Development and Evaluation of Pre-trained Language Models for Historical Danish and Norwegian Literary Texts

Ali Al-Laith[1,2], **Alexander Conroy**[1], **Jens Bjerring-Hansen**[1] **and Daniel Hershcovich**[2]
Department of Nordic Studies and Linguistics, University of Copenhagen[1]
Department of Computer Science, University of Copenhagen[2]
alal@di.ku.dk, alc@hum.ku.dk,
jbh@hum.ku.dk, dh@di.ku.dk

Abstract

We develop and evaluate the first pre-trained language models specifically tailored for historical Danish and Norwegian texts. Three models are trained on a corpus of 19th-century Danish and Norwegian literature: two directly on the corpus with no prior pre-training, and one with continued pre-training. To evaluate the models, we utilize an existing sentiment classification dataset, and additionally introduce a new annotated word sense disambiguation dataset focusing on the concept of *fate*. Our assessment reveals that the model employing continued pre-training outperforms the others in two downstream NLP tasks on historical texts. Specifically, we observe substantial improvement in sentiment classification and word sense disambiguation compared to models trained on contemporary texts. These results highlight the effectiveness of continued pre-training for enhancing performance across various NLP tasks in historical text analysis.

Keywords: Pre-trained Language Models, Digital Humanities, Sentiment Analysis, Word Sense Disambiguation

1. Introduction

The wealth of digitized historical texts enriches research in Natural Language Processing (NLP) and Digital Humanities. Embedded with archaic terminology, context-sensitive variations, and cultural idiosyncrasies, these deviate vastly from contemporary texts (Piotrowski, 2012). For example, the Danish term 'skæbne' (fate; see §3.2) evolved from a divine/metaphysical sense to a secular/material one, influencing perceptions of responsibility and personal actions. This demonstrates language's dynamism and the significance of historical context in time-bounded textual analysis.

To navigate the complexities of historical texts, Pre-trained Language Models (PLMs), e.g., based on BERT (Devlin et al., 2019), have been trained on historical corpora across various languages, improving performance across a broad spectrum of tasks. While such tailored PLMs exist for a range of languages, including English (Manjavacas and Fonteyn, 2021; Beelen et al., 2021; Rastas et al., 2022), Italian (Palmero Aprosio et al., 2022), French (Gabay et al., 2022), and Greek (Yamshchikov et al., 2022), resources remain limited for others, including Danish and Norwegian. Addressing this gap, we train three PLMs on the MeMo (*Measuring Modernity*) corpus (§3.1)—a collection of Danish and Norwegian novels spanning the period 1870–1900 (Bjerring-Hansen et al., 2022). We do not distinguish written Norwegian from Danish since until 1907 they were practically identical (Vikør, 2022). We assess model adaptability using a sentiment analysis benchmark (SA; Allaith et al., 2023) and a novel word sense disambiguation (WSD) dataset, observing significant improvement over models trained on contemporary corpora. This work aims to enhance understanding of historical linguistic changes in Danish and Norwegian, and bolster tools for Digital Humanities researchers. The developed models and datasets are available on HuggingFace[1].

2. Related Work

Historical literary PLMs. PLMs have been adapted to historical literary texts through two key strategies: modifying modern PLMs to fit them or directly training models on them (Manjavacas and Fonteyn, 2022). The success of these methods in historical language-specific NLP tasks is noted. However, the focus has been predominantly on English, with models like a historical BERT adaptation, MacBERTh and ECCO-BERT demonstrating effectiveness (Palmero Aprosio et al., 2022; Manjavacas and Fonteyn, 2021; Rastas et al., 2022). BERToldo, created for historical Italian, uniquely combines both training strategies and shows superior performance on Dante Alighieri's texts (Palmero Aprosio et al., 2022). For Early Modern French, D'AlemBERT, trained on a corpus from the 16th to the 18th centuries, exhibits enhanced part-of-speech tagging (Gabay et al., 2022). In Ancient Greek studies, a BERT model fine-tuned on Plutarch's corpus showed promise in text classification and sentiment analysis, suggesting value

[1] https://huggingface.co/MiMe-MeMo

LREC-COLING 2024, pages 4811–4819
20-25 May, 2024.

in a transfer learning approach (Yamshchikov et al., 2022). To fill the gap for historical Danish and Norwegian, we leverage both adaption of contemporary PLMs and direct training.

