; 'København']	[]	[]
den			['Strindbergs']	['Stockholm'; 'København']	[]	[]
godhed			['Strindbergs']	['Stockholm'; 'København']	[]	[]
at			['Strindbergs']	['Stockholm'; 'København']	[]	[]

Figure 1 - An overview of the different initial file formats. The raw XML files (upper right) were used to create TXT files (lower left) and CSV files for annotation (lower right).

A second set of CSV files were created from the TXT files for the evaluation of the models' NER predictions. These were gathered into a CSV file per model. Thus, two sets of CSV files (i.e., annotation files and NER model prediction files) were created. Finally, the evaluation CSV file was created from both of the aforementioned CSV files. The evaluation CSV file included the ground truth from the annotation files and the model's predictions for each letter. See figure 2 for an illustration of these files.



The two CSV files used to create the final evaluation CSV file:

Annotation CSV file:

0	human_anno	who
Under		
mit		
ophold		
i		
Stockholm	(LOC; Stockholm)	2.0
i		
afvigte		
September		
måned		
havde		
De		
den		

Model prediction CSV file:

	text	dacy_small
		[]
		[]
	Du maa virkelig undskylde, at jeg først nu be	[('ORG', 'Dit Brev'), ('PER', 'Mangel'), ('ORG', 'Bre
	Tilgiv at jeg ikke før har besvaret Dit Brev, m	[('PER', 'Sommer'), ('PER', 'Reimann'), ('MISC', 'S
	Jomfru M: Wahl hilses venskabeligst fra	[('PER', 'M'), ('PER', 'Wahl')]
	Opfordret af Deres Velbaarenhed til at erkl	[('PER', 'Else Sophie Jensdatter Birkedalen'), ('M
	Dit sidste Brev har i dobbelt Henseende glæ	[('MISC', 'Sagen'), ('LOC', 'Noget der'), ('PER', 'Du
	Gjennem din sidste Skrivelse har jeg modtag	[('PER', 'Catilinas'), ('PER', 'C.'), ('ORG', 'Direction
	Nogle Ord i største Hast. –Indlagte Digt som	[('ORG', 'Oehlenschlägers Død'), ('MISC', 'Chr'), (
	Jeg har herved den Fornøielse at indfrie mit	[('LOC', 'Grimstad')]
1	Henrik Johan Ibsen ansøger underdanigst or	[('PER', 'Henrik Johan Ibsen'), ('ORG', 'Gebyhr'), (
	Undertegnede opgiver herved at have læst :	[('ORG', 'Homers Iliade')]
	Som Gjensvar paa Ovenstaaende har jeg ku	[('ORG', 'Gjensvar'), ('ORG', 'den omhandlede Op

Evaluation CSV file:

filename	texts	ground_truths	dacy_small
BREV_B1844-1871ht_B18450801AWE	Jomfru M: Wahl h	[('PER', 'M: Wahl')]	[('PER', 'M'), ('PER', 'Wahl')]
BREV_B1844-1871ht_B18461207JCP	Opfordret af Der	[('PER', 'Else Sophie Jensdatter Birke	[('PER', 'Else Sophie Jensdatter Birke
BREV_B1844-1871ht_B18491015OS	Dit sidste Brev ha	[('PER', 'Due'), ('MISC', 'C.')]	[('MISC', 'Sagen'), ('LOC', 'Noget der')
BREV_B1844-1871ht_B18500105OS	Gjennem din sidst	[('MISC', 'Catilinas'), ('MISC', 'C.'), ('M	[('PER', 'Catilinas'), ('PER', 'C.'), ('ORG
BREV_B1844-1871ht_B18500728AkKol	Undertegnede op	[('MISC', 'Homers Iliade')]	[('ORG', 'Homers Iliade')]
BREV_B1844-1871ht_B18640307BD	Jeg tillader mig he	[('PER', 'Bjørnsons'), ('ORG', 'Theater	[('PER', 'Hr:'), ('ORG', 'Advokaten'), ('
BREV_B1844-1871ht_B18651025FH	Idet jeg aflægger	[('PER', 'B. Bjørnson'), ('PER', 'Bjørns	[('PER', 'B. Bjørnson'), ('MISC', 'Manu
BREV_B1844-1871ht_B18660304FH	Mod denne min A	[('PER', 'Bjørnstjerne Bjørnson')]	[('MISC', 'Digteren'), ('PER', 'Bjørnstje
BREV_B1844-1871ht_B18660415GS	Uagtet jeg desvæ	[('MISC', 'Aftonbladet'), ('LOC', 'Nor	[('LOC', 'Landets'), ('ORG', 'Hans Maj
BREV_B1844-1871ht_B18660425FH	Hermed følger en	[('LOC', 'Danmark'), ('LOC', 'Norge'),	[('MISC', 'Opplag'), ('PER', 'Staar Bog
BREV_B1844-1871ht_B18660505BB	Idag kun nogle fa	[('PER', 'Birkeland'), ('PER', 'Birkelan	[('PER', 'Sindsro'), ('ORG', 'Birkeland')

Figure 2 - An overview of the creation of the final evaluation CSV file. The annotation CSV files (upper left) and each model's NER predictions (upper right) were used to create an evaluation CSV file for each model containing the ground truth and the model's predictions.

2.2 Annotations

The annotation guidelines predominantly referred to during the annotation of the Ibsen letters, was the CoNLL-2003 scheme introduced by Sang & De Meulder (2003). This scheme has the advantage that it is meant to be multilingual. It was also used for the creation of DaNE (Hvingelby et al., 2020). Additionally, it includes the four classes of NE's, location (LOC), person (PER), organization (ORG),
and miscellaneous (MISC) that are relevant in this context. The MISC category is a special case, as it is

not always included in the evaluation. However, in this specific use case, the MISC category is highly relevant, because it ought to catch the named mentions of Ibsen's plays and poems. Below, the guidelines for the four classes are summarized:

- PER – includes unique names of people or fictional characters. May also include abbreviations or aliases.
- LOC – includes locations like countries, cities, roads, mountains, or specific buildings. May also be an abstract or fictional location.
- ORG – includes all sorts of organizations and collections of people, like companies, brands, political movements, and governmental bodies.
- MISC – includes everything else that is a NE, like events, literary works, languages and religions. It also includes words that are derived from the other categories, including MISC.

The following section will provide examples of all categories from Ibsen's letters. Additionally, entities that were difficult to either observe or to annotate correctly will be described.

2.2.1 Person

The person category included some interesting issues. Ibsen often used titles of different varieties in front of unique names. These were capitalized, even though they are not considered part of the NE.

Postmester [PER [Reimann]]

Jomfru [PER [M: Wahl]]

Hr: Advokaten – Not a named entity

He also often used colons instead of periods in abbreviations. These are included in the annotations, as the meaning is still clear. Abbreviations of names are always part of the NE, while abbreviations of titles are not.

[PER [J: J:]]
Hr: [PER [Bjørnsons]]

He also referred to people by their surnames a great deal, which is always considered an NE. Additionally, fictional people are included when referred to by name. These were found when Ibsen mentioned characters from his plays.

> ”[PER [Birkeland]] har tilskrevet mig to breve; …”. (Birkeland has written me two letters; …)
>
> “… kan overtage [PER [Noras]] rolle …”. (… can take over Nora’s roll …)

2.2.2 Location

The identification of entities in the location category relied somewhat on previous knowledge of various countries and cities in Scandinavia as many were spelled differently.

