Annotating and Classifying Direct Speech in Historical Danish and Norwegian Literary Texts

Ali Al-Laith[1,2], **Alexander Conroy**[1], **Kirstine Nielsen Degn**[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, knd@hum.ku.dk, alc@hum.ku.dk,
bspedersen@hum.ku.dk, jbh@hum.ku.dk, dh@di.ku.dk

Abstract

Analyzing direct speech in historical literary texts provides insights into character dynamics, narrative style, and discourse patterns. In late 19th century Danish and Norwegian fiction direct speech reflects characters' social and geographical backgrounds. However, inconsistent typographic conventions in Scandinavian literature complicate computational methods for distinguishing direct speech from other narrative elements. To address this, we introduce an annotated dataset from the MeMo corpus, capturing speech markers and tags in Danish and Norwegian novels. We evaluate pre-trained language models for classifying direct speech, with results showing that a Danish Foundation Model (DFM), trained on extensive Danish data, has the highest performance. Finally, we conduct a classifier-assisted quantitative corpus analysis and find a downward trend in the prevalence of speech over time.

1 Introduction

The analysis of direct speech in literary texts provides valuable insights into narrative style, character dynamics as well as aesthetic developments and other broader discourse patterns. In the context of literary history, it has been argued that direct speech, understood as a narrative element that purports to quote a character's speech (Cohen and Green, 2019), is one of the most distinctive components of modern Danish fiction from the late 19th century (Kristensen, 1955). Realist authors of the period use direct speech to reflect characters' social and geographical backgrounds through dialogue rather than explicit description, aiming to portray the fictional world with verisimilitude, i.e. a touch of the real. In Scandinavian literature, typographic marking of speech–such as quotation marks, dashes, and colons–is often inconsistent, complicating the task of distinguishing direct speech from narrative text. This is especially true for Danish and Norwegian novels from the late 19th century, where typographic conventions are highly variable. While readers can often intuitively recognize direct speech, computational approaches require structured annotation to accurately capture these nuanced typographic and linguistic features (Stymne, 2024).

We introduce a newly annotated dataset derived from the MeMo corpus (Bjerring-Hansen et al., 2022), which includes annotations of speech markers, speech tags and speech separated from other narrative elements across Danish and Norwegian novels from the late 19th century. This dataset, annotated on the word level, facilitates the segmentation of direct speech from other narrative elements, enabling sequence tagging model training for the automated detection of these elements. We evaluate several pre-trained language models tailored for Danish and Norwegian, including the Danish Foundation Models (DFM; Enevoldsen et al., 2023) and MeMo-BERT (Al-Laith et al., 2024a), to assess their ability to detect direct speech in historical Scandinavian texts, and find DFM particularly effective. Our findings are of importance to not only literary scholars, but also (socio)linguists who are allowed an indirect access to spoken language from before modern recording technologies (Culpeper and Kytö, 2010).

Our contributions are threefold: (1) we present an annotated dataset that captures the typographic and linguistic indicators of direct speech in 19th century Danish and Norwegian literature, (2) we conduct an empirical evaluation of state-of-the-art language models fine-tuned on this dataset, and (3) we provide insights into the performance and generalization capabilities of these models for classifying direct speech.

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 1–7
March 3-4, 2025

Author	Novel	Type	Example
Kamillo Karstens	Grevinde Danner	German quotation marks	„Læs , “udbrød han
Michael Rosing	En Romantiker	Guillemet-form, Danish	»Kom Jomfru! lad os faa en Dans til Afsked «
Ragnhild Goldschmidt	En Kvindehistorie	Guillemet-form, French	«Laura, Din Kjole er vaad; regner det? »
Herman Bang	Tine	Dash	— Farvel!
Holger Drachmann	Forskrevet	Unmarked	Jeg husker Dem meget godt ! svarede han

Table 1: Excerpts from five novels from the MeMo corpus with different quotation styles.

2 Related Work

Direct speech identification. Identifying direct speech in literary texts has been a focal area in NLP, with various resources and methodologies addressing typographic and linguistic challenges across languages. The Swedish Literary corpus of Narrative and Dialogue (SLäNDa) exemplifies these efforts, providing annotated excerpts from Swedish novels between 1809 and 1940 that capture speech segments, tags, and speaker identification (Stymne and Östman, 2020, 2022). In a similar vein, Troiano and Vossen (2024) introduced CLAUSE-ATLAS, a corpus designed to study narrative structure in 19th and 20th century English novels, leveraging large language models for clause-based annotation.

