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

Ali Al-Laith, Alexander Conroy, Kirstine Nielsen Degn, Jens Bjerring-Hansen and **Daniel Hershcovich**

NoDaLiDa/Baltic-HLT
3 March 2025

UNIVERSITY OF COPENHAGEN

Context: the MiMe/MeMo Projects

Mining the Meaning & Measuring Modernity:

Literary and Social Change in Scandinavia 1870-1900

Jens Bjerring-Hansen

Associate Professor & Co-PI

Nordic Studies and Linguistics Department, KU

Daniel Hershcovich

Assistant Professor & Co-PI

Department of Computer Science, KU

Bolette Sandford Pedersen

Professor

Nordic Studies and Linguistics Department, KU

Ali Al-Laith

Postdoc

Nordic Studies and Linguistics & Computer Science Departments, KU

Alexander Conroy

PhD Student

Nordic Studies and Linguistics Department, KU

Kirstine Nielsen Degn

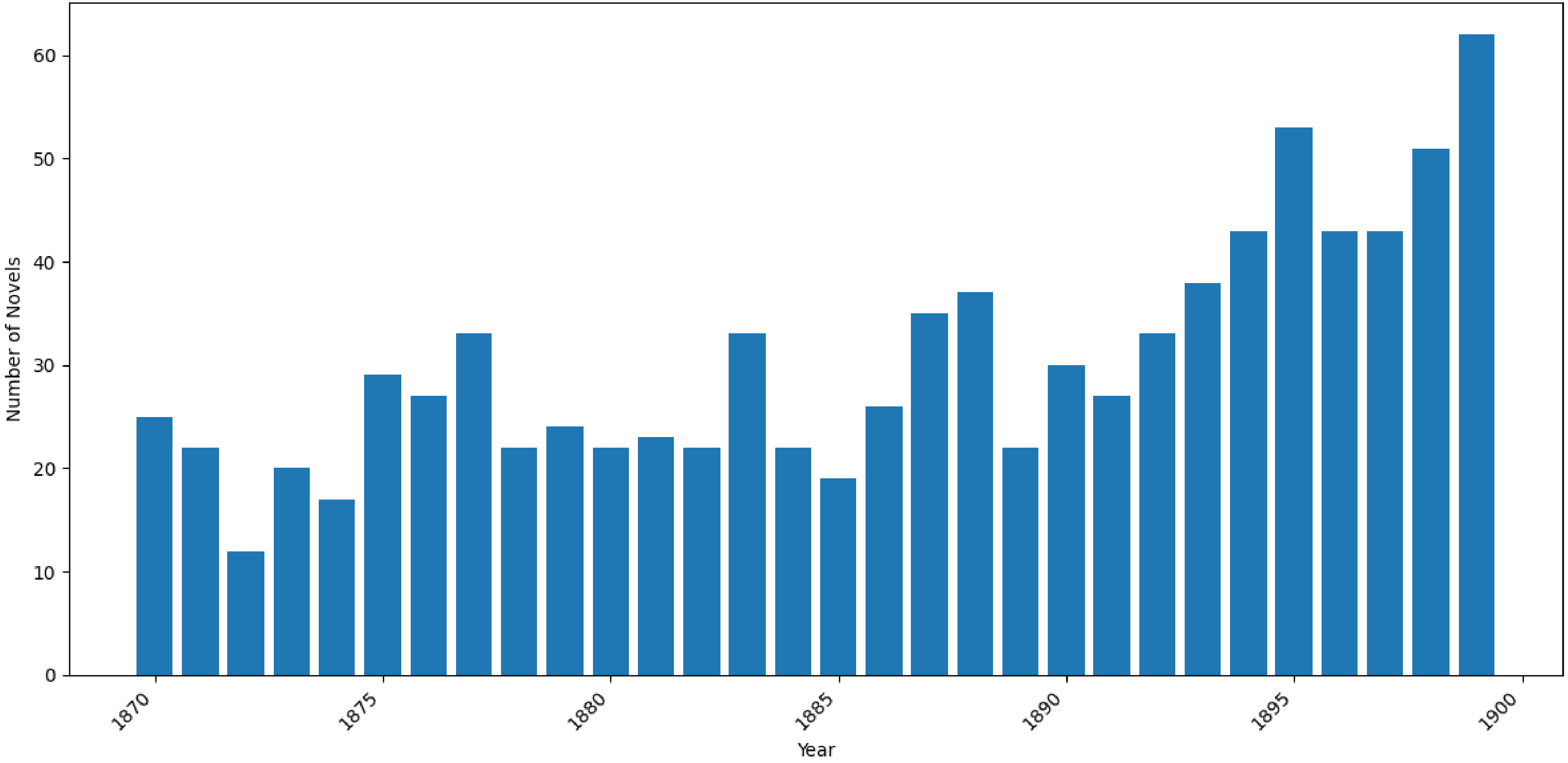
PhD Student

