Direct Speech Identification in Swedish Literature – an Exploration of Training Data Type, Typographical Markers, and Evaluation Granularity

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 with quotation marks. 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 and of using evaluation metrics of different granularity. We perform experiments on Swedish literary texts and find that using gold and silver data has different strengths, with gold data having stronger results on token-level metrics, whereas silver data overall has stronger results on span-level metrics. 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 Introduction

The main narrative of literary works is typically interspersed with dialogues representing direct speech utterances by the characters in the work. Distinguishing narrative and direct speech is important for work on digital literature studies, for tasks including identifying the social networks of novels (Elson et al., 2010) and analyzing the sentiment of characters towards each other (Nalisnick and Baird, 2013). In addition to speech segments, we are also interested in speech tags, or reporting clauses, Speech tags can have different lengths and positions with respect to the speech, exemplified in (1–3).[1] Speech tags are also relevant for work in literary studies, such as Allison (2018) who study them as a means of analyzing Dickens' narrative perspective.

(1) – Står morsan och drömmer? sade hon skarpt. Raska på.
'– Are you dreaming, mum? she said sharply. Hurry up.'
(M. Sandel, *Hexdansen*, p. 46)

(2) Han sa:
Varför står du här och skräpar?
'He said:
Why are you idling here?'
(H. Bergman, *Chefen fru Ingeborg*, p. 15)

(3) – Min chef, sade jag till domaren med en röst som jag förgäves sökte göra stadig, får jag ge honom en spruta till?
'– My boss, I said to the judge with a voice that I tried to keep stable to no avail, may I give him another shot '
(K. Boye, *Kallocain*, p. 264)

Speech segments are often marked typographically to distinguish them from the narrative. In English, the standard is to mark them with quotation marks, which makes both the start and end of such segments easily identifiable. However, in other languages, there is a variety of ways to mark speech, such as using a dash at the start, but not at the end of speech segments or at the restart after speech tags, as in Example 1. In some works, speech is not marked at all, as in Example 2. In these cases, it is much more challenging to identify speech segments, since the typography is not enough, and there is a need to use textual cues, such as reporting verbs and tense shifts. In this work, we focus on Swedish literary works from 1809–1940, containing a mix of speech marking styles.

Most previous work on direct speech identification for literature is based on different types of machine learning. Such systems have been trained

[1] All translations into English from the original Swedish are our own. In examples, we mark direct speech in blue and speech tags in purple.

Proceedings of LaTeCH-CLfL 2024, pages 253–263

on two types of data: gold or silver. Gold data consists of humanly annotated data. Such data is typically of high quality and may contain a variety of typographical markings. However, it is typically relatively limited in size. By silver data, we mean data that has been automatically extracted from literary texts, normally by identifying works that use quotation marks, and assuming that text within quotation marks constitutes a speech segment. The advantage of such data is that it is easy to collect large annotated corpora. However, the quality is typically lower than for gold data since quotation marks can also be used for other purposes, such as marking quotations, irony, and unusual usage of words or terms. To the best of our knowledge, all previous work on direct speech identification for a single target language has either used gold or silver data, which means that a direct comparison between the usefulness of the two types of data is lacking. The only exception is Kurfalı and Wirén (2020) who worked on zero-shot cross-lingual classification, and compared English silver data to an in-language gold baseline. In this work, we fill that gap, by contrasting the use of a large silver dataset with a smaller manually annotated gold dataset, taken from the SLäNDa corpus (Stymne and Östman, 2022), for the same language, Swedish. As far as we are aware, this is also the first effort to use silver data for the identification of speech tags. While extraction of speech is straightforward in texts using quotation marks, automatically extracting speech tags requires additional heuristics.

There is also little previous investigation of the impact of the use of different typographical markers in the training and test data of classifiers. Stymne and Östman (2022) provided separate test sets for different types of marking but performed only a small pilot experiment. In this work, we extend their study and explore the issue in more detail. The task setup as well as the metrics used in previous works have also varied between studies. In this work, we model the task of identifying direct speech and speech tags as a token-classification problem. Unlike previous studies, we evaluate it on two levels of granularity, both at the token level and at the span level, which requires the exact matching of a full span. This allows us to investigate the effect of metric choice on the results.

