Direct Speech Identification in Swedish Literature without Graphical Speech Marking

Sara Stymne
Department of Linguistics and Philology
Uppsala University
sara.stymne@lingfil.uu.se

Abstract

Identifying direct speech in literary fiction is challenging for cases that do not mark speech segments graphically. Such efforts have previously been based either on smaller manually annotated gold data or larger automatically annotated silver data, extracted from works with quotation marks. However, no direct comparison has so far been made between the performance of these two types of training data. In this work, we address this gap. We further explore the effect of different types of typographical speech marking in the training data. Generally, we see stronger performance with small gold data, than with considerably larger silver data for training. If the training data contains some data that matches the typographical speech marking of the target, that is generally sufficient for achieving good results, but it does not seem to hurt if the training data also contains other types of marking.[1]

1 Introduction

In research on literary text, it is often useful to be able to distinguish between the frame narrative and character speech (Elson et al., 2010; Nalisnick and Baird, 2013) and to identify speech tags (Allison, 2018). Identifying speech and speech tags is relatively easy in the presence of quotation marks, but much more challenging when speech is marked by a dash at the beginning, or not at all, as in (1). In this work, we are interested in the latter case, targeting data sets without any speech marking, and with dashes automatically stripped from the text.

(1) Ja, lefvande! ropade patron Lack. Lefver inte svinet?
'Yes, alive! Patron Lack shouted. Isn't the pig alive? '
(Nordström, *Borgare*, p. 20)

[1]This work is based on work previously presented in Stymne (2024), which contains additional experiments. See also Section 6.

In most previous work on direct speech identification, training data has either been gold data, annotated by humans, or silver data, automatically extracted from texts with quotation marks. While gold data typically has higher quality than silver data, silver data can easily be collected in considerably larger quantities. However, we are not aware of any direct comparison between these two types of data. We also believe this is the first work where silver data is used to identify speech tags.

2 Related Work

Table 1 summarizes related work on the identification of direct speech in literary works. This excludes some distantly related work, e.g. targeting other genres such as news texts and languages that predominantly use quotation marks, such as English. For a detailed review, see Stymne (2024).

All papers but one use either existing gold data or collect silver data for their experiments. Only one paper, Kurfalı and Wirén (2020) use both variants. However, their main point of investigation is to explore the feasibility of cross-lingual zero-shot training for direct speech identification, so they compare using English silver data, to using in-language gold data for German, Portuguese , and Swedish, which does not constitute a fair comparison with respect to only data type.

While not always clearly spelled out, in most previous work, there seems to be a mixture of typographical markers in both training and test data, with the exceptions of Stymne and Östman (2022) and Durandard et al. (2023), who used separate test sets for different types of typographical markings. In several works, though, it is noted that varying typographical markings in training data is a major source of misclassification and that data without quotation marks and dashes is considerably harder to classify.

Work	Language	Data type	Modelling/Eval.	Method	Marks	Miscellaneous
Brunner (2013)	German	Gold	Sentence, work	Rule, Random forest	Mixed, incl. QM	STWR
Schöch et al. (2016)	French	Gold	Sentence	SVM, MaxEnt, ...	Dash/Mix(?)	Applied
Jannidis et al. (2018)	German	Silver	Sentence, token	Log. regr., LSTM, ...	Mixed	Applied
Ek and Wirén (2019)	Swedish	Gold	Token	Log. regr., rule	Stripped speech lines	
Tu et al. (2019)	German	–	Sentence, token	Rule	No-QM	
Brunner et al. (2020)	German	Gold	Token	BiLSTM-CRF+BERT/FLAIR	Mixed (often QM)	STWR
Byszuk et al. (2020)	9 languages	Gold	Token	BERT-ft, rule	Mixed	
Kurfalı and Wirén (2020)	4 languages	Silver (En)	Token	mBERT-ft	Stripped	Cross-lingual
Dahllöf (2022)	Swedish	Silver	Segment	Multi-layer perceptron	Stripped dash lines	Applied
Stymne and Östman (2022)	Swedish	Gold	Token/Span	BERT-ft	Mixed	Speech tags
Durandard et al. (2023)	French	Gold	Token/Several	Rule, BERT-ft, BiLSTM-CRF	Mixed	