Nordic Studies and Linguistics Department, KU

Approach: Combining computational methods with historical expertise to derive insights into modernization.

MeMo Corpus

- 900 Danish and Norwegian novels published between 1870-1900



Other MiMe/MeMo Sub-Projects

Sentiment Classification of Historical Danish and Norwegian Literary Texts. **NoDaLiDa 2023**

Development and Evaluation of Pre-trained Language Models for Historical Danish and Norwegian Literary Texts. **LREC-COLING 2024**

Noise, Novels, Numbers. A Framework for Detecting and Categorizing Noise in Danish and Norwegian Literature. **EMNLP 2024**

Literary Time Travel: Distinguishing Past and Contemporary Worlds in Danish and Norwegian Fiction. **CHR 2024**

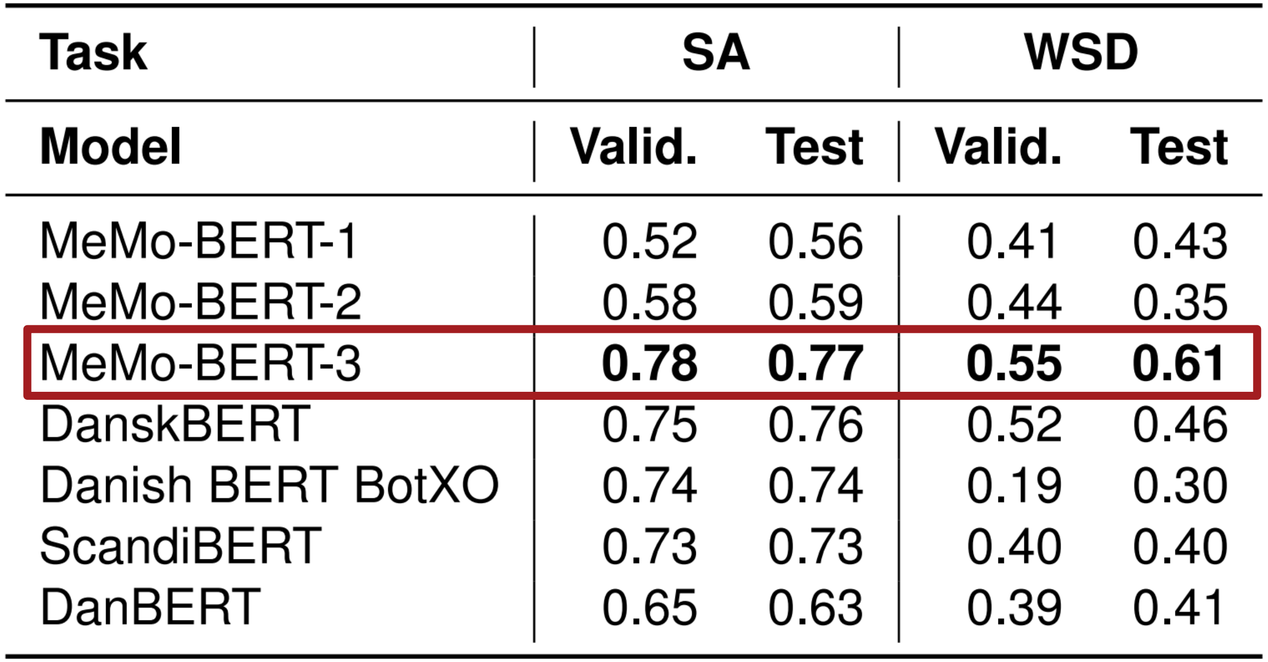
Dying or Departing? Euphemism Detection for Death Discourse in Historical Texts. **COLING 2025**