Contemporary Danish PLMs. Several PLMs have been trained on (mostly) contemporary Danish (and Norwegian) texts. DanBERT[2] is a Danish PLM based on the BERT-Base architecture, trained on more than 40 millions Danish words of unknown source. DanskBERT[3] is a Danish PLM trained by Snæbjarnarson et al. (2023) on the Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021), a billion-word corpus containing a wide variety of time periods (mostly contemporary texts, but also including 25.6 million words from Danish literature spanning 1700–now), domains (mostly legal and social media), socio-economic status, and dialects. Danish Model BotXO[4] is a Danish PLM trained by Certainly (previously known as BotXO) on 1.6 billion contemporary Danish words from Common Crawl, Danish Wikipedia, Danish debate forums and OpenSubtitles. ScandiBERT[5] is a PLM trained on concatenated data from five Scandinavian languages including Danish and Norwegian (Snæbjarnarson et al., 2023), including the Danish Gigaword Corpus and the Norwegian Colossal Corpus (Kummervold et al., 2022), which contains over 7 billion Norwegian words from various sources and time periods, including books.[6] Since no PLM was trained specifically on historical Danish and Norwegian data, we address this gap.

3. Datasets

We use the MeMo corpus for pretraining and two annotated subcorpora for benchmarking.

3.1. MeMo Corpus

Bjerring-Hansen et al. (2022) curated the MeMo corpus, named after the *Measuring Modernity* project. The corpus comprising 839 novels (with a corpus size of 690 MB) written in historical Danish and Norwegian languages, the MeMo corpus spans the final three decades of the 19th century, constituting a vast collection of over 52 million words. This comprehensive and diverse corpus holds promise for providing valuable insights into NLP tasks focused on historical texts. Detailed statistical information about the corpus is provided in Table 1, while Figure 1 illustrates the year-by-year distribution of the novels within our corpus. The full corpus is available online.[7] A data statement is included in Appendix A.

Total novels	839
Total sentences	3,229,137
Total words	52,724,457
Average sentences per novel	3,849
Average words per novel	62,842
Average words per sentence	16

Table 1: Summary of MeMo Corpus Statistics.

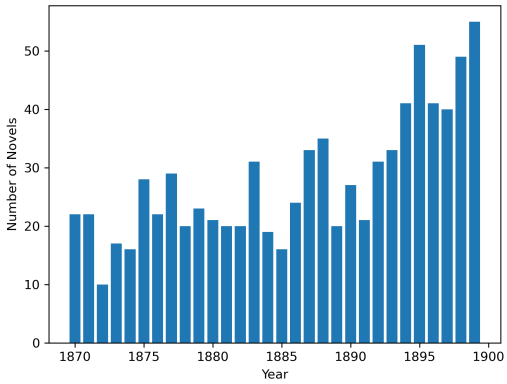

Figure 1: Distribution of Novels in the MeMo Corpus Over Time

3.2. Annotated Datasets

Sentiment analysis. The first task we consider is sentiment analysis of sentences from historical Danish and Norwegian novels. A particularly suited dataset for our evaluation purpose is the sentiment classification dataset developed by Allaith et al. (2023). The dataset consists of 2,748 sentences from the MeMo corpus, annotated manually with three sentiment classes, negative (41.4%), neutral (28.7%), and positive (29.9%).