> [LOC [Sverig]]
>
> [LOC [Kjøbenhavn]]

It was also important to research suspected LOC entities as Ibsen referred to locations in different countries. Because these were foreign, only additional research allowed for correct classification.

> [LOC [Castelamare]]
>
> [LOC [Albergo di Luna]]
>
> [LOC [Bracchio nuovo]]

A particular issue for some of the locations was that they might also function as organizations. In this scenario, the context of the sentence determined whether the entity was annotated as LOC or ORG.

> ”… skulde være antaget til opførelse på [LOC [det kgl: theater]]”. (… should be appropriated for performance at the Royal Theater.)

For this specific entity, it was also difficult to determine whether the article 'det' (the) should be kept as part of the entity name. It was decided to keep the article, because it appeared that the entity was generally written with the article capitalized contemporarily.

2.2.3 Organization

There were very few organizations in the data compared to the other categories. The identification of these also relied on the context of the sentence.

> ”Du maa endelig indlevere «Et rigt Parti» til [ORG [det kgl: Theater]] i Stockholm.”
> (You must absolutely submit «A rich Party» to the Royal Theater in Stockholm).

Generally, the names of organizations often suggested that they should be classified as ORG. The names might include society, fund, or union.

> [ORG [Kristiania Videnskabsselskab]]
>
> [ORG [Benneches Fond]]
>
> [ORG [Studenterforeningen]]

2.2.4 Miscellaneous

The MISC category was often easy to identify because special characters were present around literary works (e.g., «Et dukkehjem»). It is assumed that Ibsen wrote these himself, as original orthography is preserved. Nevertheless, these are not part of the NE and therefore not included in the annotation.

> «[MISC [Et dukkehjem]]»
>
> [MISC [Kritiker og portraiter]]
>
> [MISC [Sophoklesstatuen]]

Besides this type, the miscellaneous category covers many different types of entities that are derivatives from the other categories.

[MISC [norske]]

[MISC [Framexpeditionen]]

[MISC [Olafs-ordenen]]

These were often researched when suspected of being NE's, as the capitalization was not a trustworthy hint due to the historic orthography.

2.3 Annotation Procedure

The letters were annotated one by one in their annotation CSV file. When a NE was found in the letter, it was annotated in the adjacent column of the CSV file. The annotation was compared to the tag of the Ibsen Center. If these were in agreement the process continued and the annotator was marked as the Ibsen Center (i.e., 2.0). If these were not in agreement, the source of the error was investigated. The most frequent cause was that the annotation was of the miscellaneous category but was not a literary work (E.g., danske, Diskuskasteren (Danish, the Discobolus)). In these cases, the annotator was marked as myself (i.e., 1.0). In all cases where my annotation was not present in the tags, the entity was also researched on the internet. Often, the NE lacking as a tag was a fictional person which the search revealed. Reversely, when tags were not found in the letter, the raw XML file were compared to the annotation CSV file. This revealed that some letters were signed by multiple people, which was not part of the body of the letter and therefore removed in the preprocessing steps. Furthermore, in the Ibsen Center's tags of the letters from 1890 – 1905 only the person and organization categories were present. Therefore, most annotations in the letters from 1890 – 1905 are marked as mine.

In some cases, the correct annotation was not immediately clear. For example, the doubt would be whether an entity was a NE, what the type of the NE was or where the delimitation of the NE should be. Here, different guidelines than the main CoNLL-2003 were used to look-up definitions and examples. For instance, the entity of an address (e.g., Pilestrædet no 7) could be annotated as a location with and without the house number. However, according to Plank (2021), only the street name itself should be annotated. Other guidelines referred to in doubt were the "ACE (Automatic Content Extraction)" (2008) and the OntoNotes 5.0 by Weischedel et al. (2013). Other types of conflicts were investigated on the internet which often proved helpful. For example, unbeknownst to myself, Ibsen wrote a play called De Unges Forbund (The League of Youth). The context of the sentence implied that it was a work. This was verified by a simple internet search of the term, so the entity was annotated as MISC, and not ORG as the name otherwise implied. This example also exemplifies the need for domain specific knowledge when creating annotations. In this data, the tags created by the Ibsen Center caught many of the annotations that would otherwise have been missed.

2.4 Annotation Results

In total, 1253 human annotations were created in the 120 files. This number may seem arbitrary, but it is comparable to other studies that have based their results on a similar number of annotations (e.g. Plank (2020); 1079 tags, Hvingelby et al. (2020); 1471 tags). In total, 739 (59%) of the annotations were in agreement with the tags made by the Ibsen Center. Therefore, only 514 of the annotations were created by myself. Given the scope of this paper and as a proxy for annotator control, this score is acceptable.

	Distribution of the letters		Distribution of the named entities					
Data	Words	Sentences	Misc. +	Misc. -	Person	Location	Org.	Misc.
All data	34.696	1.831	1.253	918	460	387	78	331
1844 – 1871	9.919	510	373	281	150	118	19	89
1871 – 1879	9.483	446	318	214	84	117	13	104
1880 – 1889	7.382	422	316	208	93	98	18	107
1890 – 1905	7.912	453	246	215	133	54	28	31

Table 1 - Metadata of the letters and annotations. Note that the number of sentences was calculated through number of periods. The true number of sentences is therefore a bit lower than reported here. Org. = Organization, Misc. = Miscellaneous.

The metadata of all the letters and the distribution of the annotations are illustrated in table 1. There were 34.696 words across the letters. The NE’s consisted of 1596 words. Therefore, approximately 4.5% of the words across all letters were part of a NE. Most notably, there were quite few organizations in the data. Otherwise, the numbers found of the other entity types were relatively close to each other. The person type is also overrepresented in the letters from 1890 – 1905 while the location and miscellaneous types are underrepresented.

2.5 Evaluation

The results of all models are evaluated using the SemEval 2013 metrics (Segura-Bedmar et al., 2013). The metrics are based on the definitions from the MUC-5 conference (Chinchor & Sundheim, 1993). The metrics employ the standard measures of precision, recall and F1-score. Precision is the percentage of predicted answers that were correct. This means that precision measures how many of the predictions made were actually correct. It is defined as follows:

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)}$$

Thus, a model that does not produce any false positives has a precision of 1.0. Recall is the percentage of predicted answers that were actually correct. Recall measures the sensitivity of a model's predictions, meaning what proportion of the human annotations were actually identified correctly. It is defined as follows:

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

Thus, a model that does not produce any false negatives has a recall of 1.0. A model has a high precision score if it does well relative to the number of possible hits. It has a high recall if it does well relative to the number of actual predictions. These scores are connected in such a way that usually, if recall goes up, precision goes down and vice versa. For example, a model that predicts that every other word is a NE, would have a very high recall as it finds most of the human annotations, but a very low precision because many of the predictions are wrong.