To sum up, we investigate the following research questions, in the context of identification of direct speech segments and speech tags in literary works:

RQ1 Is it preferable to use smaller gold data or larger automatically annotated silver data for direct speech identification?

RQ2 Can heuristically constructed silver data be useful for speech tag identification?

RQ3 Is it possible to improve speech and speech tag identification by mixing gold and silver data?

RQ4 What is the effect of different typographical marking of speech in training and test data?

RQ5 What is the effect of using span-level versus token-level evaluation metrics for direct speech identification?

In addition, we provide a detailed overview of related work for the task of direct speech identification.

2 Related Work

In this section, we focus on reviewing related work on the identification of direct speech in literary works for cases where quotation marks are not predominant. This excludes some distantly related work, e.g. targeting other genres such as news texts (e.g. Pouliquen et al., 2007; Quintão, 2014), and work on languages that predominantly use quotation marks, such as English (e.g. Elson et al., 2010; Muzny et al., 2017). Table 1 gives an overview of a selection of relevant work, and summarizes the main setup of each study. In the following, we will go through and discuss each category of Table 1.

Language Most work focuses on one language, in most cases either German or Swedish, with one study on French. Two works explore multiple languages, including a cross-lingual setup (Kurfalı and Wirén, 2020) and multilingual training (Byszuk et al., 2020). The latter found that for many languages, including English, a rule-based system based on punctuation marks gave near-perfect accuracy. However, for other languages, especially Norwegian, which is closely related to Swedish, the rule-based system performed poorly, due to mixed graphical speech marking.

Training data 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,

Work	Language	Training data	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. (2020b)	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

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 in the paper (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). 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. 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.

to using in-language gold data for German (Brunner et al., 2020a), Portuguese (Quintão, 2014) and Swedish (Stymne and Östman, 2022), which does not constitute a fair comparison with respect to only data type. They do also use English gold data (Papay and Padó, 2020), for which they found that the performance is better with the gold data, with a token-level F1-score of 0.89 compared to 0.85 with the silver data. The silver data contained English books from Project Gutenberg, where speech was extracted based on quotation marks.

Modelling and Evaluation The direct speech identification task has been set up in different ways, the two most common options being either to identify lines containing speech or a token-level classification of tokens as being part of a speech segment or not. For token-level classification, some works feed only speech lines to the classifiers, while some feed the full text, also including non-speech lines. The modeling of slightly different tasks also affects the evaluation choice. The metrics presented for each granularity are accuracy or precision, recall, and/or F1-score. Normally the evaluation granularity follows the modeling, with the exception of (Stymne and Östman, 2022) where token-level classification was evaluated on the span-level. None of the surveyed studies evaluated on more than one level of granularity for a single task formulation. The variety of task formulations, metrics, languages, and datasets used, makes direct comparisons between different papers difficult.

Method The methods used mainly follow the evolution of the computational linguistics field, with some older work using rule-based methods, followed by classical machine learning approaches like logistic regression and SVM, while the majority of newer studies use neural methods, mainly fine-tuning of transformer models. Relatively few works directly compare different types of approaches. Ek and Wirén (2019) found that an SVM-based method worked considerably better than a rule-based baseline, and Brunner (2013) found that a random forest approach was better than a rule-based approach at identifying direct speech, especially for unmarked cases, and noted that the rule-based method suffered in the absence of quotation marks. Brunner et al. (2020b) found that FLAIR character embeddings performed better (for direct speech, but not for other types) than BERT-embeddings as input to a BiLSTM-CRF. Two works also compared different classical machine learning approaches (Schöch et al., 2016; Jannidis et al., 2018).

Typographical marks In Table 1 we attempt to describe the typographical markers of speech in each article. However, it is typically not clearly stated what the mix is, more than at a very high level, as indicated in the table summary. In a few studies, typographical markers are stripped from the data to investigate how well the task can be done in their absence. However, in most other studies, there seems to be a mixture of typographical markers in both training and test data. The exception is our previous work (Stymne and Östman, 2022) where we used a mixed training dataset, but with separate test sets for works with quotation marks, dashes, and no consistent marking. In addition, we explored stripped versions of these datasets. Our

overall finding was that it was preferable for both speech and speech tag identification to use training data that matched the graphical speech marking of the intended target data. However, the experiments were limited, and only strict span-level metrics were used. In several other works, the analysis of the results reveals insights relating to typographical markers. Brunner et al. (2020b) noted that a main source of misclassification is the absence of quotation marks and Byszuk et al. (2020) noted that mixing data using quotation marks and dashes may have introduced noise, affecting the performance negatively.