Unhappy Texts? A Gendered and Computational Rereading of The Modern Breakthrough. **Scandinavian Studies, 2025**

MeMo-BERT

Development and Evaluation of Pre-trained Language Models for Historical Danish and Norwegian Literary Texts. **LREC-COLING 2024**

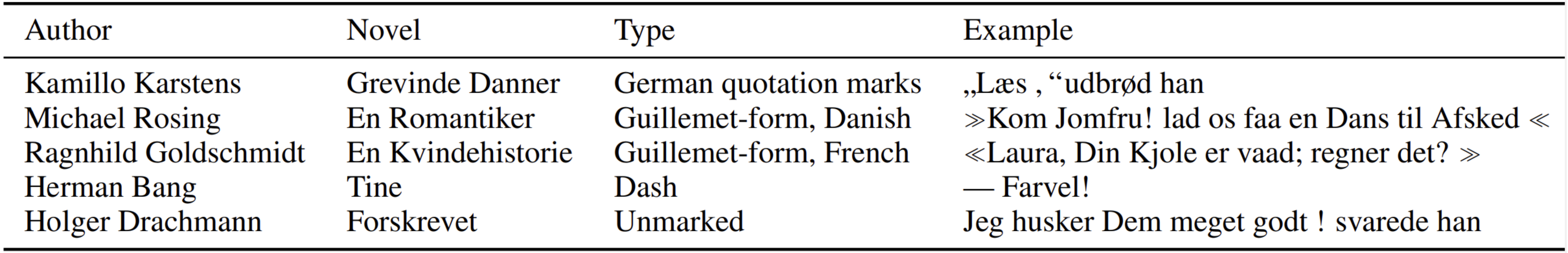
The first BERT-based models for historical literary Danish and Norwegian texts.
Outperformed other models in sentiment analysis and word sense disambiguation.



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

Direct Speech

A narrative element that purports to quote a character’s speech



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

Substantial variation in typographic conventions

Speech-related Elements

- Speech ("**SP**"): direct speech itself
- Speech Marker ("**SM**"): typographical markers of speech
- Speech Tag ("**ST**"): lexical markers of speech, verb+subject (inquit phrases)
- Other ("**O**"): indirect speech, free indirect discourse, anything else

Example:

Her stood hun lidt stille og snusede Luft til sig, idet hun sagde: Jeg kan saa godt lide Skomagerlugt.

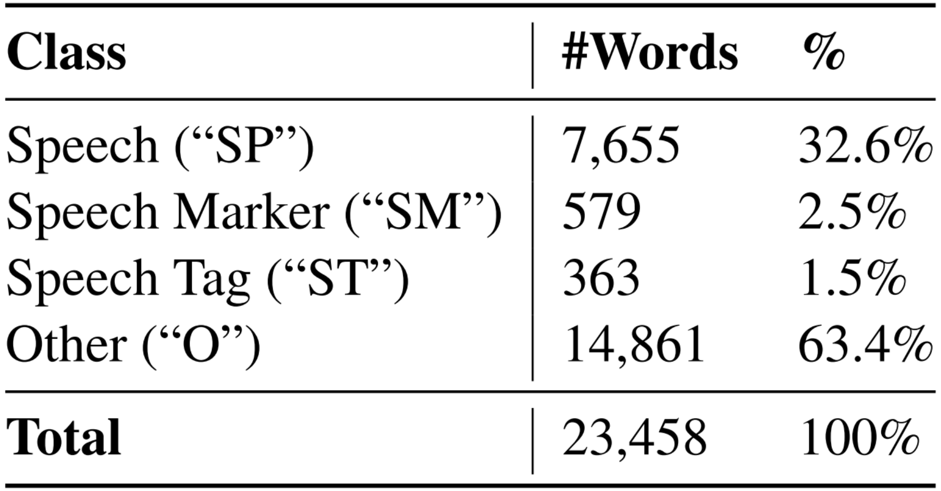
("Here she stood a little still and sniffed the air as she said: I like the smell of a shoemaker so much.")

