Speaking on Their Behalf: Detecting Indirect Speech in Historical Danish and Norwegian Texts

Anonymous ACL submission

Abstract

Indirect speech is a fundamental yet understudied form of reported speech that plays a crucial role in literary texts and communication. While direct speech detection has received significant attention in computational linguistics, the automatic identification of indirect speech remains a challenge due to its nuanced linguistic structure and contextual dependencies. This paper focuses on the detection of indirect speech in late 19th-century Scandinavian literature, where its presence has been linked to shifting aesthetic ideals. We present an annotated dataset of 150 segments, each randomly selected from 150 different novels, designed to capture indirect speech in Danish and Norwegian literature. We evaluate four pre-trained language models for classifying indirect speech, with results showing that a Danish Foundation Model (DFM Large), trained on extensive Danish data, has the highest performance. Finally, we conduct a classifier-assisted quantitative corpus analysis and find that the prevalence of indirect speech exhibits fluctuations over time.

1 Introduction

The way speech is rendered in writing shapes everything from how we interpret literary texts to everyday communication. Reported speech has been described as essential to human society, with direct speech being a universal feature and non-direct constructions also highly frequent (Goddard and Wierzbicka, 2018). Consequently, automatic detection of speech in written text is a fundamental challenge in linguistic analysis and has applications in various fields, including epidemiology (Klein et al., 2020), communication studies, and journalism (Newell et al., 2018). This paper focuses on a particular non-direct construction: indirect speech. Indirect speech is a way of reporting the utterance of someone else, typically without quoting it verbatim and with adjustments to verb tense, pronouns, and adverbials to reflect the reporter's perspective (Aarts, 2014). While direct speech identification has received significant computational attention, indirect speech remains comparatively understudied. Our empirical focus is on Scandinavian literature from the late 19th century, where indirect speech has been analyzed only for a limited number of authors (Brix, 1911). Moreover, it has been argued that the presence of indirect speech conflicts with certain aesthetic ideals of the time (Kristensen, 1955), making its automatic detection a valuable tool for reexamining Scandinavian literary history.

2 Related Work

Indirect speech is common in both spoken and written language, shaping how we interpret the content, connotations, and reliability of an utterance. Linguistic and psychological research highlights that the choice between indirect and direct speech significantly affects how we perceive, recall, and process reported statements (Eerland and Zwaan, 2018). However, distinguishing indirect speech from related phenomena is challenging in both spoken and written form. As a result, we rely on contextual cues such as pronouns, verb tense, discourse particles, exclamatives, and emotives (Eckardt, 2020). This complexity requires careful annotation to produce well-performing models.

Although computational research in this area remains limited, some studies have explored related approaches. Krestel et al. (2008) introduced a Reported Speech Tagger for the GATE framework, demonstrating an effective approach to automatically annotating reported speech in newspaper articles. Similarly, Asr et al. (2021) has successfully measured reported speech in the news media as part of its investigation into the gender representation gap. However, both studies classify all reported speech instances without distinguishing between direct and indirect speech. Pareti et al. (2013) conducted the first large-scale study on indirect speech

and mixed quotation extraction. Their findings indicated that traditional machine learning methods, such as the Maximum Entropy Classifier and Conditional Random Fields, were less effective in predicting indirect quotations compared to direct ones. Furthermore, Kathirgamalingam et al. (2023) evaluated three off-the-shelf tools — CoreNLP (Manning et al., 2014), QSample (Scheible et al., 2016), and rsyntax (Welbers et al., 2021) — across two data sources: news articles and social media communication. Their results aligned with previous research, confirming that indirect speech is more challenging to detect automatically than direct speech. Regarding literary studies specifically, Muzny et al. (2017) developed a deterministic sieve-based system for quote attribution, which effectively classifies their three example novels. However, the focus is primarily on who is speaking rather than how the speech is reported. Brunner et al. (2020) analyzed a corpus of German fictional and non-fictional texts from the 19th century and the early 20th century, demonstrating that BERT-based models outperformed models trained within the Flair framework in detecting indirect speech. In Scandinavian Studies, computational research has so far focused exclusively on direct speech, as seen in studies such as Stymne (2024) and Al-Laith et al. (2025). This paper is therefore the first to examine indirect speech in Scandinavian literary history.