Miscellaneous Several papers classify not only direct speech, but also indirect, free, and free indirect speech, thought, and writing, marked with STWR in Table 1. These papers use the German corpus REDEWIEDERGABE (Brunner et al., 2020a) for training, which contains all these levels, based on principles for English (Leech and Short, 1981). None of the corpora used for training of the other languages contain STWR annotations.

The only study that attempted to identify speech tags in addition to speech segments is Stymne and Östman (2022). In the training data of most other papers, speech tags are not annotated, and can thus not be extracted. The German REDEWIEDERGABE corpus do include annotations of speech tags. However, we are not aware of any work that has used this corpus for speech tag identification.

In Table 1, we also marked works that apply the direct speech identification to analyze a high number of additional literary works. Schöch et al. (2016) investigated the proportion of direct speech in different genres and over time in French novels and Jannidis et al. (2018) investigated the proportion of direct speech over time in German low- versus high-brow novels. Dahllöf (2022) performed a stylometric exploration of differences between the narrative and direct speech in modern Swedish novels.

3 Data

In this section, we give an overview of the SLäNDa corpus that we used for evaluation and gold training data. We then go on to describe the extraction of a new large silver training dataset, based on literary works with quotation marks.

	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 training	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.

3.1 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 main narrative, such as thoughts, quotes, and letters. Since all classes except speech and speech tags are rare, they grouped all non-speech classes into an *other* class for their experiments. We use the suggested training and development splits[2] and further adapt it by not considering the *other* class, and only distinguishing between speech segments, speech tags, and narrative (including the *other* class). The training data of SLäNDa contains a mix of typographical markings and is available in two versions, the original version, which contains a mix of quotation marks, dashes, and no marking, which we will call *Gold-mix*, and a stripped version, with quotation marks and dashes removed, which we call *Gold-strip*. We also experimented with a concatenated version containing both variants: *Gold-combo*.

We use the recommended test sets in SläNDa v2.0, which contains two main sets: data from works with dash marking, and from works with with no consistent marking, mainly with no marking at all. We refer to these datasets as *Dash* and *None* respectively, and further also use the provided stripped version of the dash test set: *Dash-strip*. Table 2 summarizes the size of these sets in the number of tokens (for stripped versions), number of speech segments, and number of speech tags.

3.2 Silver Dataset

We collect a new silver dataset by gathering novels and collections of short stories from the same period as the SLäNDa data from Litteraturbanken, a publicly available collection of Swedish literature

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

Grefvinnan	log	och	betraktade	henne	innerligt	.	Var	icke	rädd	för	mig	,
The countess	smiled	and	watched	her	dearly	.	Be	not	scared	for	me	,
O	O	O	O	O	O	O	B-SPE	I-SPE	I-SPE	I-SPE	I-SPE	I-SPE
mitt	barn	–	kom	närmare	!	sade	hon	.				
my	child	–	come	closer	!	said	she	.				
I-SPE	I-SPE	I-SPE	I-SPE	I-SPE	I-SPE	B-TAG	I-TAG	I-TAG				

Figure 1: IOB2 scheme for a sample paragraph, with English glosses for clarity ('The countess smiled and watched her dearly. Don't be afraid of me, my child – come closer! she said.' C. J. L. Almqvist, *Syster och bror*. p. 27).

works.[3] From Litteraturbanken, we selected 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 (dashes can be used for other purposes, but typically sentence-internally). This filtering resulted in 141 works from 1821–1931.

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. These two heuristics would cover instances similar to examples (1–3). To further improve the quality, and have data that is not overly imbalanced, we applied two filtering strategies, based on the extracted entities. We only kept works where speech tokens constituted at least 20% of the total number of tokens and where there was at least a ratio of 20% speech tags, compared to speech segments. After this filtering, we were left with 88 works. The proposed heuristics are not perfect and the silver data still contains some noise. However, it is considerably larger than the gold data, as shown in Table 2.