Annotated Dataset

100 text segments from different novels.
Token-level annotation: 3 literary scholars.

IAA:

Average pairwise Cohen's Kappa: 0.92

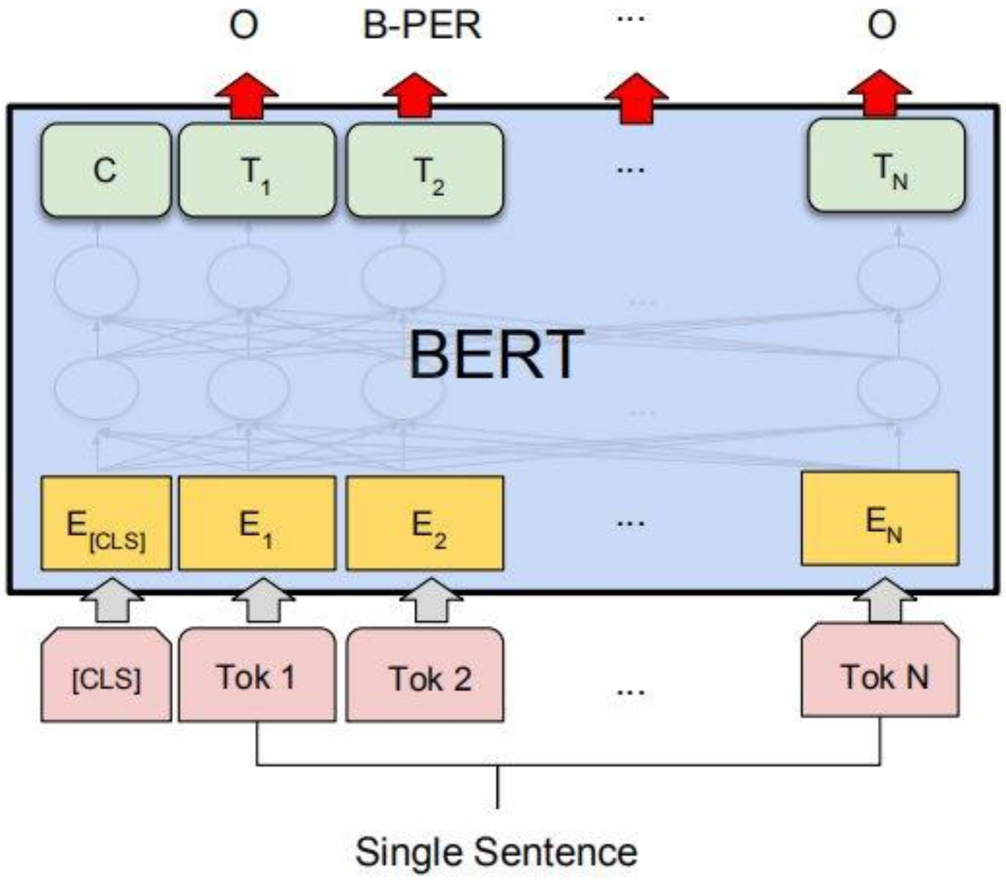


Class	#Words	%
Speech (“SP”)	7,655	32.6%
Speech Marker (“SM”)	579	2.5%
Speech Tag (“ST”)	363	1.5%
Other (“O”)	14,861	63.4%
Total	23,458	100%