3 Dataset

3.1 Main Corpus

We use the MeMo corpus (Bjerring-Hansen et al., 2022), consisting of 859 Danish and Norwegian novels (64M+ tokens) from the last 30 years of the 19th century.[1] 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. To address the low frequency of indirect speech in our main corpus, we use a linguistically informed regular expression targeting communication verbs followed by a complementizer as a seed pattern to extract candidate passages (Appendix A). This method ensures sufficient positive examples. From 150 randomly selected novels, we retrieve three consecutive paragraphs surrounding a randomly selected seed pattern match.

Annotation guidelines. To address the challenges described in §2, we develop clear annotation criteria to ensure consistency and accuracy in identifying speech-related elements:

1. **Indirect Speech ("IS")**: All words and punctuation that are part of indirect speech are labeled as "IS". We do not differentiate embedded speech (e.g., quotations within speech) within passages of indirect speech. We understand indirect speech as a way of reporting speech by using an introductory report verb (e.g. say, ask, tell) and a subordinate clause, for example: *"Anna asked if Kramer could speak with her"* or *"Jørgen suggested that they should leave."* Contrary to direct speech, which repeats the used words verbatim, indirect speech typically involves changes to the original speaker's words, such as adjustments of pronouns, time and place adverbials, and verb tenses to reflect the perspective of the reporter (Aarts, 2014).

2. **Direct Speech ("DS")**: All words and punctuation that are part of direct speech are labeled as "DS". We again do not differentiate embedded speech (e.g., quotations within speech) as both the outer and inner quotations are labeled as "DS".

3. **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".

4. **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 phrases 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 within the same sentence is also considered part of the "ST" if it is not eligible to be marked as "SM".

5. **Other ("O")**: All other words and punctuation not categorized under the above labels

[1] Released with Creative Commons Attribution 4.0 license: https://huggingface.co/datasets/MiMe-MeMo/Corpus-v1.1.

Class	#Words	%
Indirect Speech (“IS”)	537	1.70%
Direct Speech (“DS”)	14,010	44.17%
Other (“O”)	14,962	47.19%
Speech Marker (“SM”)	1,083	3.42%
Speech Tag (“ST”)	1,115	3.52%
Total	31,707	100%

Table 1: Distribution of annotated dataset.

are marked as “O”. This includes free indirect discourse. Additionally, inner thoughts and citations from letters or documents are also labelled as “O”.

Annotation process. The annotation is conducted on the INCEpTION platform (Klie et al., 2018) by three scholars with domain expertise in late 19th century Scandinavian literature. The annotation is done on a token level. For agreement calculation and in order to obtain a high-quality testing set, we select 20% of samples for multiple annotation by all three experts. These consist of 30 random segments from each year.

Annotation results. Annotation results show that most words fall under "Other" (47.19%), while direct speech ("DS") accounts for 44.17%, highlighting the prominence of dialogue. However, due to our extraction method—using a regular expression to target communication verbs—DS is likely overrepresented compared to its actual share in the main corpus, previously measured at 35% (Al-Laith et al., 2025). Indirect speech is rare (1.70%), while "Speech Marker" ("SM") and "Speech Tag" ("ST") are unsurprisingly low (3.42% and 3.52%), given their dependence on speech and minimal token length. This distribution reflects the dataset's complexity, shaped by diverse literary styles and typographical conventions, underscoring the need for precise annotation. Table 1 provides detailed statistics on the manually 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 0.88, indicating substantial agreement among annotators in classifying indirect speech from other representations of speech and narrative elements.

4 Experiment and Results

We model indirect speech detection 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., 2024) and ScandEval (Nielsen, 2023). We experiment with both models not trained primarily on *historical/literary* Danish or Norwegian: DanskBERT (Snæbjarnarson et al., 2023)[2] and DFM (Large), the Danish Foundation Models sentence encoder (Enevoldsen et al., 2023),[3] both trained on the Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021); and NB-BERT-base (Kummervold et al., 2021),[4] trained on the extensive digital collection at the National Library of Norway. Finally, MeMo-BERT-03 (Al-Laith et al., 2024),[5] developed by continued pre-training of DanskBERT on the MeMo corpus.

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 20% of the dataset annotated by all three experts, and randomly split the rest such that 66% of the overall annotated dataset is used for training and 14% for development.

4.3 Classification Results

Fine-tuning results in notable performance variations, as shown in Table 2. 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.