The F1-score is the balanced mean of the precision and the recall. It comes in two formats, the micro-F1 and the macro-F1. The difference between the two, is that the macro-average will compute the metrics for each class and then take the average, whereas a micro-average will aggregate the contributions of all classes to compute the average metric. This means that the macro-average will treat all classes equally, whereas a micro-average will calculate the importance/prevalence of each class. Therefore, the micro-average is preferable if there is a class imbalance in the data (Swalin, 2018). This is often the case in NER, because person and location entities occur more frequently than organizations in natural language. Thus, the micro-F1 is predominantly used in NER evaluation, and was used by both DaCy and spaCy. For these reasons, it is also what will be used in this paper. The micro-F1 is calculated in the following way, and will hence be referred to as the F1-score:

$$Micro\ F1 = \frac{(\beta^2 + 1.0) * Precision * Recall}{(\beta^2 * Precision) + Recall}$$

If equal importance is attributed precision and recall, the beta value is 1. The SemEval13 introduced four ways of calculating the F1-score based on the flexibility allowed for a true positive (Segura-Bedmar et al., 2013). These four modes are:

- Strict match. The boundary and type of the entity must be correct.
- Exact match. The boundary of the entity must be correct, regardless of type.
- Partial match. The boundary of the entity must be partially overlapping, regardless of type.
- Type match. The type must be correct, and there must be partial overlap.

For each mode the F1-score is calculated across both axes (type and span) to evaluate each separately. The advantage of using the four modes is that they present a more nuanced picture of the various model's strengths and weaknesses. The implementation of this evaluation was conducted with the python package Eval4Ner (Chai, 2019).

3 Results

The following section firstly presents the overall results of the models. Secondly, it presents the results by year. Finally, it presents a direct comparison between the results obtained in this paper on historic Danish and the previously obtained results on contemporary Danish. The F1-values are rounded to two digits. If the rounded value is equal between models, the unrounded value takes precedence. The full F!-values can be found in the linked GitHub repository.

	Misc. −				Misc. +			
Model	Strict	Exact	Partial	Type	Strict	Exact	Partial	type
DaCy large	**0.80**	**0.83**	**0.86**	**0.85**	**0.76**	**0.82**	**0.86**	**0.83**
DaCy medium	0.69	0.73	0.78	0.79	0.68	0.74	0.79	0.76
DaCy small	0.67	0.73	0.76	0.73	0.66	0.75	0.79	0.71
SpaCy trf	0.74	0.78	0.83	0.84	0.72	0.76	0.82	0.81
SpaCy large	0.64	0.70	0.76	0.74	0.55	0.64	0.71	0.62
SpaCy medium	0.58	0.61	0.66	0.66	0.55	0.65	0.70	0.61
SpaCy small	0.47	0.57	0.60	0.50	0.45	0.58	0.62	0.48

	Type			
	Person	Location	Org.	Misc.
DaCy large	**0.82**	0.87	0.62	**0.73**
DaCy medium	0.71	0.86	0.59	0.65
DaCy small	0.69	0.80	0.55	0.63
SpaCy trf	0.71	**0.89**	**0.63**	0.68
SpaCy large	0.56	0.83	0.43	0.53
SpaCy medium	0.55	0.80	0.50	0.51
SpaCy small	0.44	0.66	0.40	0.54

Table 2 - NER performance on all data. The best scores are marked ***bold*** *and second best are underlined. The F1-values are rounded to 2 digits. In case of equal values, the unrounded value takes precedence. The type F1-scores are calculated by the 'strict' mode's criteria.*

Table 2 shows the performance of all models with and without the MISC category. The F1-values range between 0.45 and 0.86 which is a very wide. The models that are employing transformers (i.e., all DaCy models and SpaCy trf) are outperforming the ones that are not (i.e., SpaCy small, medium and large). Across all evaluation modes, DaCy large obtains the highest scores. However, it is closely followed by the SpaCy transformer model. Generally, the F1-values of all models decrease due to the inclusion of the MISC category. The only notable difference is that SpaCy large's performance decreased the most, dropping 0.11 points in the strict mode. This means that it performs similarly to SpaCy medium. Considering the F1-values by type, SpaCy trf only just achieves the highest score on the location and organization categories. From these it also appears that the non-transformer SpaCy models fall behind on all categories except for locations.

	1844 – 1871				1871 – 1879			
Model	Strict	Exact	Partial	type	Strict	Exact	Partial	Type
DaCy large	0.67	0.77	0.82	0.76	**0.80**	**0.83**	**0.86**	**0.83**
DaCy medium	**0.70**	0.76	0.80	0.77	0.71	0.75	0.79	0.76
DaCy small	0.48	0.61	0.67	0.58	0.69	0.76	0.79	0.72
SpaCy trf	0.69	**0.77**	**0.82**	**0.78**	0.78	0.82	0.85	0.81
SpaCy large	0.42	0.53	0.61	0.51	0.60	0.70	0.74	0.64
SpaCy medium	0.40	0.50	0.58	0.49	0.63	0.72	0.75	0.66
SpaCy small	0.25	0.37	0.42	0.31	0.45	0.63	0.66	0.46
	1880 – 1889				1890 – 1905			
Model	Strict	Exact	Partial	Type	Strict	Exact	Partial	Type
DaCy large	**0.84**	**0.90**	**0.93**	**0.88**	**0.76**	**0.77**	**0.82**	0.84
DaCy medium	0.68	0.75	0.80	0.77	0.61	0.70	0.78	0.75
DaCy small	0.76	0.86	0.89	0.79	0.70	0.75	0.80	0.75
SpaCy trf	0.69	0.72	0.79	0.78	0.71	0.75	0.82	**0.85**
SpaCy large	0.60	0.69	0.75	0.67	0.58	0.65	0.74	0.68
SpaCy medium	0.60	0.69	0.74	0.66	0.58	0.67	0.73	0.65
SpaCy small	0.61	0.73	0.75	0.64	0.47	0.61	0.64	0.52

Table 3 - NER performance across the four time periods. The best scores are marked ***bold*** *and second best are underlined. The F1-values are rounded to 2 digits. In case of equal values, the unrounded value takes precedence. The miscellaneous type is included in these scores.*

Table 3 illustrates the models performance by year while including the MISC category. Note however, that the smaller amount of data might cause these results to fluctuate more (i.e., 30 letters per time-period). Again, the large transformer models (i.e., DaCy large and spaCy trf) are generally superior. However, all models are clearly struggling with the oldest letters from 1844 – 1871.

SpaCy small's performance is most notably affected by the historic text with an F1-score of 0.25 in strict mode. DaCy medium performs surprisingly good on the oldest data, achieving the highest F1-score in strict mode. DaCy Small also performs quite good on all time periods except for the oldest.

The accuracy of the models incrementally increases when moving towards the more recent letters from 1880 – 1889. It is surprising, that most models drop a bit in performance on the letters from 1890 – 1905. This warranted a further inspection of those letters, as the pattern is otherwise the opposite. This finding will be expanded upon in the discussion.