Word sense disambiguation. The second task we approach is word sense disambiguation in historical texts. In order to address it, we introduce a novel dataset established and annotated by one of the authors (a Danish-speaking literary scholar) investigating how the concept of fate ('skæbne') is transformed in the latter part of the 19th century from its pre-modern sense, which is religiously

[2] https://huggingface.co/alexanderfalk/danbert-small-cased

[3] https://huggingface.co/vesteinn/DanskBERT

[4] https://huggingface.co/Maltehb/danish-bert-botxo

[5] https://huggingface.co/vesteinn/ScandiBERT

[6] https://github.com/NbAiLab/notram/blob/master/guides/corpus_description.md#publish-periode-7

[7] https://huggingface.co/datasets/MiMe-MeMo/Corpus-v1.1

Category	Example from the dataset	English translation
Pre-modern	Vær vis paa, at der er nøje Sammenhæng mellem Skæbne og Guds Kjærlighed, og det skal klares en Gang, men ikke ved Forstandens Kløgt.	Be sure that there is a close connection between fate and the love of God, and it must be settled once, but not by reason.
Modern	Thi ingen ung Pige havde Skyld i sin Skæbne. Det var hans Tro, at for hver Kvinde, der gik til Grunde, maatte der findes en Mand, som bar Ansvaret, hvis Brøde det var.	For no young girl was to blame for her fate; he believed that for every woman who perished, there must be a man who bore the responsibility, whose offense it was.
Figure of speech	Det var ikke, som han hidtil havde troet, en Fremmed, ingen ligegyldig, udeltagende Sømand, men derimod Doktor Hansen, han, der bestandig havde næret en saa varm og oprigtig Interesse for hans Skjæbne.	It was not, as he had previously believed, a stranger, no indifferent sailor, but rather Doctor Hansen, he who had always harbored such a warm and sincere interest in his fate.
Ambiguous	Nu vidste hun det... nu var det bleven sagt hende for sidste Gang. Der var taget Bestemmelse om hendes Skæbne... hendes Liv skulde blive forspildt og ulykkeligt.. - hun kunde ingen Naade vente.	Now she knew it... it had been told to her for the last time. A decision had been made about her fate... her life would be wasted and unhappy... she could expect no mercy.

Table 2: Abbreviated Word Sense Disambiguation examples demonstrating the four categories of the word 'skæbne'.

and metaphysically inflected, to a modern meaning where the concept incorporates a secular and material understanding of the world. The WSD dataset was established using a regular expression that captured all variations of the word 'skæbne' (fate) in the MeMo Corpus. This includes all inflections of the word as well as derivatives and compounds. This means that words like 'vanskæbne' (misfortune) and 'skæbnesvanger' (fateful), which are semantically closely related to the keyword 'skæbne,' are also included in the dataset. To capture the surrounding context of the word, we have utilized the Danish pipeline in spaCy (Honnibal et al., 2020). Specifically, we used spaCy's sentence segmentation to extract segments of nine sentences each, with the word 'skæbne' appearing in the middle sentence. This procedure resulted in a total of 8031 segments, which were subsequently shuffled to ensure that the annotation process did not follow a chronological order but rather spans across the novels and the period. The annotated dataset consists of the first 650 segments, which have been annotated according to four predefined categories: Pre-modern notions of fate, Modern notions of fate, Figures of speech, and Ambiguous cases. The first two labels apply to the dichotomy between the religious/metaphysical notions and secular/material notions, while the third label applies to cases where the word 'skæbne' appears in idioms and narrative clichés, i.e., segments that do not express a particular worldview. The first two labels reflect the historical senses of the word, while the third label does not represent a distinct word sense, although it reflects some linguistic properties. The fourth and final label pertains to the linguistic and cultural-historical complexity of the task. There are indeed several segments where, based on the given context, it is difficult to determine which of the other three labels is the most obvious. Abbreviated examples are listed in Table 2. In total, the dataset comprises 109 instances labeled as Pre-modern fate (label 0), 87 instances labeled as Modern fate (label 1), 275 instances labeled as Figures of speech (label 2), and 179 instances labeled as Ambiguous (label 3). An Inter-Annotator Agreement (IAA) was conducted on a subset of the data (100 annotated segments). The IAA result showed that the two literary scholars agreed on 69 out of the 100 annotations and disagreed on 31, resulting in a Cohen's Kappa value of 0.56. The dataset was split based on the year of each segment, with the training, validation, and testing sets containing 70%, 15%, and 15% of the data, respectively.