[2] https://huggingface.co/vesteinn/DanskBERT
[3] https://huggingface.co/KennethEnevoldsen/dfm-sentence-encoder-large-exp2-no-lang-align
[4] https://huggingface.co/NbAiLab/nb-bert-base
[5] https://huggingface.co/MiMe-MeMo/MeMo-BERT-03

Model	Validation	Testing
DanskBERT	0.65	0.66
DFM (Large)	**0.93**	**0.97**
MeMo-BERT-03	0.65	0.66
NB-BERT-base	0.85	0.88

Table 2: Fine-tuned models' word-level F1-score results on validation and testing sets, of 21 and 30 segments respectively.

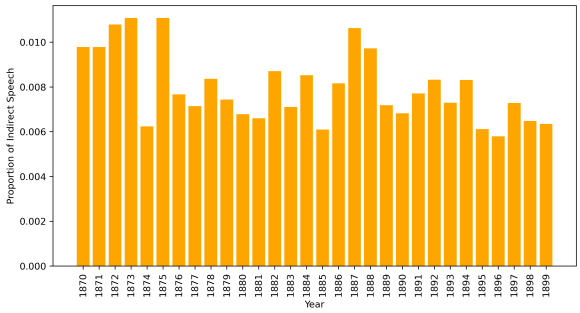

Figure 1: Proportion of indirect speech tokens, predicted by fine-tuned DFM (Large), by publication year.

The classification results indicate strong performance for most tags in the testing set, with Speech Marker (SM), Direct Speech (DS), Speech Tag (ST), and Indirect Speech (IS) achieving high F1-scores above 0.94, suggesting excellent model precision and recall for these categories. However, the Other (O) category has a significantly lower F1-score (0.52), indicating difficulty in distinguishing this class, possibly due to class imbalance or overlapping features with other categories. Overall, the model performs well in identifying speech-related tags but struggles with the broader "Other" category.

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 37.77% of words labeled as Direct Speech (DS), 0.79% as Indirect Speech (IS), 56.51% as Other (O), 2.54% as Speech Marker (SM), and 2.38% as Speech Tag (ST). Figure 1 shows the proportion of indirect speech label over time from 1870 to 1899. The trend appears to be fluctuating rather than showing a consistent increase or decrease. While no clear temporal pattern emerges, indirect speech usage appears linked to the social status and aesthetic position of authors. The 20 works with the highest proportion of indirect speech (7.4%–2.5%) come from non-canonized or lesser-known authors in popular genres like crime fiction and historical novels. In contrast, the 20 works with the lowest proportion (0.0%–0.1%) are by canonized authors such as Viggo Stuckenberg, Johannes Jørgensen, Holger Drachmann, and Jonas Lie. This pattern is further reinforced when examining the 'Other' category ("O"). Among the works with the highest percentage in this category—ranging from 91.9% to 83.4%, well above the corpus average of 56.51%—male canonized authors dominate, including Karl Gjellerup, Jonas Lie, Johannes Jørgensen, Herman Bang, Henrik Pontoppidan, and Edvard Brandes. Our results suggest that canonized authors favored other narrative techniques than indirect and direct speech—perhaps using other ways of representing speech (e.g., free indirect speech) or focusing primarily on representing other types of events such as actions, thoughts, and sensations. These questions will need further examination.

6 Conclusion

In this study, we explored the detection of indirect speech in late 19th-century Danish and Norwegian literature, an understudied aspect of reported speech with significant implications for linguistic and literary analysis. Our work introduces a new annotated dataset and evaluates multiple pre-trained language models for indirect speech classification. The results highlight the superior performance of the Danish Foundation Model (DFM Large), suggesting that domain-specific linguistic resources enhance the accuracy of the model in historical Scandinavian texts.

Beyond technical advancements, our findings reinforce the argument that indirect speech patterns reflect broader aesthetic and literary shifts, particularly in the Scandinavian literary tradition. By allowing for systematic study of these patterns, our approach provides a new computational lens for examining historical discourse. Future work should expand on this foundation by incorporating additional linguistic features, refining annotation strategies, and extending the analyses to other genres and languages. Ultimately, this research underscores the importance of computational methods in uncovering nuanced linguistic phenomena and advancing literary studies.