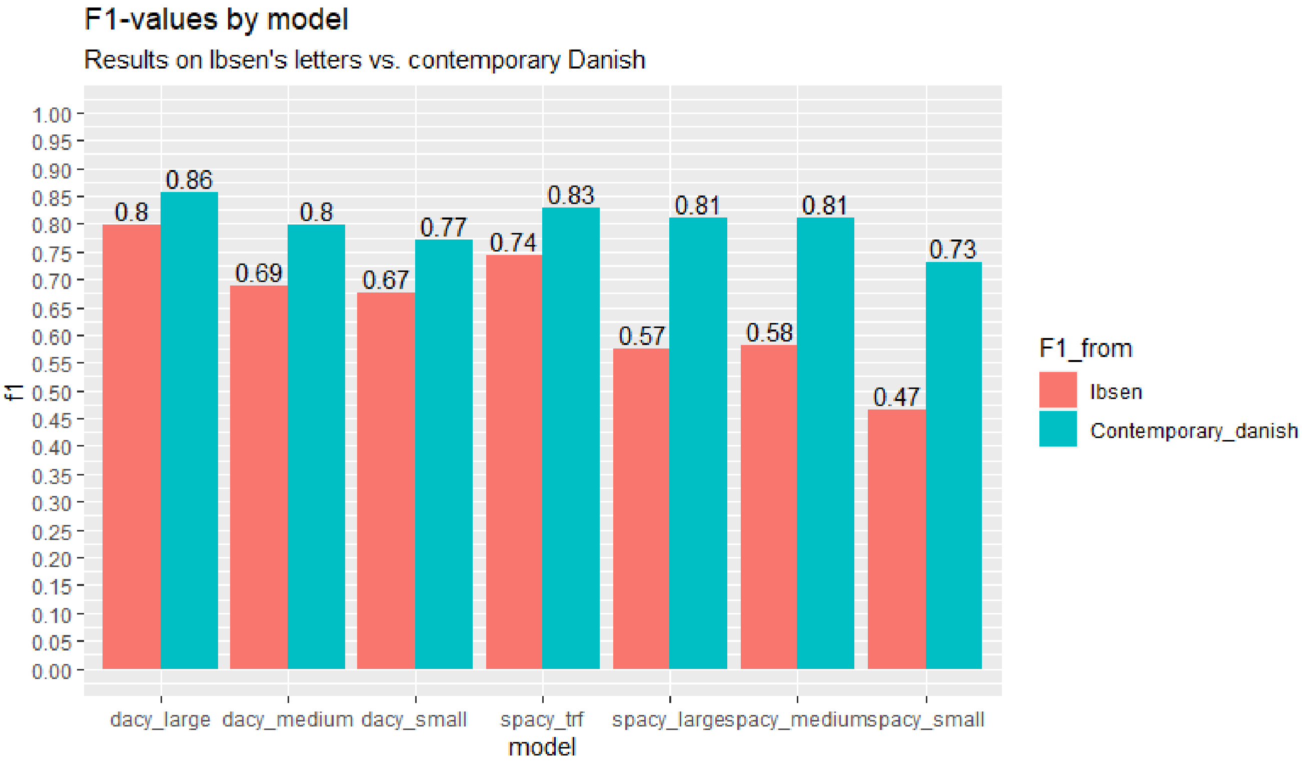


Figure 3 - Illustrates the difference in F1-scores obtained on contemporary Danish versus historic Danish from Ibsen's letters. The F1-scores are in the strict mode and excluding the MISC category. Note that the DaCy scores have been divided by 100, so that they are on the same scale as the other F1-scores.

Figure 3 illustrates the F1-scores achieved on Ibsen's letters compared to the reported results of each model on contemporary Danish. The reported results of the models can be found in the DaCy paper and on SpaCy's GitHub (Enevoldsen et al., 2021; Honnibal & Montani, 2021). The F1-scores correspond to the strict mode, and the MISC category is excluded. All models experience a performance drop. All models originally score around 0.8. The non-transformer models experience a marked performance drop compared to the transformer models. This makes it clear that historic Danish, as presumed, is more difficult for the models than contemporary Danish.

4 Discussion

The following section will discuss each of the three questions of interest specified in the introduction of this paper. Firstly, a description of the extent to which models decrease in accuracy on historic Danish text. Secondly, a discussion of the specific challenges that are causing the models to decrease in accuracy. Thirdly, future recommendations for achieving solid NER results on historic Danish text will be put forth. Finally, limitations of this paper will be defined.

4.1 Q1 - To what extent does contemporary Danish NLP models decrease in accuracy on historic text?

It is clear from the results that the larger models are obtaining better results than the smaller models, which is in line with current empirical NLP research (Kaplan et al., 2020). DaCy large achieved the best result across all parameters. Including the MISC category it achieved an F1-score of 0.76. As one would probably not use NER on Ibsen's letters without being interested in the entities of his poems and plays, including the MISC category provides a more holistic estimate of the models' performance. Employing the model in practice, this is the result that would be achieved. DaCy large performs at an F1-score equaling 0.84 while including the MISC category on contemporary text. Therefore, DaCy large's accuracy drops by 0.08 points. According to these results, DaCy Large is the most robust model, when applied to historic data. However, it is the largest model available, meaning that it takes considerable time and power to employ. The non-transformer models experience a severe decrease in F1-score, with SpaCy small dropping 0.26 points and SpaCy medium dropping 0.23 points. Such a severe decrease in accuracy means that these models would not be usable for historic Danish text. DaCy small is notably performing better than SpaCy large, which is impressive, and exemplifies the superiority of the transformer models. Considering only the data with the oldest orthography from 1844 – 1871, SpaCy trf performs best with an F1-score of 0.69 in strict mode. However, it still experiences a performance drop of 0.14 points, scoring substantially lower than on contemporary text. Even if one knew beforehand to employ a transformer model for higher accuracy, DaCy small only scores 0.48, with a severe drop of 0.29 points. Therefore, this conclusion alone cannot serve as the recommendation for future historic NLP research.

Summarized, the extent to which the accuracy of state-of-the-art NLP models decrease on historic text is quite serious. Especially for the text of the oldest orthography. Considering that there is much Danish text older than 1844, it is conjectured that the results would be even worse if contemporary models were applied. The orthography is clearly important for NLP models, meaning that the decrease in accuracy on older data could quickly make even the state-of-the-art models unusable.

4.2 Q2 - Which challenges does historic text pose to the models?

The results in this paper suggests that there are several challenges for NLP models when applied to historic text. From the partial mode of the evaluation, it appears that the models suffer from incorrect delimitation. Considering the results excluding the MISC type, DaCy large performs at F1 equal to 0.86 in partial mode and 0.85 in type mode. This means that the drop to 0.80 in strict mode is at least partially caused by improper delimitation. It is even more evident considering the data from 1844 – 1871. Here, DaCy large achieves an F1-score of 0.65 in strict mode and 0.80 in partial mode, which is a drop of 0.15 points. These results imply that different orthographic conventions have a severe impact on the performance of state-of-the-art NLP models.

It also appears that especially the ORG and MISC types are difficult for the models to predict. In these categories, the models experience the greatest drop in performance. These categories generally are more difficult for NER models. One reason the performance within these two categories may have dropped more than in the other categories, is that the entities are completely unknown to the models. Only the LOC category remains at a relatively high F1-score across all models. This is surprising considering that the PER category usually scores highest. This is the case for the DaCy models on contemporary Danish, while the results by category is unavailable for SpaCy. However, this suggests that the orthography of the person NE's is troublesome for the models. Recall that Ibsen often used colons as abbreviations in names and used job descriptions in front of surnames. It appears that this orthography influenced the models' ability to correctly detect person NE's.

It was also noteworthy that the best performance achieved across time periods was not for the data from 1890 – 1905. This prompted an inspection of the language of each letter in the period, which revealed that the oldest orthography was present in four of the letters. These included capitalizations of nouns, colon abbreviations, and altered spelling. The same orthography was only present in one of the letters from 1880 – 1889. Based on the number of predictions made by each model, something within these letters is causing the models similar problems as in the oldest data (see appendix E). This could explain why most models experienced a drop in performance on that time-period.