4. MeMo-BERTs Models

We introduce MeMo-BERTs, three PLMs designed to support historical Danish and Norwegian text. We employ two different approaches to train the models, one directly on the MeMo Corpus, (§3.1) and the other using continued pre-training based on a contemporary-language PLM.

The first two models (MeMo-BERT-1 and MeMo-BERT-2) are randomly initialized Transformer (Vaswani et al., 2017). MeMo-BERT-1 the BERT architecture (Devlin et al., 2019): 12 layers, a hidden dimension of 768, 12 attention heads, and a vocabulary size of 30,000. MeMo-BERT-2 uses the XLM-RoBERTa architecture (Conneau et al.,

2020), with 24 layers, a hidden dimension of 1024, 16 attention heads, and a subword vocabulary size of 50,000 subwords.

For the third model (MeMo-BERT-3), we start with and continue pre-trainined the Transformer PLM DanskBERT (Snæbjarnarson et al., 2023), trained on the contemporary Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021). It was in turn based on continued pre-trainined of XLM-RoBERTa, with 24 layers, a hidden dimension of 1024, and 16 attention heads. We use a subword vocabulary size of 250,000.

For all models, we use the masked language modelling objective in pre-training, where 15% of the input tokens are masked, and the model is trained to predict the original tokens. The models are all encoder-only Transformers with case sensitivity. The corpus is randomly split into 80% for training and 20% for validation. We set the batch size for training to 16 and validation to 32, the number of gradient accumulation steps to 8, the learning rate to 1e-4, the number of training epochs to 3, the maximum number of training steps to 12500, and the number of warm-up steps to 1250. We select the best checkpoint based on validation loss. These parameters have a significant impact on the convergence and performance of the trained model. The training process is performed in a distributed manner, utilizing two A100 GPUs, and takes 44, 36, and 32 hours for training the three models respectively.

5. Experiments

To evaluate the developed historical models (§4) and the baselines trained on contemporary Danish (§2), we use two downstream tasks, sentiment analysis and word sense disambiguation. These tasks were selected on the basis of their relevance for historical text processing. Both are represented by datasets annotated over text from the same MeMo corpus that we use for pre-training the PLMs. We select the comparison models based on their popularity and accuracy in similar tasks. The models were tested on diverse NLP benchmark datasets (Nielsen, 2023). We utilize them to assess the performance of our developed historical models.

Sentiment analysis. We use the same split as Allaith et al. (2023) for training, development, and testing (86%, 10%, and 4% respectively).

Word sense disambiguation. We split the dataset into training, development, and testing with the proportions 70%, 15%, and 15% respectively.

Each model is fine-tuned for 20 epochs and evaluated on each of the task separately, repeated 5 times with different random seeds. Average results are reported in Table 3.

Task	SA		WSD	
Model	**Valid.**	**Test**	**Valid.**	**Test**
MeMo-BERT-1	0.52	0.56	0.41	0.43
MeMo-BERT-2	0.58	0.59	0.44	0.35
MeMo-BERT-3	**0.78**	**0.77**	**0.55**	**0.61**
DanskBERT	0.75	0.76	0.52	0.46
Danish BERT BotXO	0.74	0.74	0.19	0.30
ScandiBERT	0.73	0.73	0.40	0.40
DanBERT	0.65	0.63	0.39	0.41

Table 3: F1-Score results of fine-tuning the selected models for sentiment analysis (SA) and word sense disambiguation (WSD) classification tasks on validation and test sets.