Table 1: Summary of work on direct speech identification of literary works. Data type distinguishes training on human annotated gold data, and automatically extracted silver data. Method refers to the main method used (ft: fine-tuning, rule: rule-based modeling). For modeling and evaluation, it is stated if it is performed on the token level, span level (i.e. for each speech sequence), segment level (i.e. segment between punctuation marks), sentence level (i.e. does a specific sentence contain speech), or on the work level (i.e. based on the percentage of speech predicted for a full work); Several refers to usage of more than one granularity. We also make a best effort to categorize the type of typographical marking used in each study, which is challenging since it is not always directly stated; here QM stands for quotation mark. Miscellaneous notes some additional aspects of the work, where STWR stands for speech, thought, and writing representation, works marked as such are not restricted to only identifying direct speech. Works marked with Applied, apply the classifiers to a large set of literary works, for further analysis.

	Tokens	Speech	Tags
Gold train	110K	1881	863
Gold dev	17K	201	90
Gold test:dash	38K	883	325
Gold test:none	25K	577	336
Silver train	6290K	88097	34114

Table 2: Size of data in total number of tokens (for stripped versions), number of speech (segments) and number of (speech) tags.

3 Data

Table 2 summarizes the size of the data in number of tokens (for stripped versions), number of speech segments, and number of speech tags.

3.1 Gold Dataset: SLäNDa

Our gold training data comes from the SläNDa corpus version 2.0 (Stymne and Östman, 2022), a collection of excerpts from 19 novels from 1809–1940, manually annotated for speech and other features not forming part of the frame narrative, such as thoughts, quotes, and letters, which we merge with the other class in this work. We use the suggested training and development splits.[2] The training data of SLäNDa contains a mix of typographical markings and no markings. We use the original version (*Gold-mix*), as well as a stripped variant where speech marking punctuation is removed (*Gold-strip*), as well as mixing these two variants (*Gold-combo)*. The test sets from SLäNDA are separated by the graphical speech marking. We use the testsets without marks (*None*) and the testset with stripped dashes (*Dash-strip*).

3.2 Silver Dataset

We collect a new silver dataset[3] by gathering novels and collections of short stories from the same period as the SLäNDa data from Litteraturbanken.[4] We select works of high-quality proofread OCR, which we filtered to only keep those that use quotation marks for speech marking and do not have dashes at the start of lines. From this data, we extracted speech segments by selecting all sequences surrounded by quotation marks. Speech tags are identified using two heuristics, in relationship to the first speech segment in a paragraph. (1) If the first speech segment is preceded by a colon (either within the paragraph, or in the previous paragraph), we search for the preceding punctuation mark or the start of a line, and mark the tokens in this stretch as a speech tag. (2) If the first speech segment of a line is not followed by a period, we mark any tokens up until a sentence-final punctuation mark or another quotation mark as a speech tag. We filter the extracted data to exclude works with few speech segments and speech tags, resulting in data from 88 works. We prepare two versions of the silver data: *Silver-quote*: with original quotation marks kept (not matching the SLäNDa test data)

[2]Available at https://lindat.cz/repository/xmlui/handle/11372/LRT-4739

[3]Available at https://github.com/UppsalaNLP/LitDialogSilver/.

[4]https://litteraturbanken.se/

Test data→ Training data↓	Dash-strip P	Dash-strip R	None P	None R
Gold-mix	**94.25**	87.93	93.42	89.13
Gold-strip	93.41	92.02	**94.18**	91.51
Gold-combo	92.47	**92.57**	94.14	**92.12**
Silver-quote	28.97	2.04	32.03	5.53
Silver-strip	**94.04**	**86.74**	**87.58**	**76.63**

Table 3: Macro-average results with different variants of gold or silver training data.

and *Silver-strip*, where all quotation marks are removed.

3.3 Mixed Dataset

We also perform experiments where we mix gold and silver data. For these experiments, we combine *Gold-combo* and *Silver-strip*.

4 Experimental Setup

We model the task of identifying direct speech segments and speech tags as a token classification task. Based on previous work, summarized in Table 1, we choose to fine-tune a BERT model for token classification based on the IOB2-schema of our data, which has been used in the majority of the most recent works. We use the Machamp toolkit (van der Goot et al., 2021), a toolkit for various NLP tasks, based on fine-tuning an LLM. When mixing gold and silver data, we use the smoothing feature of Machamp to give higher weight to the gold data (α=0.25). As the base LLM, we use the Swedish BERT-model KBBert (Malmsten et al., 2020). In all our experiments, we use the development set from SLäNDa to select the best model across all epochs, to be used for testing.