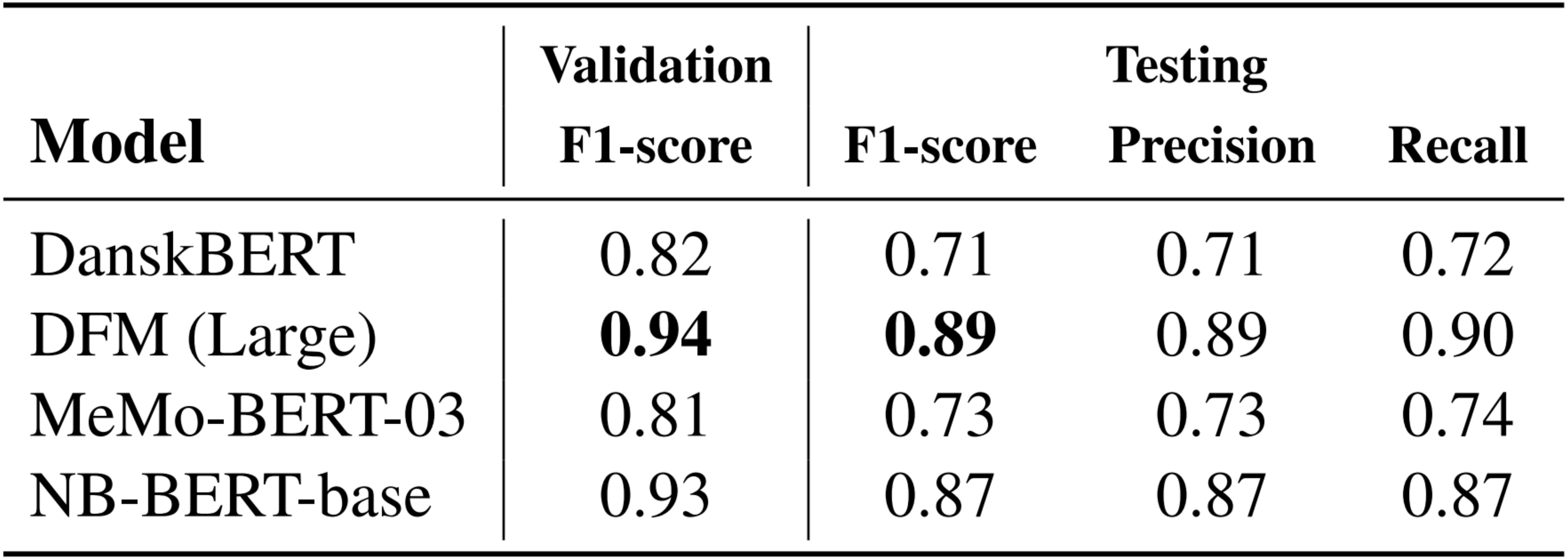
Experiments

- Approach: sequence tagging (token classification)

- Fine-tuning and evaluating pre-trained language models:
  - DanskBERT
  - DFM (Danish Foundation Models)
  - MeMo-BERT-03
  - NB-BERT-base