Limitations

This study presents several limitations that should be acknowledged. First, the annotated dataset is relatively small, consisting of only 150 segments drawn from 150 different novels. While this sampling strategy ensures literary diversity, it limits the robustness of training data, particularly for rare phenomena like indirect speech. Second, our extraction method, based on regular expressions targeting communication verbs and complementizers, likely introduces selection bias and overrepresents certain syntactic constructions of reported speech. Third, while we achieved high inter-annotator agreement, the inherent ambiguity of indirect speech, especially in cases involving free indirect discourse, remains a source of uncertainty both for annotators and models. Fourth, our experiments focused on a limited set of Danish and Norwegian language models. Although we selected state-of-the-art models suited to the task, we did not explore cross-lingual transfer or few-shot prompting strategies. Lastly, the classifier-assisted corpus analysis assumes consistent performance across time and text types, which may not hold due to evolving orthographic conventions, genre-specific styles, and shifting linguistic norms during the late 19th century. These limitations open avenues for future work, including expansion of the dataset, improved sampling strategies, and more nuanced modeling of temporal and stylistic variation.

References

Bas Aarts. 2014. Indirect speech. In *The Oxford Dictionary of English Grammar*, 2nd edition. Oxford University Press.

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024. 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, Alexander Conroy, Kirstine Nielsen Degn, Jens Bjerring-Hansen, and Daniel Hershcovich. 2025. Annotating and classifying direct speech in historical danish and norwegian literary texts. In *Proceedings of NoDaLiDa/Baltic-HLT 2025*.

Fatemeh Torabi Asr, Mazraeh Mohammad, Alexandre Lopes, Vagrant Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The gender gap tracker: Using natural language processing to measure gender bias in media. *PLoS One*, 16(1).

Jens Bjerring-Hansen, Ross Deans Kristensen-McLachlan, Philip Diderichsen, and Dorte Haltrup Hansen. 2022. Mending fractured texts. a heuristic procedure for correcting ocr data. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference*, volume 3232, pages 177–186, Uppsala, Sweden. DHNB Proceedings.

Hans Brix. 1911. *Gudernes Tungemaal*. Gyldendal, Copenhagen, DK.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2020. To bert or not to bert–comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing representation. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*.

R. Eckardt. 2020. The parameters of indirect speech. pages 1–25.

Anita Eerland and Rolf A. Zwaan. 2018. The influence of direct and indirect speech on source memory. *Collabra: Psychology*, 4(1):5.

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. *Preprint*, arXiv:2311.07264.

Cliff Goddard and Anna Wierzbicka. 2018. Direct and indirect speech revisited: Semantic universals and semantic diversity. In Alessandro Capone, Manuel García-Carpintero, and Alessandra Falzone, editors, *Indirect Reports and Pragmatics in the World Languages. Perspectives in Pragmatics, Philosophy & Psychology, vol 19*, pages 173–199. Springer, Cham.

Ahrabhi Kathirgamalingam, Fabienne Lind, and Hajo G. Boomgaarden. 2023. Automated detection of voice in news text – evaluating tools for reported speech and speaker recognition. *Computational Communication Research*, 5(1):85.

Ari Z. Klein, Haitao Cai, Davy Weissenbacher, Lisa D. Levine, and Graciela Gonzalez-Hernandez. 2020. A natural language processing pipeline to advance the use of twitter data for digital epidemiology of adverse pregnancy outcomes. *Journal of Biomedical Informatics*, 112:100076. Articles initially published in Journal of Biomedical Informatics: X 5-8, 2020.

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.

Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In *Proceedings of the International Language Resources and Evaluation Conference*, LREC, pages 2823–2828. European Language Resources Association (ELRA).

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.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain. Association for Computational Linguistics.

Chris Newell, Tim Cowlishaw, and David Man. 2018. Quote extraction and analysis for news. In *Proceedings of KDD Workshop on Data Science, Journalism and Media (DSJM)*.

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.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, Washington, USA. Association for Computational Linguistics.

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model architectures for quotation detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745, Berlin, Germany. 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.

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.

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

Kasper Welbers, Wouter van Atteveldt, and Jan Kleinnijenhuis. 2021. Extracting semantic relations using syntax. *Computational Communication Research*, 3(2):180–194.

A Regular expression

```
# Regex:
[word != "˙"]*
[word = "(sige|fortælle|spørge|påstå|tro)r|
(sagde|fortalte|spurgte|påstod|nævnede|troede)|
(svare|indrømme|bemærke|forklare|understrege|tilføje
bekræfte|erklære|anmode|hævde|advare)(r|de)|
(men|nævn|forlang|råb)(er|te)"]
[]0,12 [word = ","]0,1
[word = "at|(hvem|hvad|hvilke|hvorledes|hvor|
hvornår|hvordan|hvorfor)"]
[word != "˙"]* [word = "˙"]
```