6. Results and Discussion

Table 3 displays the F1-Score performance of various models in both Sentiment Analysis (SA) and Word Sense Disambiguation (WSD) downstream tasks. In the SA task, MeMo-BERT-3 demonstrates superior performance, achieving an F1-Score of 0.78 and 0.77 on the validation and test sets respectively. This outperforms MeMo-BERT-1 and MeMo-BERT-2, which attained F1-Scores of 0.52/0.56 and 0.58/0.59 respectively on the same datasets. Other models such as DanskBERT, Danish BERT BotXO, ScandiBERT, and DanBERT also participated in the SA task but achieved lower F1-Scores compared to MeMo-BERT-3.

Similarly, in the WSD task, MeMo-BERT-3 exhibits remarkable performance, achieving an F1-Score of 0.55 on the validation set and 0.61 on the test set, surpassing all other models. MeMo-BERT-1 and MeMo-BERT-2 achieved F1-Scores of 0.41/0.43 and 0.44/0.35 respectively on the validation and test sets, indicating MeMo-BERT-3's superiority. DanskBERT also performed competitively in the WSD task with F1-Scores of 0.52 and 0.46 on the validation and test sets respectively.

DanskBERT, pretrained on a mixed of Danish contemporary text and historical literature, demonstrates strong performance on sentiment analysis and word sense disambiguation dataset from historical and literary text, showcasing its ability to accurately resolve the intended meanings of words in the context of archaic or literary language.

These results underscore the effectiveness of MeMo-BERT-3 in both SA and WSD tasks, highlighting its capability to capture nuanced linguistic features in the datasets, particularly benefiting from its pre-training on historical text. Moreover, the comparison against other state-of-the-art models

demonstrates MeMo-BERT-3's superiority across both tasks, validating its efficacy in natural language understanding and processing tasks.

Overall, the results indicate that language models pretrained on historical text or a combination of contemporary and historical text consistently outperformed models pretrained mostly on contemporary text. This suggests that incorporating historical language data has a beneficial impact, enhancing the model's comprehension and accurate classification of text from historical periods.

7. Conclusion

We presented the first PLMs for historical literary Danish and Norwegian, trained on the MeMo corpus that consists of 839 novels from the last 30 years of 19th century. Experiments on a novel word sense disambiguation dataset and a sentiment analysis dataset revealed our models outperform counterparts trained mostly on contemporary texts.

In future work, we will collect further historical documents (including newspapers and novels from other time periods), leveraging datasets such as the Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021) and Norwegian Colossal Corpus (Kummervold et al., 2022) for training new PLMs. Our goal is to develop PLMs that capture richer historical context. Additionally, we aim to assess the generalizability of our models by utilizing a more diverse set of historical corpora for testing, enabling us to evaluate their transferability to unseen data not encountered during pre-training. This approach will provide a more robust assessment of their generalization capabilities. Furthermore, we plan to expand annotated datasets and develop new ones for tasks such as named entity recognition and event extraction. This will allow us to assess the performance of our PLMs and enable more nuanced literary analysis using them.

8. Limitations

Our models were specifically trained and evaluated on the MeMo corpus for the purpose of addressing literary research questions on that corpus. Other notable Danish/Norwegian corpora from different historical time periods include H. C. Andersen's folk legends from the early 19th century (Tangherlini, 2022). Since we did not test the generalizability of our models to such similar corpora, nor the relationship to other close time periods, genres or geographic/cultural regions, we cannot make any claims about their fit for use in such contexts. This must be addressed by future work.

Furthermore, even for generalization and applicability for studying texts within the scope of the MeMo corpus, the annotated datasets are relatively small and may not be representative of the overall set of novels in the corpus. Nevertheless, the fact they were driven and created by literary scholars invested in the analysis of these texts means that they are relevant to the literary research questions that can be addressed by applying our PLMs to the whole corpus and proceeding with computer-assisted literary analysis.