Recent studies have also explored annotation challenges in texts lacking quotation marks. Stymne (2024) compared manual (gold) and automated (silver) annotation methods, finding that gold data yields better model performance at the token level, while silver data often excels at capturing speech spans. Despite these advancements, most methods rely on monolingual, genre-specific corpora that may not extend well to historical Scandinavian languages.

Historical literary Scandinavian NLP. Our study builds on recent advances in computational approaches for analyzing historical Scandinavian literature, emphasizing the need for tailored datasets and models suited to under-resourced languages. Allaith et al. (2023); Al-Laith et al. (2024b) developed NLP methods specifically adapted to the unique linguistic characteristics of 19th century Danish and Norwegian texts, addressing challenges such as archaic vocabulary, inconsistent orthography, and noisy data. Further studies, such as Feldkamp et al. (2024) and Lindhardt Overgaard et al. (2024), underscore that models and datasets customized for genre-specific nuances enhance the analysis of specialized text types. Bjerring-Hansen et al. (2024) also contributed by distinguishing between contemporary and historical novels, underscoring the value of domain-specific resources for genre classification in historical corpora. These efforts highlight the importance of customized NLP frameworks for advancing computational humanities, particularly in historical Scandinavian literature.

3 Dataset

3.1 Main Corpus

We use the MeMo corpus (Bjerring-Hansen et al., 2022), comprising 859 Danish and Norwegian novels spanning the last 30 years of the 19th century, with more than 64 million tokens. We refer to this corpus as the ‘main corpus’. It should be noted that, until 1907, written Norwegian was practically identical to written Danish (Vikør, 2022).

3.2 Speech Corpus

Segment extraction. We randomly extract 100 segments, each consisting of three consecutive paragraphs, from 100 different novels in the MeMo corpus. For the selection of the target novels, five novels are handpicked by literary experts specifically to represent diverse quotation styles (see Table 1), while the remaining 95 are selected at random, ensuring diverse and comprehensive coverage of quotation styles.

Annotation guidelines. To address the challenges described in §1, we develop clear annotation criteria to ensure consistency and accuracy in identifying speech-related elements:

1. **Speech (“SP”)**: All words and punctuation that are part of direct speech are labeled as “SP”. We do not differentiate embedded speech (e.g., quotations within speech) as both the outer and inner quotations are labeled as “SP”.

2. **Speech Marker (“SM”)**: Any typographical markers indicating speech, such as quotation marks, colons, or dashes, are labeled

as "SM". If a colon appears directly before quotation marks, it is also labelled "SM". For example, in the following:
He shook his head and said: "Certainly, but the stones must be examined first",
both the colon and quotation marks are labeled as "SM".

3. **Speech Tag ("ST")**: Speech tags (or inquit phrases), such as "he said," "she asked," or "they replied," are labeled as "ST". This label applies only to the verb and subject, excluding any adverbs or adverbial phrases, e.g., in
And then he whispered almost inaudibly
only "he whispered" is labeled as "ST". Punctuation immediately preceding or following the tag is also considered part of the "ST" if it is not eligible to be marked as "SM".

4. **Other ("O")**: All other words and punctuation not categorized under the above labels are marked as "O". This includes indirect speech and free indirect discourse. Additionally, inner thoughts and citations from letters or documents are also labelled as "O".

Annotation process. The annotation is carried out on the INCEpTION platform (Klie et al., 2018) by three literary scholars with domain expertise in late 19th century Scandinavian fiction. For agreement calculation and in order to obtain a high-quality testing set, we select 15% of samples for multiple annotation by all three experts. These consist of 15 segments from 15 different novels from the last four years of the period, 1896–1899. In total, they contain 2,530 words. After separate annotation by the three experts, these are consolidated by word-level label majority vote for the final testing set. The rest of the segments in the dataset (75 segments from 75 different novels) are equally split among annotators to be annotated individually.