4.3 Q3 – How should future research on historic Danish proceed?

Retrospectively, knowing the importance of the orthography for the results, the letters should have been divided based on orthography rather than the arbitrary time-period it was written in. This observation relates back to the requirement of clear delimitation between historic texts. For future research, it would be relevant to quantify how different the language in each text is from contemporary Danish. This could theoretically be achieved by encoding each orthographic convention and summarizing the scores per text. It should also be feasible to create an algorithm for this purpose, as the result does not have to be precise. Depending on the amount of data available, these delimitations could then be used to fine-tune a larger model trained on historic text. Clearer orthography-based delimitations would also allow a researcher to decide which of the hypothetical historic models to use in the first place, given the orthography of his text.

Overall, the results suggest that NLP on historic text could be improved. This makes sense as contemporary language models are trained on contemporary language data. While the results suggest that the currently available DaCy large is usable on the historic text, there is a general drop-off in accuracy. Especially while considering the oldest text from 1844 – 1871. Therefore, the best way to improve the models would be to train new language models on historic language data. However, NLP models require a lot of input data to be trained on. Currently, digitized historic Danish text is not easy to come by. Even though the Royal library is in the process of digitalizing historic Danish text, there is still a long way to go before a resource is available to train models on. The data needs to be gathered, and a fair bit also needs to be annotated as in this paper. Therefore, a model trained on historic Danish data won't be available in the immediate future.

An alternative that could perhaps improve the current models, would be to fine-tune them on augmented Danish text. The creators of DaCy tested their models on this type of text, to provide an idea of how the models would perform on different domains. These augmentations manipulate the text they are used on by introducing spelling errors, removing whitespace, and creating abbreviations of names (Enevoldsen et al., 2021). They could be similar to how historic Danish appears to a NLP model, thus serving as a proxy for a corpus of annotated historic Danish text. One might even create specific augmentations based on a sample of historic Danish, allowing contemporary Danish to mimic historic Danish. Therefore, fine-tuning DaCy or SpaCy models on augmented contemporary Danish text, could provide an intermediate model for historic Danish, while a specified model is under development. Naturally, this would have to be investigated further, before knowing if it would improve the results of the different models. Whether fine-tuning an existing model, or training a new one, the results in this paper suggests DaCy Large should be referenced as a starting point. Therefore, it should be a transformer model that is quite large. These two features together, made for the most robust model which is essentially what is needed for historic NLP models.

4.4 Limitations

Initially, the state of manual annotation guidelines serves as a limitation. These were found to be generally vague, meaning that situations of doubt were difficult to solve. This problem has also been previously reported (see Fort et al. (2009)). This means that the annotations contain subjective decision-making resulting in decreased replicability (see differing results of the second annotation in Hvingelby et al., (2020)). This is generally mitigated by employing several annotators and thereby, inter-annotator control. It would also have been preferred to use inter-annotator control in this paper. The tags provided by the Ibsen Center provided an overview of the scope of this problem and mitigated a certain degree of it. The tags also exemplified the need for domain specific knowledge of Ibsen. Unfortunately, the tags were not present for all NE's, which serves as a limitation on the lacking domain specific knowledge of myself. Additionally, when describing the orthographic conventions, some practices may have been missed in Ibsen's letters. Linguists are often employed in annotation, especially when it is also relevant to distinguish between word types. If future research should train a new language model on historic text, the involvement of a linguist would be preferred.

In this paper, only the performance of NER was investigated. The evaluation of the models based on NER gives an idea of the models' capabilities on historic Danish text. NER is a token-level classification task, which means that the models are tested on the sentence level. As this paper was meant to gauge the capabilities of Danish NLP on historic text, NER was well suited. However, current NLP models can do many things. To properly say which model is superior and which elements are lacking, it is prudent to conduct a fuller evaluation. I am also confident that more results could have been found from this data, by investigating the construction of the SpaCy and DaCy models in depth and describing how their predictions were off. However, this was not within the scope of this paper.

5 Conclusion

This paper investigated how well Danish state-of-the-art models performed on historic Danish text from 1844 – 1905. The text was annotated using the categories: person, location, organization and miscellaneous. The models were then evaluated by comparing their NE predictions to the ground truth annotations. The results show that all models experienced a drop-off in F1-scores by being applied to historic text. DaCy large performed best overall, experiencing the mildest drop of 0.06 points. SpaCy small's performance decreased the most with a drop of 0.26 points. Generally, the transformer models performed better than the non-transformer models. Additionally, the larger models also performed better than the smaller models. The F1-scores achieved were markedly lower on the oldest data from 1844 – 1871. This suggests that the orthographic conventions of the 19th century affected the performance of the models. Due to the drop-off in accuracy, which is severe for some models, it is recommended that new language models are trained on a corpus of historic text for future research.

6 Bibliography

ACE (Automatic Content Extraction) English Annotation Guidelines for Relations. (2008, June 13). Linguistic Data Consortium, Version 6.6. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

Chai, Y. (2019). *Eval4ner: An All-Round Evaluation for Named Entity Recognition* (0.0.4) [Python]. GitHub. https://github.com/cyk1337/eval4ner

Chinchor, N., & Sundheim, B. (1993). MUC-5 Evaluation Metrics. *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore*, 10.

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2021). Named Entity Recognition and Classification on Historical Documents: A Survey. *ArXiv:2109.11406 [Cs]*, 39.

Enevoldsen, K., Hansen, L., & Nielbo, K. (2021). DaCy: A Unified Framework for Danish NLP. *ArXiv:2107.05295 [Cs]*. http://arxiv.org/abs/2107.05295

Fort, K., Ehrmann, M., & Nazarenko, A. (2009). Towards a methodology for named entities annotation. *Proceedings of the Third Linguistic Annotation Workshop on - ACL-IJCNLP '09*, 142–145. https://doi.org/10.3115/1698381.1698406

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. *Morgan & Claypool Publishers*, 311. https://doi.org/10.2200/S00762ED1V01Y201703HLT037

Hjorth, E. (2018). *Dansk Sproghistorie—Ord for ord for ord* (H. Galberg Jacobsen, B. Jørgensen, B. Jacobsen, M. Korvenius Jørgensen, & K. Kristian Fahl, Eds.; 1st ed., Vol. 1–6). Aarhus universitetsforlag.

Honnibal, M., & Montani, I. (2021, July 6). *Danish · spaCy Models Documentation*. Danish. https://spacy.io/models/da https://github.com/explosion/spacy-models/releases?q=danish&expanded=true

Hvingelby, R., Pauli, A. B., Barrett, M., Rosted, C., Lidegaard, L. M., & Søgaard, A. (2020). DaNE: A Named Entity Resource for Danish. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 8.