9. Ethics Statement

The annotation work conducted in the process of creating our word sense disambiguation dataset was done as part of the literary research project of one of the authors, and was not specifically compensated for monetarily. We release the dataset under the Creative Commons Attribution 4.0 International license.[8] The sentiment dataset from (Allaith et al., 2023) is released under the Creative Commons Attribution 4.0 International license and we use it for evaluation in this study without redistributing it.

10. Bibliographical References

Ali Allaith, Kirstine Degn, Alexander Conroy, Bolette Pedersen, Jens Bjerring-Hansen, and Daniel Hershcovich. 2023. Sentiment classification of historical Danish and Norwegian literary texts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 324–334, Tórshavn, Faroe Islands. University of Tartu Library.

Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 2021. When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2751–2761.

Jens Bjerring-Hansen, Ross Deans Kristensen-McLachlan, Philip Diderichsen, and Dorte Haltrup Hansen. 2022. Mending fractured texts. a heuristic procedure for correcting OCR data.

Jens Bjerring-Hansen and Sebastian Ørtoft Rasmussen. 2023. Litteratursociologi og kvantitative litteraturstudier: Den historiske roman i det moderne gennembrud som case. *Passage-Tidsskrift for litteratur og kritik*, 38(89):171–189.

Jens Bjerring-Hansen and Matthew Wilkens. 2023. Deep distant reading: The rise of realism in

[8] http://creativecommons.org/licenses/by/4.0/

Scandinavian literature as a case study. *Orbis Litterarum*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From FreEM to D'AlemBERT: a large corpus and a language model for early modern French. *arXiv preprint arXiv:2202.09452*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Per Kummervold, Freddy Wetjen, and Javier de la Rosa. 2022. The Norwegian colossal corpus: A text corpus for training large Norwegian language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3852–3860, Marseille, France. European Language Resources Association.

EMA Manjavacas and Lauren Fonteyn. 2022. Adapting vs. pre-training language models for historical languages. *Journal of Data Mining & Digital Humanities*, pages 1–19.

Enrique Manjavacas and Lauren Fonteyn. 2021. Macberth: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2022. BERToldo, the historical BERT for Italian. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis lectures on human language technologies*, 5(2):1–157.

Iiro Rastas, Yann Ciarán Ryan, Iiro Lassi Ilmari Tiihonen, Mohammadreza Qaraei, Liina Repo, Rohit Babbar, Eetu Mäkelä, Mikko Tolonen, and Filip Ginter. 2022. Explainable publication year prediction of eighteenth century texts with the BERT model. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*. The Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. The Danish Gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Timothy R Tangherlini. 2022. The accidental folklorist: Thiele's collection of Danish folk legends in early nineteenth-century Denmark. In *Grimm Ripples: The Legacy of the Grimms' Deutsche Sagen in Northern Europe*, pages 70–105. Brill.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Lars S. Vikør. 2022. Rettskrivingsreform i store norske leksikon på snl.no. In *https://snl.no/rettskrivingsreform*.

Ivan P Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. BERT in Plutarch's shadows. *arXiv preprint arXiv:2211.05673*.

A. Data Statement

1. Header.

Dataset Title: MeMo Corpus

Dataset Curators/Authors: Jens Bjerring-Hansen, University of Copenhagen; Philip Diderichsen, University of Copenhagen; Dorte Haltrup Hansen, University of Copenhagen

Dataset Version: Version 1.1, August 15, 2023

Dataset Statement Version: 1, September 25, 2023

2. Executive summary. The MeMo corpus is established to investigate literary and cultural change in a seminal epoch of Scandinavian cultural and social history (known as 'the modern breakthrough') using natural language processing and other computational methods. The corpus consists of original novels by Norwegian and Danish authors printed in Denmark in the period 1870-99. It includes 858 volumes, totaling 4.5 million sentences and 65 million words.