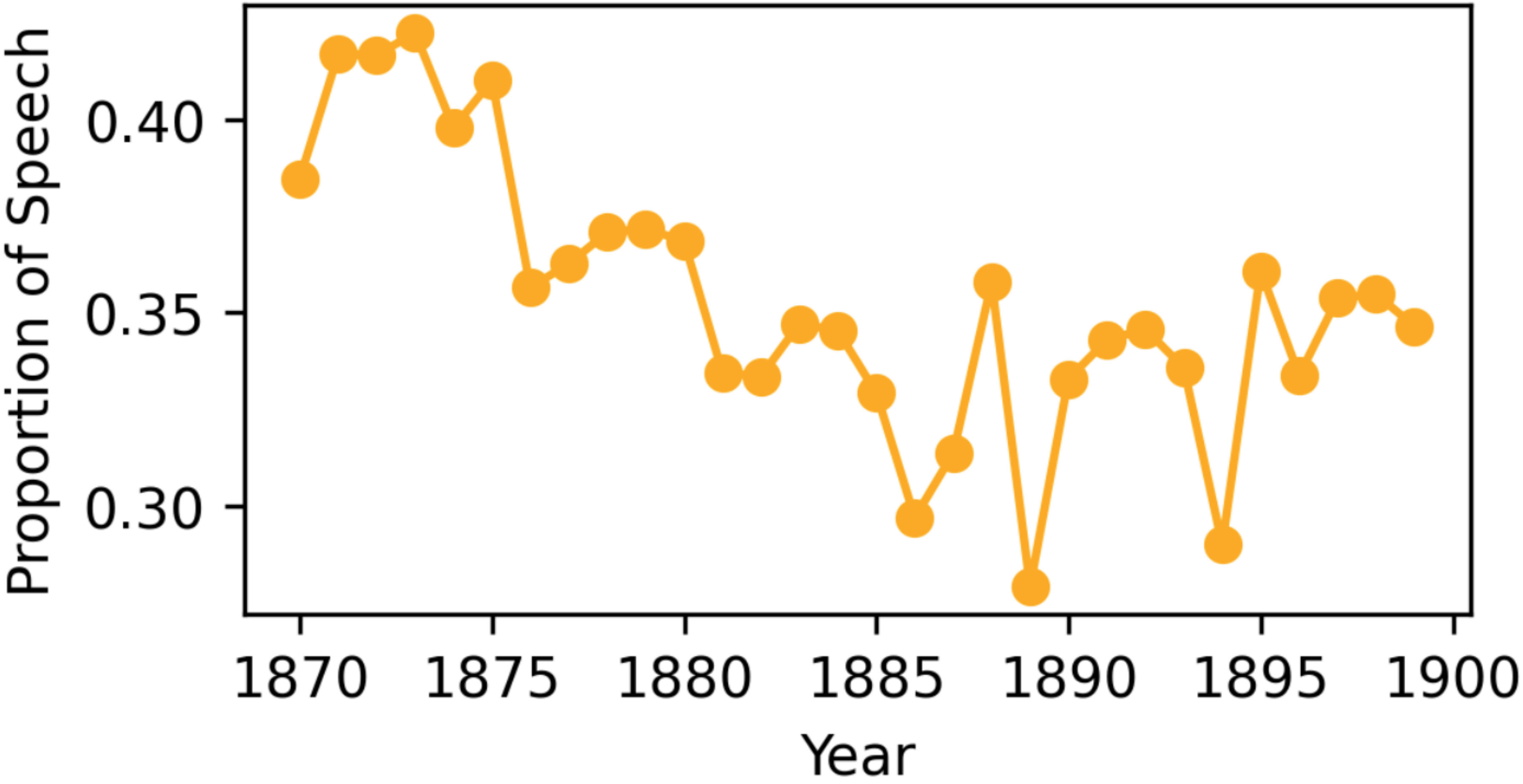
Results



	Validation	Testing		
Model	**F1-score**	**F1-score**	**Precision**	**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

Danish Foundation Models substantially outperform other models!

Classifier-assisted Corpus Analysis



Decline in the prevalence of direct speech over time

Significance

- Realist authors of the period use direct speech to reflect characters' social and geographical backgrounds through dialogue rather than explicit description. Our models enable deep analysis of these dialogues.
- Hypothesis for future work: 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.