Johannsen, A., Martínez Alonso, H., & Barbara, P. (2015). *The Danish UD treebank* (Version 2) [Computer software]. Universal Dependencies. https://github.com/UniversalDependencies/UD_Danish-DDT (Original work published 2015)

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *ArXiv:2001.08361 [Cs, Stat]*. http://arxiv.org/abs/2001.08361

Microsoft. (2021). *Visual Studio Code* (1.62.2) [Computer software]. https://code.visualstudio.com/docs

Montani, I., Honnibal, M., Boyd, A., Landeghem, S. V., Peters, H., McCann, P. O., Samsonov, M., Geovedi, J., O'Regan, J., Orosz, G., Altinok, D., Kristiansen, S. L., Roman, Explosion Bot, Fiedler, L., Howard, G., Phatthiyaphaibun, W., Tamura, Y., Bozek, S., … Dubbin, G. (2021). *SpaCy: V3.2.0: Registered scoring functions, Doc input, floret vectors and more* (3.2.0) [Python]. Zenodo. https://doi.org/10.5281/zenodo.5648257

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvistcæ Investigationes*, *Volume 30*(Issue 1), 24. https://doi.org/10.1075/li.30.1.03nad

Piotrowski, M. (2012). Natural Language Processing for Historical Texts. *Morgan & Claypool Publishers*, 159. https://doi.org/10.2200/S00436ED1V01Y201207HLT017

Plank, B. (2020a). Neural Cross-Lingual Transfer and Limited Annotated Data for Named Entity Recognition in Danish. *ArXiv:2003.02931 [Cs]*. http://arxiv.org/abs/2003.02931

Plank, B. (2020b). Neural Cross-Lingual Transfer and Limited Annotated Data for Named Entity Recognition in Danish. *ArXiv:2003.02931 [Cs]*. http://arxiv.org/abs/2003.02931

Plank, B. (2021). Cross-Lingual Cross-Domain Nested Named Entity Evaluation on English Web Texts. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1808–1815. https://doi.org/10.18653/v1/2021.findings-acl.158

Plank, B., Jensen, K. N., & van der Goot, R. (2020). DaN+: Danish Nested Named Entities and Lexical Normalization. *Proceedings of the 28th International Conference on Computational Linguistics*, 6649–6662. https://doi.org/10.18653/v1/2020.coling-main.583

R Core Team. (2021). *R: A language and environment for statistical computing* (4.0.3) [R]. https://www.R-project.org/

Richardson, L. (2012). *Beautifulsoup4* (4.9.1) [Python]. https://www.crummy.com/software/BeautifulSoup/bs4/doc/#

RStudio Team. (2020). *RStudio: Integrated Development Environment for R* (1.4.1103) [R]. http://www.rstudio.com/

Sang, E. F. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *ArXiv:Cs/0306050*. http://arxiv.org/abs/cs/0306050

Sang, E. F. T. K., & De Meulder, F. (2005, December 5). *Language-Independent Named Entity Recognition (II)*. https://www.clips.uantwerpen.be/conll2003/ner/ https://www.clips.uantwerpen.be/conll2003/ner/annotation.txt

Segura-Bedmar, I., Martinez, P., & Zazo, M. H. (2013). SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 10.

Strømberg-Derczynski, L., Ciosici, M. R., Baglini, R., Christiansen, M. H., Dalsgaard, J. A., Fusaroli, R., Henrichsen, P. J., Hvingelby, R., Kirkedal, A., Kjeldsen, A. S., Ladefoged, C., Nielsen, F. Å., Petersen, M. L., Rystrøm, J. H., & Varab, D. (2021). The Danish Gigaword Project. *ArXiv:2005.03521 [Cs]*. http://arxiv.org/abs/2005.03521

Swalin, A. (2018, May 17). Choosing the Right Metric for Evaluating Machine Learning Models—Part 2. *USF-Data Science*. https://medium.com/usf-msds/choosing-the-right-metric-for-evaluating-machine-learning-models-part-2-86d5649a5428

University of Oslo. (2017, December 14). *Home—Centre for Ibsen Studies*. https://www.hf.uio.no/is/english/index.html

van Rossum, G., & L. Drake Jr., F. (2009). *Python 3 Reference Manual*. Createspace.

Vi digitaliserer samlingerne—Det Kongelige Bibliotek. (n.d.). Retrieved December 3, 2021, from http://www5.kb.dk/da/nb/digitalisering

Vikør, L. S. (2021). Rettskrivingsreforma av 1907. In *Store norske leksikon*. http://snl.no/Rettskrivingsreforma_av_1907

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., & Houston, A. (2013). *OntoNotes Release 5.0* (p. 2806280 KB) [Data set]. Linguistic Data Consortium. https://doi.org/10.35111/XMHB-2B84

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wiingaard, J. (2012, August 31). *Henrik Ibsen i Gyldendals Teaterleksikon | lex.dk*. Gyldendals Teaterleksikon. https://teaterleksikon.lex.dk/Henrik_Ibsen

7 Appendix

7.1 Appendix A

Orthographic convention	Example from letters	Contemporary	English
Capitalized nouns	Luerne	ilden	The fire
	Aften	aften	evening
	Regnskab	regnskab	account
Double vowels	Nødvendigviis	nødvendigvis	necessarily
	Afviisning	Afvisning	Denial
	Maaskee	Måske	Maybe
Accent signs. Only found for: é	måské	måske	maybe
	Sér	Ser	Sees
	Véd	Ved	Know
	Imødesés	Imødeses	Accommodate
Double consonants were used until 1871	Oppgjør	Opgør	Showdown
	Till	Til	For
	Forsikkring	Forsikring	Ensurance
	Rett	Ret	Right
	Brudd	Brud	Break
The letter 'q' was replaced by 'k' in 1871	Æqvivalent	Ækvivalent	Compensation
	Qvartals	Kvartals	Quarterly
	Samqvem	Samkvem	Companionship
The letter 'j' was used in front of 'gj' and 'kj'	Afgjørelse	Afgørelse	Decision
	Gjentager	Gentager	Repeats
	Kjøbe	Købe	Buy
	Forskjellige	Forskellige	Different
Foreign words did not have a Danish spelling before 1889	Convolut	Konvolut	Envelope
	Theater	Teater	Theater
	Manuscriptet	Manuskriptet	The manuscript
	Buxetøiet	Buksetøjet (bukserne)	The pants

The letter ‘x’ is replaced by ‘ks’ in 1889	Vexel	Veksel	Change
	Voxede	Voksede	Grew
	Strax	Straks	Immediately
The rules for conjugation of words are altered in 1900	Saameget	Så meget	So much
	Saagodtsom	Så godt som	As good as
	Forlængesiden	For længe siden	Long ago
	Paany	På ny	Anew
The use of ‘aa’ was altered to ‘å’ from 1826 (officially in 1948)	Daaben	Dåben	Baptism
	Maa	Må	Must
	Aar	År	Year
	Middelmaadige	Middelmådige	Average
The word ending ‘nd’ and ‘ld’ in simple past tense is changed to ‘nn’ and ‘ll’ in 1948.	Skulde	Skulle	Should
	Vilde	Ville	Would
	Kunde	Kunne	Could
Spelling words with ‘ie’ instead of ‘je’ (unknown rule, but observed in letters)	Eie	Eje	Own
	Høiere	Højere	Higher

Table 4 – An overview of the orthographic conventions of historic Danish observed in Henrik Ibsen’s letters. Column 2

7.2 Appendix B

Software metadata	
Computing platform / Operating System	Microsoft Windows 11 x64
Current software version	- Python 3.6.6, JupyterLab 2.1.5 - Anaconda3 - R 4.0.3, RStudio 1.4.1103. - Visual Studio Code 1.62.2
Packages and models	- Beautifulsoup4 (4.9.1) - DaCy (1.1.4) o da_dacy_large-trf_0.1.0 o da_dacy_medium-trf_0.1.0 o da_dacy_small_trf-0.1.0 - SpaCy (3.0.6) o da_core_news_trf-3.1.0 o da_core_news_sm-3.1.0 o da_core_news_md-3.1.0 o da_core_news_lg-3.1.0 - Eval4ner (0.0.4) - Pandas (1.0.5) - Glob2 (0.7) - Ast - Os - Pprint - Tidyverse (1.3.0)

Table 5 – An overview of the software, including versions, employed in this paper.