3. Text characteristics. The corpus consists of novels, i.e. long works of narrative fiction, usually written in prose and published as a book. The novels contain both dialogue and description. As instances of imaginative literature they are infused with ambiguity, interpretational confounding, rhetorical sophistication, and narrative layerings between author, narrator, and characters. The cultural diversity of the texts in the corpus is pronounced. From a genre perspective, we have contemporary novels as well as historical novels and other forms of genre fiction such as romance, crime, and war stories (cf. Bjerring-Hansen and Rasmussen, 2023). And from an aesthetic perspective we have both avant-garde forms of realism, including instances of naturalism and impressionism, and more traditional prose with a preference for abstract or generalized over concrete specification (cf. Bjerring-Hansen and Wilkens, 2023).

4. Curation Rationale. The MeMo Corpus was created as the basis for a research project, MeMo – Measuring Modernity: Literary and Social Change in Scandinavia 1870-1900, investigating how processes of social change in late nineteenth century Scandinavia were reflected and discussed in the novels from the period.[9] As opposed to traditional historiography on the period, which has focused on selected texts by a few prominent, male authors, our digital corpus, with rich metadata on texts and authors, allows for the capturing of robust literary and sociological trends and for new insights into the processes of modernization in this formative period in the literary and social history of Scandinavia. To this corpus we thus ask questions such as: How did this breakthrough of new ways of thinking and writing actually unfold? Who were the actors? And to what extent did newness relate to literature at large? Also, the corpus acts as the empirical foundation of an interrelated methodological project, Mining the Meaning, which aims to develop state-of-the-art computational semantic methods and training large language models towards written late 19th-century Danish and Norwegian.[10] Included in the corpus are all original (i.e. newly written) novels by Danish and Norwegian authors published in Denmark 1870-99. The list of texts was compiled on the basis of *Dansk Bogfortegnelse* (a continuous list of books published in Denmark since 1841; from 1861 published annually) supplemented with literary handbooks and special bibliographies. Not included (mainly due to pragmatic reasons and for the sake of coherence) in the corpus are: reprints translations serializations (i.e. serialized novels from newspapers and magazines) diasporic literature (i.e. novels by Danish emigrant authors in the U.S.) Around 20% of the novels are produced by female authors. Thus, highlighting and exploring the often overlooked female literary production of the period is a distinctive ambition of the corpus and the explorations based on it.

5. Language Varieties. The language of the novels in the corpus is late nineteenth century Danish (BCP-47: da). On the whole, we are dealing with a more or less linguistically coherent body of texts. However, the following circumstances must be acknowledged: The texts contain a pronounced spelling variation, partly on an individual level, partly explained by an ongoing orthographic standardization, which is most clearly expressed in the Spelling Reform of 1892. Here, forms such as 'Kjøbenhavn' and 'Familje' became 'København' and 'Familie'. Some books are written in dialect (e.g. Jutlandic or West Norwegian) or contain dialectal features to create psychological individualism in the dialogue. Approximately 16% of the books are written by Norwegian authors. In this regard it should be noted that, until 1907, written Norwegian was practically identical to written Danish. 'Norvagisms' (i.e. distinct Norwegian words, not used by Danes) do appear.

6. Preprocessing and data formatting. OCR scans: The book volumes were scanned with optical character recognition (OCR) by the Royal Danish Library's Digitization on Demand (DoD) team.