Annotation results. The annotation results demonstrate a clear prevalence of non-speech elements in the dataset, with a majority of words categorized as "Other". Despite the lower representation of speech-related annotations, the presence of direct speech is still significant, indicating that dialogue plays an important role in the corpus. The minimal occurrences of "Speech Marker" and "Speech Tag" highlight the challenges in identifying these features. This distribution underscores the complexity of the dataset, as a result of diversity in both literary styles and typographical conventions, and the necessity for careful annotation to capture the nuances of speech within the text. Table 2 shows statistics about the manually annotated dataset.

Class	**#Words**	**%**
Speech ("SP")	7,655	32.6%
Speech Marker ("SM")	579	2.5%
Speech Tag ("ST")	363	1.5%
Other ("O")	14,861	63.4%
Total	23,458	100%

Table 2: Distribution of annotated dataset.

Agreement. We use pairwise Cohen's Kappa to assess Inter-Annotator Agreement (IAA) on the subset annotated by all three experts prior to consolidation. The pairwise comparisons between annotators resulted in an average Cohen's Kappa score of 0.92, indicating substantial agreement among annotators in classifying direct speech from other narrative elements.

4 Experiments and Results

We model direct speech identification as token classification, i.e. sequence tagging, with the tags described in §3. We fine-tune and evaluate pre-trained language models for token classification.

4.1 Pre-trained Language Models

We select models pre-trained on Danish and Norwegian text, based on their performance on Danish and Norwegian literary benchmark datasets (Al-Laith et al., 2024a) and ScandEval (Nielsen, 2023). We experiment with models that are not primarily trained on historical/literary Danish or Norwegian. These include DanskBERT and DFM (Large), the Danish Foundation Models sentence encoder, both trained on the Danish Gigaword Corpus; NB-BERT-base, trained on the extensive digital collection at the National Library of Norway; and MeMo-BERT-03, which was developed through continued pre-training of DanskBERT on the MeMo corpus. The following provides an explanation of each model used in this research.

DanskBERT. DanskBERT,[1] a top-performing Danish language model noted for its success on the ScandEval benchmark (Snæbjarnarson et al., 2023), is based on the XLM-RoBERTa architecture and trained on the Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021). It features 24 layers, a hidden dimension of 1024, 16 attention heads, and a subword vocabulary of 250,000. The model was trained with a batch size of 2,000 for 500,000 steps on 16 V100 GPUs over two weeks.

Danish Foundation Models sentence encoder. A sentence-transformers model (Enevoldsen et al., 2023) based on the BERT architecture, featuring 24 layers, 16 attention heads, and a hidden size of 1024. It incorporates a dropout rate of 0.1 for attention probabilities and hidden states, using GELU activation and supporting up to 512 position embeddings. With a vocabulary size of 50,000 tokens, this model, referred to as DFM (Large), excels in some NLP downstream tasks such as sentiment analysis and named entity recognition.[2]

MeMo-BERT-03. Developed by continuing the pre-training of the pre-trained Transformer language model DanskBERT (Al-Laith et al., 2024a).[3] This foundation allows MeMo-BERT-3 to leverage extensive linguistic knowledge for NLP tasks in historical literary Danish including sentiment analysis and word sense disambiguation. The model outperformed different models in sentiment analysis and word sense disambiguation tasks (Al-Laith et al., 2024a).

NB-BERT-base. A general-purpose BERT-base model was developed using the extensive digital collection at the National Library of Norway (Kummervold et al., 2021).[4] It follows the architecture of the BERT Cased multilingual model and has been trained on a diverse range of Norwegian texts, encompassing both Bokmål and Nynorsk from the past 200 years. This comprehensive training allows the NB-BERT-base to effectively handle a wide array of NLP tasks in Norwegian. The model achieved the second-highest performance ranking in the Norwegian Named Entity Recognition task compared to other models listed on the ScandEval benchmark for Norwegian natural language understanding.