We prepare three versions of the silver data: *Silver-quote*: with original quotation marks kept (not matching the SLäNDa test data), *Silver-dash* with quotation marks replaced by an initial dash, and *Silver-strip*, with all quotation marks removed. The data was prepared in the same format as SLäNDa. The silver data is publicly available under the CC BY-NC-SA license.[4]

[3] https://litteraturbanken.se/

[4] https://github.com/UppsalaNLP/LitDialogSilver/

4 Experimental Setup

In this section, we describe how we model the task, the system we use, and the evaluation metrics used.

4.1 Modelling

We model the task of identifying direct speech segments and speech tags as a token classification task. We follow Stymne and Östman (2022) and use the IOB2-scheme for representing the data, with tags for speech segments, *SPE* and speech tags *TAG*, and all other tokens marked with an *other* tag, *O*.

Figure 1 exemplifies the IOB2-scheme used, from a novel without speech marking. In case there would have been a dash or quotation marks indicating speech, they would have been included in the speech segment annotation. Also, note that dashes can be used for other purposes than speech marking; here one is used within a speech segment.

4.2 System

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, with support for using multiple datasets. Machamp has given competitive results on several tasks and has features that suit our experimental design. We use their *seq_bio* encoder, which is a CRF model enforcing consistency with IOB-schemes. As the base LLM, we use the Swedish BERT-model KBBert (Malmsten et al., 2020).

For experiments where we train on both gold and silver training data, with very different sizes, we take advantage of the dataset smoothing feature of Machamp, used to control how to sample instances from different datasets. The sampling is based on a multinomial distribution, controlled by the variable α, where $\alpha = 1.0$ means that the original

Test data→	Span-level						Token-level macro					
	Dash		Dash-strip		None		Dash		Dash-strip		None	
Training data↓	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
Gold-mix	**82.52**	**87.73**	74.12	74.35	**76.66**	**81.05**	92.41	**93.01**	92.47	**92.57**	94.14	**92.12**
Gold-strip	46.48	50.39	**75.70**	**80.06**	74.09	79.45	**93.32**	91.54	**93.41**	92.02	**94.18**	91.51
Gold-combo	79.49	84.36	74.56	78.46	73.49	78.28	92.41	**93.01**	92.47	**92.57**	94.14	**92.12**
Silver-quote	**67.55**	6.00	15.98	3.00	14.17	5.33	78.29	41.68	28.97	2.04	32.03	5.53
Silver-strip	59.93	35.24	**92.15**	**87.68**	**85.88**	**79.60**	91.72	51.95	94.04	**86.74**	87.58	**76.63**
Silver-dash	52.79	**50.02**	44.66	29.78	49.90	33.39	**95.57**	**65.52**	**95.11**	54.99	**95.31**	52.58

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

data sizes are used, and $\alpha = 0.0$ means that an equal amount of data from each dataset is used. We experiment with different values of α for mixing gold and silver data. We use the default Machamp settings for *seq_bio* for all other hyper-parameters. With gold data we run for 20 epochs, and when using the much larger silver data, for 10 epochs. In all our experiments, we use the development set from SLäNDa to select the best model across all epochs, to be used for testing.

4.3 Evaluation

For evaluation, we use both span-level and token-level metrics. For span-level evaluation, which is a strict metric requiring the exact matching of a span, including any graphical marker of speech, we use the conlleval script, originally used to evaluate chunking in CoNLL 2020 (Tjong Kim Sang and Buchholz, 2000).[5] For the token-level evaluation, we ignore punctuation and the distinction between *B*- and *I*-tags. The reason for ignoring punctuation in token-level evaluation is to ensure a fair comparison between the original and stripped data versions, which differ in the punctuation marks used for speech marking. We use our own implementation of token-level evaluation. For both granularities, we report precision, recall, and F1-score for speech segments and speech tags separately, as well as macro-averaged scores over the two classes. We repeat all experiments three times with different random seeds and report average results

5 Results

In this section, we present and discuss the results, followed by a summary of our main findings.

6 High-Level Results

In our first experiment, we compare macro-average precision and recall for all variants of gold and silver training data. For F1-scores, we refer to the detailed results in Tables 4–7. Results are shown in Table 3.