[9] https://nors.ku.dk/english/research/projects/measuring-modernity/

[10] https://mime-memo.github.io/

The data were delivered as full volume PDF files with the OCR'ed text as an invisible searchable, copyable text layer, as full volume text files, and as single page text files (one text file per page for each volume). OCR correction: The text files were automatically post-corrected for OCR errors. This involved two different processes, one for texts originally typeset in Antikva (Roman) typefaces, one in Fraktur (Gothic) typefaces. The Antikva files were corrected using a set of hand-crafted substitution patterns, with look-up in the dictionary Sprogteknologisk Ordbase, STO (Eng. 'Word database for language technology'). The Fraktur files were corrected using a correction procedure involving a combination of spelling correction, hand-crafted pattern substitution, and improved OCR using the pretrained 'Fraktur' Tesseract data plus an alternative OCR layer from the pretrained 'dan' Tesseract data, which was used as a corrective to problems with the Danish characters 'æ' and 'ø' in particular. This procedure improved the word error rate of the Fraktur data from 10.46% to 2.84% (cf. Bjerring-Hansen et al., 2022).

Token-level annotation: The corrected data were annotated with grammatical information using the pipeline orchestration tool Text Tonsorium, provided by the Danish CLARIN node.[11] The particular pipeline used included the LaPos part of speech tagger, the CSTLemma lemmatizer, and an implementation of the Brill tagger. Grammatical information included lemma and part of speech, plus sentence and paragraph segmentation (which are of course not strictly speaking token-level annotations). In addition to the grammatical annotations, convenience annotations with various counters were also added: word number in sentence, word number on line, word number in book volume, line number on page, page number in book volume.

Text normalization: After OCR correction, all texts were normalized to modern Danish spelling using hand-crafted substitution patterns and lookup in STO (see above). Nouns were lower cased, 'aa' changed to 'å', and frequent character patterns changed to obey modern Danish orthography. VRT transformation: After annotation with token-level categories and metadata, the data were transformed to a VRT file (vertical format) for indexing in Corpus Workbench (CWB). Format: One token per line delimited by <corpus>, <text>, and <sentence> XML elements. The XML elements contain attributes with metadata. The tokens are annotated with the above-mentioned token-level annotations, separated by tabs. For more information about the metadata, see below. The data are available as: OCR-corrected full volume text files

Normalized full volume versions of these text files A single VRT file containing the whole corpus.

[11] https://https://cst.dk/texton/

7. Limitations. A standard limitation of data preprocessed and annotated using automatic natural language processing tools and procedures is that the results are not perfect. Thus, basically all the layers of the data can be assumed to be flawed: Text data: The raw texts come from OCR scans of the physical book volumes. This process is not perfect, and although we have taken steps to mitigate errors, the basic text layer of the data can still be expected to have OCR errors (or wrong corrections) in 2-3% of tokens. Normalized data: The normalization to modern Danish spelling as such should not be expected to be perfect either. We currently do not have estimates of the error rate in the normalized data. Grammatical annotations: These are also added using automatic tools which cannot be expected to yield perfect results. We currently do not have estimates of error rates in the grammatical annotations. Metadata: The metadata are hand-curated by literary scholars and should be close to perfect. However, the occasional human error can of course not be ruled out.

8. Metadata. The metadata was curated with the help of students (Lasse Stein Holst, Lene Thanning Andersen, and Kirstine Nielsen Degn). on the basis of *Dansk Bogfortegnelse* (1861ff), www.litteraturpriser.dk, Ehrencron-Müller: *Anonym- og Pseudonym-Lexikon* (1940) as well as additional literary and bibliographical handbooks.

Among the metadata categories are the following:

fileid
filename
author firstname
author surname
author pseudonym
author gender
author nationality
title
subtitle
volume
year of publication
pages
illustrations
typeface [gothic/roman]
publisher
price

9. Disclosure and Ethical Review. Funding for the creation and curation is supplied by the Carlsberg Foundation through a Young Researcher Fellowship awarded to Jens Bjerring-Hansen, University of Copenhagen. In terms of data management, the project data (novels from 1870-1900) consist of

imaginative texts by non-living authors. The texts are out-of-copyright. From a GDPR perspective, the biographical, bibliographical and demographic data are historical as well as non-sensitive.