7.3 Appendix C

Letters annotated and used for the analysis			
1844 – 1871	1871 – 1879	1880 – 1889	1890 - 1905
BREV_B1844-1871ht_B18450801AWE	BREV_B1871-1879ht_B18720402EGo	BREV_B1880-1889ht_B18800319NL	BREV_B1890-1905ht_B18900525BBj
BREV_B1844-1871ht_B18461207JCP	BREV_B1871-1879ht_B18720411LL	BREV_B1880-1889ht_B18800716FH	BREV_B1890-1905ht_B18910104CHon
BREV_B1844-1871ht_B18491015OS	BREV_B1871-1879ht_B18730206FH	BREV_B1880-1889ht_B18810106LuDa	BREV_B1890-1905ht_B18910111JE
BREV_B1844-1871ht_B18500105OS	BREV_B1871-1879ht_B18730928PNA	BREV_B1880-1889ht_B18810327HEB	BREV_B1890-1905ht_B18910204CK
BREV_B1844-1871ht_B18500728AkKoll	BREV_B1871-1879ht_B18731004JHT	BREV_B1880-1889ht_B18811222LP	BREV_B1890-1905ht_B18910405HAB
BREV_B1844-1871ht_B18640307BD	BREV_B1871-1879ht_B18731203FH	BREV_B1880-1889ht_B18820102FH	BREV_B1890-1905ht_B18911022AuLa
BREV_B1844-1871ht_B18651025FH	BREV_B1871-1879ht_B18740723RSch	BREV_B1880-1889ht_B18820218HEB	BREV_B1890-1905ht_B18920829JH
BREV_B1844-1871ht_B18660304FH	BREV_B1871-1879ht_B18740921ThK	BREV_B1880-1889ht_B18820909FH	BREV_B1890-1905ht_B18931127OMA
BREV_B1844-1871ht_B18660415GS	BREV_B1871-1879ht_B18741023PB	BREV_B1880-1889ht_B18831227FH	BREV_B1890-1905ht_B18941018HBa
BREV_B1844-1871ht_B18660425FH	BREV_B1871-1879ht_B18750204LDa	BREV_B1880-1889ht_B18841003SuI	BREV_B1890-1905ht_B18950120AuLa
BREV_B1844-1871ht_B18660505BB	BREV_B1871-1879ht_B18750224DT	BREV_B1880-1889ht_B18851110VG	BREV_B1890-1905ht_B18950312SuI
BREV_B1844-1871ht_B18660505BD	BREV_B1871-1879ht_B18750303EGr	BREV_B1880-1889ht_B18860113EBo	BREV_B1890-1905ht_B18950413SuI
BREV_B1844-1871ht_B18660521FH	BREV_B1871-1879ht_B18750310CHo	BREV_B1880-1889ht_B18870320EK	BREV_B1890-1905ht_B18950702SuI
BREV_B1844-1871ht_B18671104JB	BREV_B1871-1879ht_B18750310JHT	BREV_B1880-1889ht_B18871006MGo	BREV_B1890-1905ht_B18950726SiI
BREV_B1844-1871ht_B18680622JB	BREV_B1871-1879ht_B18750617FH	BREV_B1880-1889ht_B18871108JHo	BREV_B1890-1905ht_B18960106NL
BREV_B1844-1871ht_B18691217JHT	BREV_B1871-1879ht_B18750918JHT	BREV_B1880-1889ht_B18871115HOst	BREV_B1890-1905ht_B18960911NN_Til_det
BREV_B1844-1871ht_B18700126JB	BREV_B1871-1879ht_B18760206JHT	BREV_B1880-1889ht_B18871117FH	BREV_B1890-1905ht_B18961129AuLa
BREV_B1844-1871ht_B18700308JHT	BREV_B1871-1879ht_B18760614LJ	BREV_B1880-1889ht_B18871122IA	BREV_B1890-1905ht_B18970210Odel

BREV_B1844-1871ht_B18700505FH	BREV_B1871-1879ht_B18760915FH	BREV_B1880-1889ht_B18871125CS	BREV_B1890-1905ht_B18970530SiI
BREV_B1844-1871ht_B18700512JPA	BREV_B1871-1879ht_B18770729FH	BREV_B1880-1889ht_B18880413KAWi	BREV_B1890-1905ht_B18971112JH
BREV_B1844-1871ht_B18700529MT	BREV_B1871-1879ht_B18770903MG	BREV_B1880-1889ht_B18880428AC	BREV_B1890-1905ht_B18980327VBogh
BREV_B1844-1871ht_B18700604MT	BREV_B1871-1879ht_B18770912SuI	BREV_B1880-1889ht_B18880521LG	BREV_B1890-1905ht_B18981230RFi
BREV_B1844-1871ht_B18701106FH	BREV_B1871-1879ht_B18770925EK	BREV_B1880-1889ht_B18880723LK	BREV_B1890-1905ht_B18991124WA
BREV_B1844-1871ht_B18710108AK	BREV_B1871-1879ht_B18771005EF	BREV_B1880-1889ht_B18880830EK	BREV_B1890-1905ht_B19010102ChrSo
BREV_B1844-1871ht_B18710208FH	BREV_B1871-1879ht_B18780719AuLa	BREV_B1880-1889ht_B18881121AuLa	BREV_B1890-1905ht_Budat1891-92HA
BREV_B1844-1871ht_B18710215CFS	BREV_B1871-1879ht_B18790421WAM	BREV_B1880-1889ht_B18890103RS	BREV_B1890-1905ht_Budat1891JLi
BREV_B1844-1871ht_B18710217GB	BREV_B1871-1879ht_B18790712BB	BREV_B1880-1889ht_B18890227BHa	BREV_B1890-1905ht_Budat18930116NN_Fest
BREV_B1844-1871ht_B18710513JB	BREV_B1871-1879ht_B18791114FH	BREV_B1880-1889ht_B18890629NL	BREV_B1890-1905ht_Budat18950427SuI
BREV_B1844-1871ht_Budat18680831JB	BREV_B1871-1879ht_Budat187409VVo	BREV_B1880-1889ht_B18891014AL	BREV_B1890-1905ht_Budat189504SuI
BREV_B1844-1871ht_Budat18700720AK	BREV_B1871-1879ht_Budat1879NN_Kare_ven	BREV_B1880-1889ht_Budat188510JLi	BREV_B1890-1905ht_Budat189509NN_Et_Tidsrum

Table 6 – Illustrates the ID of the letters annotated and used in the analysis.