Gold Versus Silver Data

For the *Dash* test set, which contains dashes for speech marking, the performance is overall higher with gold data than with silver data, except for token-level precision, which is slightly higher, but with a considerably lower recall.

For the two test sets with no (consistent) marking, *Dash-strip* and *None*, there is a precision/recall tradeoff on the token-level metrics, with considerably higher recall when trained on gold data, and higher precision with silver data. On the span-level evaluation, results are overall better with silver data.

Token- Versus Span-Level Evaluation

For the two test sets without marking, we see a clear difference between the two metrics. On span-level evaluation, matching silver data performs better than gold data, whereas the recall is considerably higher for gold data than for silver data when evaluated on the token level, however, with slightly lower precision. On the *Dash* test set, all gold training data sets perform well on token-level evaluation, whereas there is a large difference between the span-level results, which is due to dashes not being identified as speech markers, which means that the whole span is not matched.

A preliminary investigation into the difference between the two metrics, especially on the two unmarked test sets, showed that the gold SLäNDa data has some inconsistencies in the annotation of punctuation marks between speech segments and speech tags, as well as at the end of speech segments and tags. The silver data, on the other hand, is consistent in this respect, since it was annotated by rule-based heuristics. This seems to be one reason why it is harder to match full segments with gold training data since just missing a punctuation

[5] Retrieved from https://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt.

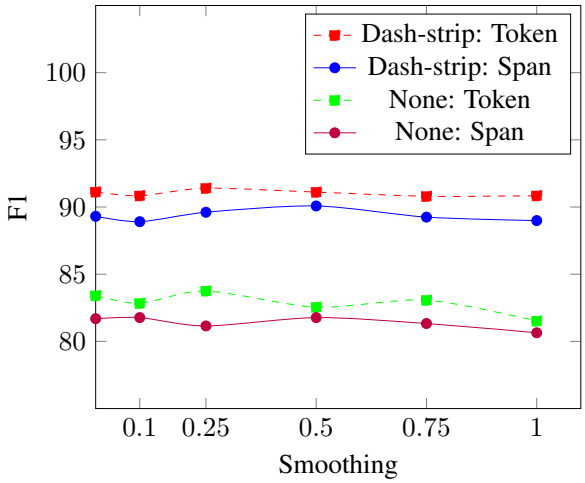

Figure 2: F1-scores at the span and token level (macro), for models trained on both gold and silver data, with different smoothing values.

mark will mean missing a full segment, whereas such an error will not be included in the token-level metrics, which ignores punctuation.

Impact of Typographical Marking

For the *Dash* test set, it is overall best to use training data containing dashes, i.e. *Gold-mix* and *Silver-dash*, with a few exceptions with higher precision at a cost of a lower recall. For the two test sets without marking, the difference between the three types of gold training data is generally very small for both types of evaluation. With silver data, the clearly best option is to use matching training data in the form of *Silver-strip*. It is interesting to see that when training on the silver training data, the recall is much better for *Dash-strip* than for *Dash*, without hurting precision. When training on gold data, these two data sets have similar results on token-level metrics, but *Dash* performs better on the span-level metrics.

We note that the very low recall with *Silver-quote* training is due to the mismatch of punctuation marks between the training and test data, leading to the system rarely predicting speech without punctuation marks. When training with *Silver-strip*, the performance is higher when testing with dashes, on the *Dash-strip* test set than without marking, on the *None* test set, indicating that literary works with some kind of graphical marks may share some similarities compared to original unmarked speech.

	Speech			Tags		
	P	R	F	P	R	F
Gold-mix	93.57	**95.82**	**94.68**	93.78	**87.52**	**90.54**
Silver-dash	93.52	77.37	84.68	**97.63**	53.67	69.25
Mixed .25	93.72	91.62	92.65	89.47	77.08	82.81
Mixed .50	**94.51**	88.36	91.33	92.57	76.38	83.68

Table 4: Token-level results for speech (segments) and (speech) tags for the best models on the *Dash* test set.