We use token-level evaluation, where we ignore punctuation and the distinction between *B-* and *I-* tags. We report precision, recall, and F1-score for speech segments and speech tags separately, as well as macro-averaged scores over the two classes.

5 Results

Table 3 shows an overview of the main results, with macro-average scores for different types of gold and silver training data on our two test sets. In all cases, we get better results when training on gold data than on silver data. However, the difference in precision is small for the *Dash-strip* test set, and we see a much larger difference in recall for both test sets. As expected, we see that training on the original silver data with quotation marks, which does not match the test data at all, gives very poor results, especially on recall. However, stripping quotation marks leads to relatively strong scores, especially on precision. This may indicate that the silver data is accurate, but may not include all types of speech segments and speech tags that are needed. For the gold data, all three variants of training data perform reasonably well, but training on the original data gives somewhat lower recall, than when including stripped data. This indicates that we need to have some data in the training that matches the test set, but it does not seem to help to exclude data that does not match, as seen by the similar performance of *Gold-strip* and *Gold-combo*.

In Table 4 we report scores per class, with the overall best-performing variants of training data, as well as a mixed variant with both gold and silver data. Again it is clear that gold training data is overall preferable to the much larger silver data. Mixing gold and silver data does not help, with the scores being equal or lower to the best gold or silver scores. For speech segment identification, we see that there is a precision/recall tradeoff, with silver data giving slightly higher precision than gold data, but at a cost of recall. For speech tag identification, silver data always performs worse than gold data, especially for the *None* test set. Artificially unmarked data, by stripping, seems to perform well on the task.

For speech tag identification, we see that the precision is similar to that for speech segment identification with gold training data, but that recall is lower. With silver and mixed data the scores are much lower for speech tag identification, especially on recall. With silver training data, scores for speech tag identification are considerably higher on the *Dash-strip* than *None* test set. We think this may indicate that the dash data, even when stripped, has more similarities to the data with quotation marks, in how speech tags are used, than the data that was originally unmarked. The low recall on speech tag identification with silver data also indicates that the quality of the heuristic speech tag identification could be improved.

Overall, we have good results when training with gold data, and testing on these challenging unmarked test sets, both for speech segments and for speech tags. We think the quality is good enough for applying these classifiers for work focusing on

	Dash-strip						None					
	Speech			Tags			Speech			Tags		
	P	R	F	P	R	F	P	R	F	P	R	F
Gold-strip	92.01	96.40	94.15	94.81	87.63	91.06	93.01	94.14	93.56	**95.36**	88.87	92.00
Gold-combo	89.53	**96.96**	93.08	**95.42**	**88.18**	**91.59**	93.47	**94.39**	**93.92**	94.81	**89.86**	**92.27**
Silver-strip	**94.91**	94.34	94.60	93.16	79.15	85.26	**96.12**	90.66	93.30	79.04	62.60	69.76
Mixed	93.88	95.37	**94.61**	90.36	80.52	85.16	96.09	89.92	92.87	72.62	65.01	68.56

Table 4: Token-level results for speech segments and speech tags for the best models on data without any typographic markers.

analyzing literary text. Training with silver data, which may be the only available data for some languages, gives reasonably good results on speech segment identifications, but further work is needed to better identify speech tags with silver training data.

6 Additional Findings

We have described experiments focused on the identification of direct speech in challenging datasets without speech marking, focusing on the impact of gold versus silver training data, and the graphical speech marking of training data. In this section, we will briefly summarize the additional experiments presented in Stymne (2024), where we also presented results on the SLäNDa testset with dashes and explored the impact of metric granularity.

For the *dash* test set, we also saw considerably higher recall when training on gold data than on silver data. However, precision was typically a little bit higher with silver or mixed training data, than for gold. The results when training with stripped gold data were slightly lower than with the original training data, that contained dashes. For the silver data, there was a low recall both for the original data with quotation marks and with the stripped variant. We also tried to replace quotation marks with dashes in the silver data, and while improving over the other two variants of silver data, it was still much lower than with gold training data on recall, while giving the highest macro-average precisison.