7.4 Appendix D

	1844-1871				1871-1879			
Model	Strict	Exact	Partial	type	Strict	Exact	Partial	Type
DaCy large	**0.72**	**0.76**	**0.80**	0.78	**0.81**	**0.85**	**0.88**	**0.84**
DaCy medium	0.71	0.74	0.78	0.77	0.73	0.76	0.80	0.81
DaCy small	0.53	0.57	0.63	0.63	0.69	0.76	0.78	0.72
SpaCy trf	0.71	0.74	0.80	**0.81**	0.80	0.85	0.87	0.84
SpaCy large	0.41	0.45	0.52	0.51	0.64	0.70	0.72	0.68
SpaCy medium	0.42	0.46	0.53	0.54	0.62	0.66	0.67	0.65
SpaCy small	0.24	0.30	0.35	0.29	0.46	0.59	0.60	0.46

	1880-1889				1890-1905			
Model	Strict	Exact	Partial	Type	Strict	Exact	Partial	Type
DaCy large	**0.85**	**0.88**	**0.90**	**0.89**	**0.82**	**0.84**	**0.87**	0.86
DaCy medium	0.71	0.74	0.79	0.81	0.65	0.67	0.74	0.78
DaCy small	0.79	0.86	0.87	0.82	0.69	0.73	0.76	0.76
SpaCy trf	0.72	0.75	0.81	0.84	0.75	0.79	0.86	**0.88**
SpaCy large	0.63	0.69	0.74	0.72	0.61	0.67	0.73	0.73
SpaCy medium	0.68	0.73	0.77	0.75	0.60	0.61	0.66	0.68
SpaCy small	0.63	0.75	0.76	0.65	0.54	0.67	0.70	0.59

Table 7 - NER performance across the four time periods. The best scores are marked ***bold*** *and second best are underlined. The F1-values are rounded to 2 digits. In case of equal values, the unrounded value takes precedence. The miscellaneous type is not included in these scores.*

7.5 Appendix E

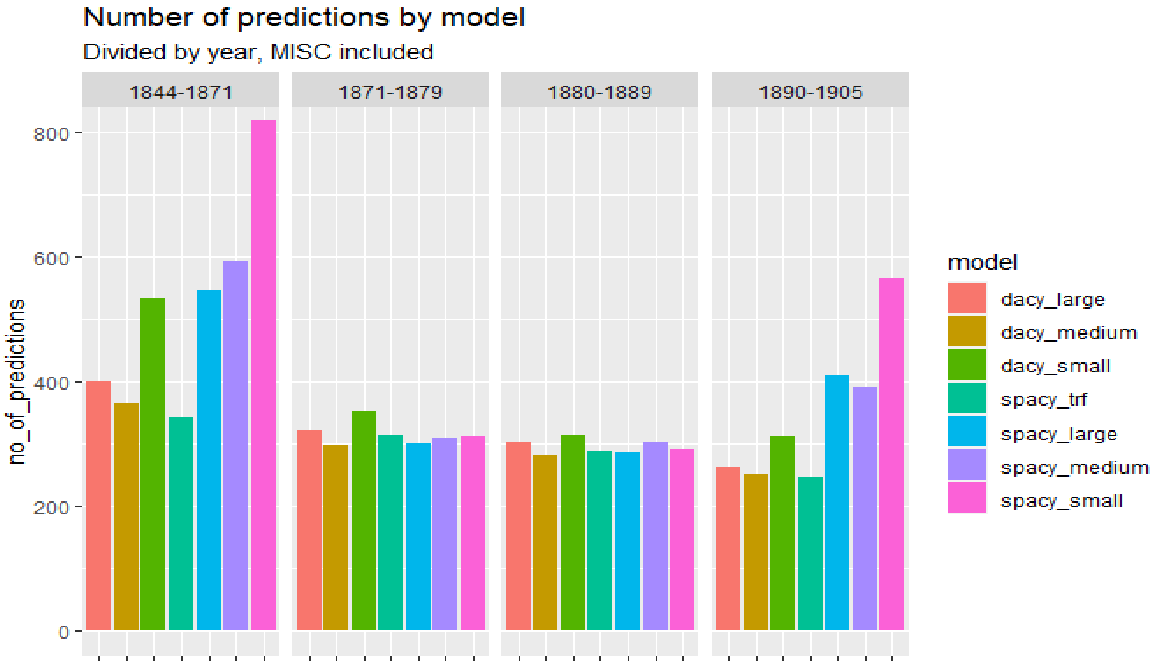


Figure 4 – An illustration of the number of predicted NE's by each model by time period. Note the similarity between the predictions on the oldest data and the newest data. This implies that they have something in common that cause the models problems.

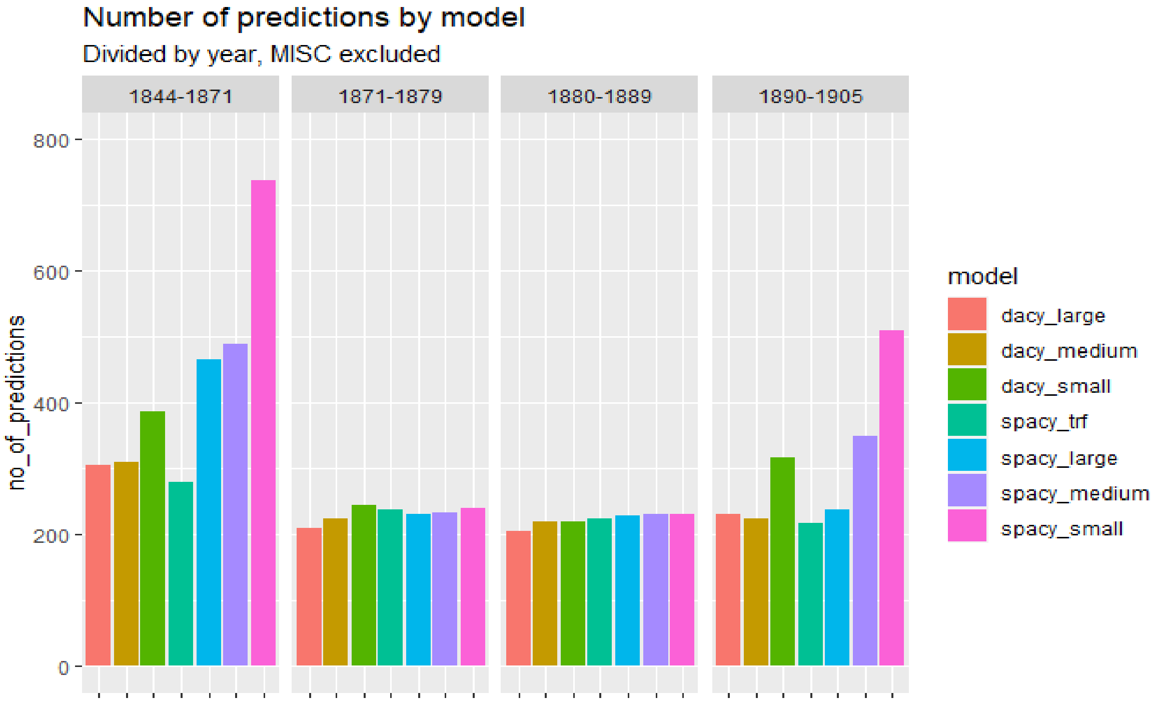


Figure 5 – An illustration of the number of predicted NE's by each model by time period. The same pattern is found as in the plot including MISC.