	Speech			Tags		
	P	R	F	P	R	F
Gold-mix	85.03	**91.47**	88.13	80.01	**84.00**	81.95
Silver-dash	67.78	69.27	68.50	37.80	30.77	33.89
Mixed .25	87.24	86.11	86.66	88.38	81.13	84.59
Mixed .50	**91.20**	88.49	**89.82**	**90.74**	81.44	**85.83**

Table 5: Span-level results for speech (segments) and (speech) tags for the best models on the *Dash* test set.

6.1 Mixing Gold and Silver Data

To further explore the usefulness of gold versus silver data, we perform an experiment where we combine *Gold-combo* and *Silver-strip* training data. We choose these variants based on initial results, with the main focus on the two test sets without marking, but with the goal of also achieving reasonable performance on the test data with dashes.

Figure 2 shows the macro-average F1-scores for the two test sets without dashes with different values for α, which controls the smoothing of dataset sizes. Overall the differences are quite small, with slightly worse results with original sizes ($\alpha = 1.0$). For both test sets, $\alpha = 0.25$ gives the best token-level scores and $\alpha = 0.5$ gives the best span-level scores. We thus chose those two values for further analysis.

6.2 Detailed Results

We now show detailed results for speech segments and speech tags separately, for the best training data for each type of test set. Tables 4–7 show these results. Note that we use different gold and silver training data for the test sets with and without dashes, to present the best option for each type of test data.

Again, we see a clear difference in results between the token-level and span-level metrics. With token-level evaluation, we always have the highest recall for a system trained on gold data, and while the precision can sometimes be slightly better with silver or mixed training, the difference in precision is quite small, whereas the difference in recall often is large, especially for speech tag identification.

	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 .25	93.88	95.37	**94.61**	90.36	80.52	85.16	96.09	89.92	92.87	72.62	65.01	68.56
Mixed .50	93.72	93.90	93.81	91.03	80.69	85.54	96.09	85.49	90.44	75.31	64.28	69.27

Table 6: Token-level results for speech segments and speech tags for the best models on data without any typographic markers.

	Dash-strip						None					
	Speech			Tags			Speech			Tags		
	P	R	F	P	R	F	P	R	F	P	R	F
Gold-strip	77.23	83.20	80.10	74.18	76.92	75.52	74.51	81.92	78.04	73.67	**76.98**	75.29
Gold-combo	73.82	79.69	76.64	75.30	77.23	76.25	76.54	82.96	79.62	70.44	73.61	71.99
Silver-strip	**93.80**	92.49	93.14	**90.50**	82.87	86.51	86.38	**87.46**	**86.91**	85.39	71.73	**77.96**
Mixed .25	92.37	92.68	92.51	89.17	**84.41**	**86.72**	85.60	86.66	86.11	**82.70**	70.63	76.19
Mixed .50	93.74	**92.75**	**93.24**	89.83	84.21	86.92	**86.86**	85.50	86.17	84.95	71.03	77.37

Table 7: Span-level results for speech segments and speech tags for the best models on data without any typographic markers.

We note that overall, the recall is lower on speech tags than on speech segments, whereas the difference in precision is smaller. The mixed models overall have a high precision, for the *Dash* test set even higher than with gold training data, but with a lower recall. However, on speech tags, the mixed models perform considerably poorer than gold, on both precision and recall.

For span-level metrics, silver data has a strong performance on the two test sets without marking, and the mixed models also do well. Only in one case, do we see a gold score having the highest value, recall for the *None* test set, which, however, has considerably lower precision than the mixed and silver models. For the *Dash* test set, silver performs poorly, even in the dashed variant, especially for speech tags. Here, gold has the highest recall for both speech segments and speech tags, whereas mixed has the highest precision and F1-score.

Across both metric types, gold in most cases has a higher recall than silver, and mixed training tends to give a higher recall than silver in such cases. However, there does not seem to be overall gains to be had over the strongest model by mixing gold and silver data; at best there is a precision/recall tradeoff. We are slightly surprised at the relatively strong performance for the mixed models on the *Dash* test set for both metric types, since we used *Silver-strip* in it, which performs worse than *Silver-dash* for *Dash*, but apparently it gives enough support in combination with the dashes seen in the gold data.