Conclusion

Investigating direct speech in 19th-century Danish and Norwegian fiction

We developed an annotated dataset and evaluated various models

The analysis reveals a decline in the prevalence of direct speech over time

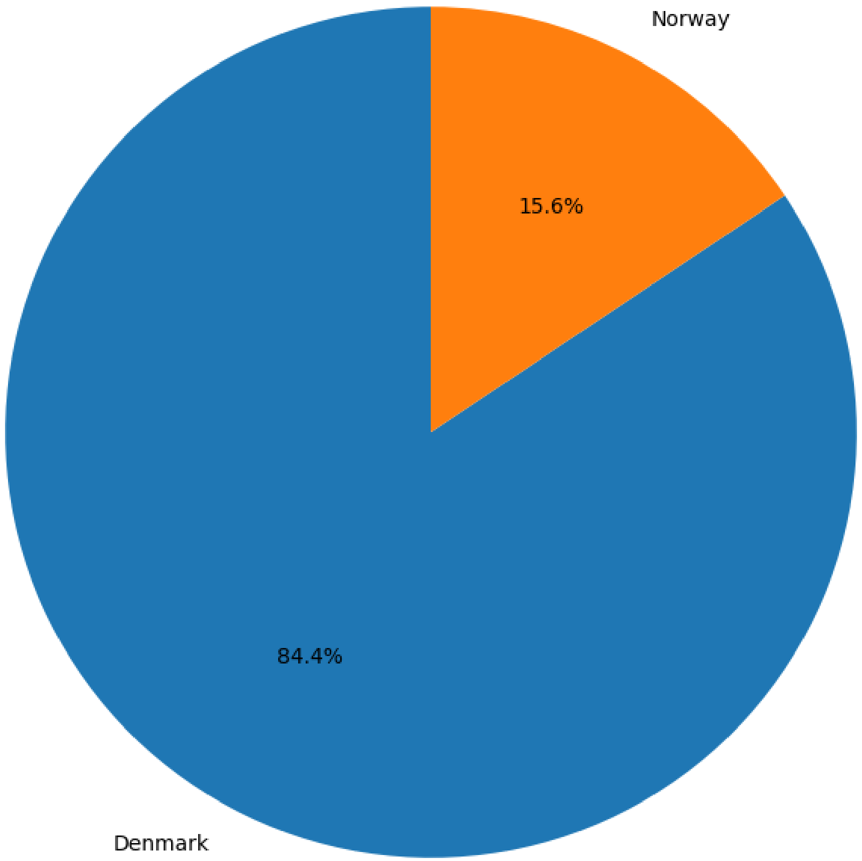
🌐 danielhers.github.io

✉ dh@di.ku.dk

Slides credits: Ali Al-Laith

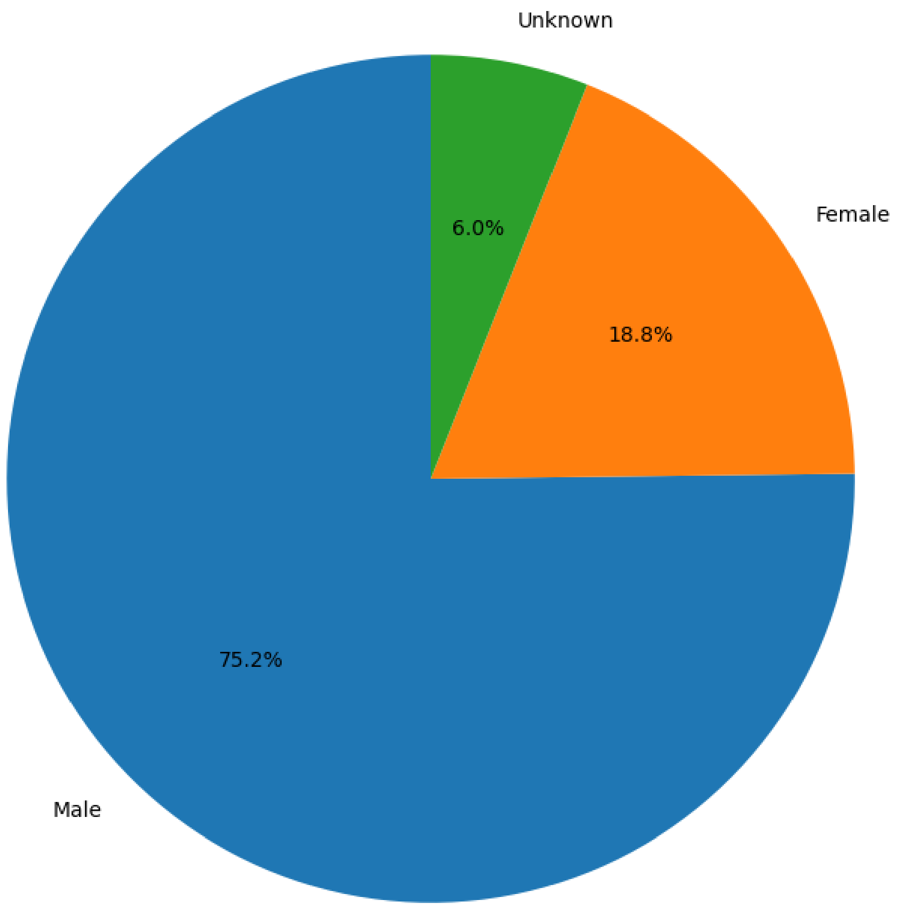
MeMo Corpus

- **Distribution of Novels over Nationalities:**



MeMo Corpus

- **Distribution of Novels over Gender:**


Unknown
Female
Male