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

[2] https://huggingface.co/KennethEnevoldsen/dfm-sentence-encoder-large-exp2-no-lang-align

[3] https://huggingface.co/MiMe-MeMo/MeMo-BERT-03

[4] https://huggingface.co/NbAiLab/nb-bert-base

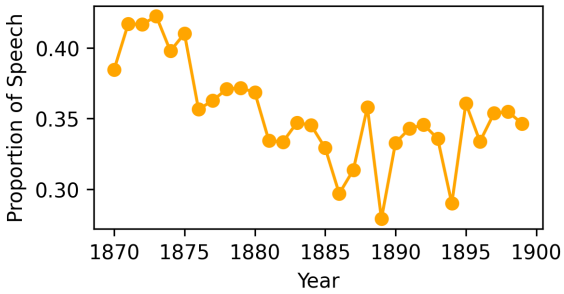

Figure 1: Proportion of speech tokens, predicted by fine-tuned DFM (Large), by publication year.

4.2 Experimental Setup

To fine-tune the models, we use a batch size of 32, and train for 20 epochs with the AdamW optimizer at a learning rate of 10^{-3}, choosing the best epoch based on validation loss. For evaluation, we employ word-level weighted average F1-score. We select for testing the 15% of the dataset annotated by all three experts, and randomly split the rest such that 70% of the overall annotated dataset is used for training and 15% for development.

4.3 Speech Classification Results

Fine-tuning results in notable performance variations, as shown in Table 3. DFM (Large) achieves the best results, indicating strong generalization. NB-BERT-base follows closely, but DanskBERT and MeMo-BERT-03 perform moderately, showing a notable drop from validation to test scores, suggesting less robust generalization. As described in §3.2, the testing set consists of segments from the last four years of the period, while (as described in §4.2) the validation set is randomly sampled from the rest of the period. The testing set therefore represents a time shift from training and is more challenging.

5 Classifier-assisted Corpus Analysis

We use the top-performing model, DFM (Large), to tag all unlabeled segments in the main corpus. This results in 35% of words labeled as speech, 61% as non-speech, 2% as speech markers and 2% as speech tags. Figure 1 shows the proportion of speech and non-speech labels over years,

Model	Validation F1-score	Testing F1-score	Testing Precision	Testing Recall
DanskBERT	0.82	0.71	0.71	0.72
DFM (Large)	**0.94**	**0.89**	0.89	0.90
MeMo-BERT-03	0.81	0.73	0.73	0.74
NB-BERT-base	0.93	0.87	0.87	0.87

Table 3: Fine-tuned models' word-level results on validation and testing sets, of 15 segments each.

illustrating a decreasing trend in the proportion of direct speech over time, with the highest point at 42% in 1874, declining to a low of 29% by 1889.[5] This downward trend stands in contrast to findings from other quantitative studies on direct speech in novels. For example, in the study of British 19th century novels by Menon (2019), the overall fraction of dialogue across her entire corpus compares roughly to ours (36%), but she finds no significant change over time.

Furthermore, our findings challenge the widely held critical assumption within literary historiography that the use of direct speech increased with the rise of the realist novel in the late 19th century, as argued by Kristensen (1955), Allison (2018), and Cohen and Green (2019). Instead, our analysis seems more consistent with the argument presented by Cohn (1978) that the French naturalist aesthetic favored free indirect speech over direct speech, leading to a decline of the latter. This perspective aligns more closely with the downward trend we observe in 19th century Scandinavian literature than with the stable levels of direct speech that Menon (2019) reports in British novels from the same period. In other words, based on these quantitative analyses of direct speech, late 19th century Scandinavian novels appear to align more closely with conventional ideas of naturalist narrative techniques than with those of more conventional realist aesthetics.

6 Conclusion

We presented a dedicated dataset and methodology for annotating direct speech in Danish and Norwegian novels from the late 19th century, useful for not only literary studies but also for linguistics by providing access to representations of 19th spoken language. By building on the MeMo corpus, we systematically annotated typographic markers, speech tags, and direct speech segments, addressing the significant variation and inconsistencies in typographic conventions within historical Scandinavian literature. Through our experiments with multiple language models, including Danish Foundation Models and MeMo-BERT, we found that DFM (Large) performed best. Using it to quantify the proportion of speech in the main corpus, we observed a decreasing trend over time.