Another interesting aspect is the performance on the *Dash* test set compared to the *Dash-strip* test set, since these test sets are identical except for the use of dashes. A difference in performance could potentially reveal how important the presence of graphical speech marking in the form of dashes is to the identification of speech segments. We find that on the token-level metrics, the performance with the gold training data differs very little between *Dash* and *Dash-strip*, suggesting that linguistic clues are good indicators of speech. On the span-level evaluation, the results are more mixed. With gold data, the results are worse on *Dash-strip* than *Dash*. With silver data, the performance is overall bad on *Dash*, but better on *Dash-strip*. With mixed training, the results are worse on *Dash-strip* than on *Dash* for speech segments, whereas the difference on speech tags is relatively small. Overall, it thus seems that the system does not solely rely on graphical marking of speech, since it can achieve good results in their absence, especially on the token level, which indicates that there are enough linguistic clues to perform well on this task. However, it seems slightly harder to identify the exact speech boundaries in the absence of dashes.

6.3 Summary of Main Findings

Here we follow up on our research questions, summarizing our main findings.

RQ1: Is it preferable to use smaller gold data or larger automatically annotated silver data for identification of direct speech identification?

According to token-level evaluation, gold data is overall preferable to silver data, especially for achieving high recall. For the two test sets without dashes, silver data gives overall better results on span-level evaluation.

RQ2: Can heuristically constructed silver data be useful for speech tag identification?

Speech tags can be identified reasonably well with silver data on the two unmarked test sets, which match the stripped silver data well, whereas the recall is very low on the test set with dashes. Overall, the performance on speech tags with silver data is lower than for speech segment identification.

RQ3: Is it possible to improve speech and speech tag identification by mixing gold and silver data?

We saw no clear gains by combining gold and silver data. Overall the mixed model performed on par with or slightly worse than the stronger of the gold and silver models for each metric and test set. In the cases where the mixed model performed best on a metric, there was a precision/recall tradeoff.

RQ4: What is the effect of different typographical markings of speech in training and test data?

The target speech marking needs to be present in the training data; training on mismatching quotation silver data always performed poorly and stripping the training data of speech marking negatively affected the test set with dashes, both for silver and gold. As long as there is some matching data, as for the original gold data with mixed marking, the performance is quite strong across test set variants, especially on token-level metrics. Graphical speech marking does not seem necessary for good results on the task since there is no large degradation when stripping dashes from the dataset with dashes.

RQ5: What is the effect of using span-level versus token-level evaluation metrics for direct speech identification?

The results vary considerably between the two metric types, giving partly different pictures of the best option for each data combination. We believe one reason could be the inconsistent annotation of punctuation marks on the border of speech segments and speech tags in the gold data, making full-span identification challenging. Predictably, span-level metrics also suffer with marked speech when the training data is stripped of speech marking. Restricting evaluation to only one metric granularity can give an incomplete picture of the full results.

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, the impact of typographical markers of speech, and the impact of evaluation granularity. We find that using gold and silver data has different strengths, with gold data giving better token-level performance, and silver data often better span-level performance. 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, 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, and specifically explore the reason for the differences between token-level and span-level metrics in more depth. A further line of work is to investigate the use of ensemble models as an alternative to data concatenation, which was not successful in this study. 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. Based on a specific use case, it is possible to choose training data that gives either high recall or high precision, on the token and/or span level. 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 their helpful discussions about this work.

The 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, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020a. Corpus REDEWIEDERGABE. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 803–812, Marseille, France. European Language Resources Association.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2020b. 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.

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.

Geoffrey N. Leech and Michael Short. 1981. *Style in fiction: a linguistic introduction to English fictional prose*. Longman, London.

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.

Grace Muzny, Mark Algee-Hewitt, and Dan Jurafsky. 2017. Dialogism in the novel: A computational model of the dialogic nature of narration and quotations. *Digital Scholarship in the Humanities*, 32(Supl. 2):ii31–ii52.

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.

Sean Papay and Sebastian Padó. 2020. RiQuA: A corpus of rich quotation annotation for English literary text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 835–841, Marseille, France. European Language Resources Association.

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*, pages 487–492, Borovets, Bulgaria.

Marta E. Quintão. 2014. Quotation attribution for Portuguese news corpora. Master's thesis, Técnico Lisboa/UTL, Portugal.

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 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.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

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.