We also compared the token-level evaluation reported in this paper, to a strict span-level metric, where only exact matches of the full span counted as correct, including any leading or trailing punctuation marks. On this metric, the results were relatively consistent for the *dash* test set. However, for the two test sets presented in this paper, the results were conflicting on the two metric types, with the span-level metric often favoring the silver training data over the gold training data. We believe this is mainly due to some inconsistencies in the human annotation of punctuation marks at span boundaries in the gold training data, which could lead to non-matching spans. This issue does not occur in the silver data, which is deterministically constructed based on heuristics. Due to this, we believe the token-level results to be more representative of the usefulness of the classifiers and thus presented them in this paper.

7 Conclusion

We explore several aspects related to the automatic identification of direct speech segments and speech tags in Swedish literary works. We focus on the usefulness of manually annotated gold data, compared to automatically annotated silver data, and the impact of different types of typographical marking of speech in the training data. We found that overall, we had the best results when using the smaller gold data, especially on speech tag identification. Mixing gold and silver data did not lead to further improvements. The training data needs to contain the type of speech marking that is used in the target data, possibly by artificially stripping speech markers, but may also contain other variants, to ensure a reasonable performance.

In future work, we plan to extend the current study with a detailed error analysis. It would be interesting to explore both the quality of the silver data, especially for speech tags, and the classification results in more detail. Another line of work would be to compare classification using silver data for the target language to gold data in a cross-lingual setting, expanding on Kurfalı and Wirén (2020). We think the current classifiers are strong enough to apply to research in digital literature studies where the identification of direct speech and/or speech tags is needed. We plan to use such a classifier to investigate changes in the Swedish written language in literary narrative and dialog over time.

Acknowledgements

This work is funded by the Swedish Research Council under project 2020-02617: *Fictional prose and language change. The role of colloquialization in the history of Swedish 1830–1930*. I would like to thank David Håkansson, Carin Östman, and Mats Dahllöf for helpful discussions.

Computations and data handling were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725, and the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

References

Sarah Allison. 2018. *Reductive Reading. A Syntax of Victorian Moralizing*. John Hopkins University Press, Baltimore.

Annelen Brunner. 2013. Automatic recognition of speech, thought, and writing representation in german narrative texts. *Literary and Linguistic Computing*, 28(4):563–575.

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)*, pages 114–118, Online.

Joanna Byszuk, Michał Woźniak, Mike Kestemont, Albert Leśniak, Wojciech Łukasik, Artjoms Šeļa, and Maciej Eder. 2020. Detecting direct speech in multilingual collection of 19th-century novels. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 100–104, Marseille, France. European Language Resources Association (ELRA).

Mats Dahllöf. 2022. Quotation and narration in contemporary popular fiction in Swedish: Stylometric explorations. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference*, pages 203–211, Uppsala, Sweden.

Noé Durandard, Viet Anh Tran, Gaspard Michel, and Elena Epure. 2023. Automatic annotation of direct speech in written French narratives. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7129–7147, Toronto, Canada. Association for Computational Linguistics.

Adam Ek and Mats Wirén. 2019. Distinguishing narration and speech in prose fiction dialogues. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, pages 124–132, Copenhagen, Denmark.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Fotis Jannidis, Leonard Konle, Albin Zehe, Andreas Hotho, and Markus Krug. 2018. Analysing direct speech in German novels. In *Abstract zur Konferenz Digital Humanities im deutschsprachigen Raum 2018*, pages 114–118, Cologne, Germany.

Murathan Kurfalı and Mats Wirén. 2020. Zero-shot cross-lingual identification of direct speech using distant supervision. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 105–111, Online. International Committee on Computational Linguistics.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the National Library of Sweden - making a Swedish BERT. *arXiv*, arXiv:2007.01658v1.

Eric T. Nalisnick and Henry S. Baird. 2013. Character-to-character sentiment analysis in Shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 479–483, Sofia, Bulgaria. Association for Computational Linguistics.

Christof Schöch, Daniel Schlör, Stefanie Popp, Annelen Brunner, Ulrike Henny, and José' Calvo Tello. 2016. Straight talk! Automatic recognition of direct speech in nineteenth-century French novels. In *Digital Humanities 2016: Conference Abstracts*, pages 346–353, Kraków, Poland.

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. 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.

Ngoc Duyen Tanja Tu, Markus Krug, and Annelen Brunner. 2019. Automatic recognition of direct speech without quotation marks. A rule-based approach. In

Proceedings of Digital Humanities: multimedial & multimodal. 6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum, pages 87–89, Frankfurt am Main, Germany.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.