Future work will extend our analysis to include other variations of speech, namely indirect discourse, i.e. reporting of character speech, and free indirect discourse, namely the incorporation of a character's speech within the narrator's language (Cohen and Green, 2019). Literary-historical research will examine the lexical variations of the speech tags within the corpus to address a hypothesis (Allison, 2018) that a narrative development from "telling" to "showing" in 19th century literature is manifested in a movement towards greater nuance and lexical variation in the speech tags. While 'telling' is a narrative style, where events are explained explicitly (e.g., 'He was angry'), 'showing' uses a more detailed narrative style to implicitly convey what is at stake in the event, as in 'He slammed his fist on the table and shouted "Enough"'. The hypothesis is that this shift is reflected in more nuanced speech tags, moving from simple terms like 'said' to varied ones like 'muttered' or 'snarled.' Our code and data are in this Github repository: `https://github.com/mime-memo/DirectSpeech`.

[5] These numbers may not be perfectly accurate as they are a result of an accurate-but-not-perfect classifier, as shown in §4.3. Moreover, they may be more reliable for some years than for others, but we are unable to quantify this with our current dataset, since our testing set consists only of segments from 1896 to 1899.

References

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024a. Development and evaluation of pre-trained language models for historical Danish and Norwegian literary texts. In

Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 4811–4819, Torino, Italia. ELRA and ICCL.

Ali Al-Laith, Daniel Hershcovich, Jens Bjerring-Hansen, Jakob Ingemann Parby, Alexander Conroy, and Timothy R Tangherlini. 2024b. Noise, novels, numbers. a framework for detecting and categorizing noise in Danish and Norwegian literature. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

Sarah Allison. 2018. *Reductive reading: A syntax of Victorian Moralizing* . Johns Hopkins University Press, Baltimore, Maryland.

Jens Bjerring-Hansen, Ali Al-Laith, Daniel Hershcovich, Alexander Conroy, and Sebastian Ørtoft Rasmussen. 2024. Literary time travel: Distinguishing past and contemporary worlds in Danish and Norwegian fiction. In *Computational Humanities Research 2024*.

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

William A. Cohen and Laura Green. 2019. Introduction: Revisiting dialogue. *Narrative*, 27(2):1013–1019.

Dorrit Cohn. 1978. *Transparent Minds: Narrative Modes for Presenting Consciousness in Fiction.* Princeton University Press, Princeton, United States.

Jonathan Culpeper and Merja Kytö. 2010. *Early Modern English dialogues: Spoken interaction as writing*. Cambridge University Press, Cambridge, England.

Kenneth Enevoldsen, Lasse Hansen, Dan S. Nielsen, Rasmus A. F. Egebæk, Søren V. Holm, Martin C. Nielsen, Martin Bernstorff, Rasmus Larsen, Peter B. Jørgensen, Malte Højmark-Bertelsen, Peter B. Vahlstrup, Per Møldrup-Dalum, and Kristoffer Nielbo. 2023. Danish foundation models.

Pascale Feldkamp, Jan Kostkan, Ea Overgaard, Mia Jacobsen, and Yuri Bizzoni. 2024. Comparing tools for sentiment analysis of Danish literature from hymns to fairy tales: Low-resource language and domain challenges. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 186–199, Bangkok, Thailand. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Sven Møller Kristensen. 1955. *Impressionismen i dansk prosa 1870-1900*. Gyldendal, Copenhagen, DK.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Ea Lindhardt Overgaard, Pascale Feldkamp, and Yuri Bizzoni. 2024. Towards a GoldenHymns dataset for studying diachronic trends in 19th century Danish religious hymns. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 55–61, Bangkok, Thailand. Association for Computational Linguistics.

Tara Menon. 2019. Keeping count: Direct speech in the nineteenth-century british novel. *Narrative*, 27(2):160–181.

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.

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.

Sara Stymne. 2024. Direct speech identification in Swedish literature and an exploration of training data type, typographical markers, and evaluation granularity. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics*

for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), pages 253–263, St. Julians, Malta. Association for Computational Linguistics.

Sara Stymne and Carin Östman. 2020. SLäNDa: An annotated corpus of narrative and dialogue in Swedish literary fiction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 826–834, Marseille, France. European Language Resources Association.

Sara Stymne and Carin Östman. 2022. SLäNDa version 2.0: Improved and extended annotation of narrative and dialogue in Swedish literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5324–5333, Marseille, France. European Language Resources Association.

Enrica Troiano and Piek T.J.M. Vossen. 2024. CLAUSE-ATLAS: A corpus of narrative information to scale up computational literary analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3283–3296, Torino, Italia. ELRA and ICCL.

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