Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5324–5333
Marseille, 20-25 June 2022

SLäNDa Version 2.0: Improved and Extended Annotation of Narrative and Dialogue in Swedish Literature

Sara Stymne and Carin Östman
Department of Linguistics and Philology, Department of Nordic Languages
Uppsala University, Sweden
sara.stymne@lingfil.uu.se,carin.ostman@nordiska.uu.se

Abstract

In this paper, we describe version 2.0 of the SLäNDa corpus. SLäNDa, the Swedish Literary corpus of Narrative and Dialogue, now contains excerpts from 19 novels, written between 1809–1940. The main focus of the SLäNDa corpus is to distinguish between direct speech and the main narrative. In order to isolate the narrative, we also annotate everything else which does not belong to the narrative, such as thoughts, quotations, and letters. SLäNDa version 2.0 has a slightly updated annotation scheme from version 1.0. In addition, we added new texts from eleven authors and performed quality control on the previous version. We are specifically interested in different ways of marking speech segments, such as quotation marks, dashes, or no marking at all. To allow a detailed evaluation of this aspect, we added dedicated test sets to SLäNDa for these different types of speech marking. In a pilot experiment, we explore the impact of typographic speech marking by using these test sets, as well as artificially stripping the training data of speech markers.

Keywords: Literary dialogue, Narrative, Direct speech, Annotation

1. Introduction

Modern computational models for text analysis have allowed large-scale studies in many fields of research, such as language studies and literary studies. Early studies in computational literary studies, as well as in other application areas, often used unsupervised methods, such as topic modelling (Boyd-Graber et al., 2017). However, for many tasks, supervised methods are preferred, necessitating the creation of annotated corpora. In this work we present a corpus, SLäNDa, which contains annotations of speech, narrative, and other non-narrative parts, as well as mentions of speakers, in Swedish literature from the 19th and 20th centuries. The period around the turn of the 19th century is a period of modernization of the Swedish written language, where literature, and especially literary dialogue, is believed to have played a role (Teleman, 2003). SLäNDa is the only available large-scale corpus of Swedish literature annotated for narrative, speech, and speakers.

While it may seem trivial to extract speech segments in a language like English, where speech is typically marked with quotation marks, this is not the case for all languages. In Swedish there are two main types of speech marking, either with quotation marks, or with a dash at the start of a speech segment, not marking where the speech segment ends, where the speech tag starts, or where the speech continues after the speech tag. Example (1) shows an example from Strindberg, page 138, which starts with the speech segment ‘My ladies’, marked with a dash, followed by a speech tag, and then the speech continues without any typographical marking with ‘should we consider …’.[1] In other works there are no typographical markings of speech at all, as in example (2) from Nordström, page 54. This variety, and especially cases that lack speech marking, makes the task of separating speech and narrative challenging for Swedish texts.

(1) – Mina damer, tog pastorn ordet, skola vi anse oss hava arbetat nog i vingården för i dag?
‘– My ladies, the pastor started, should we consider ourselves having worked enough in the vineyard for today?’

(2) Kom då! sade han, och så gingo de.
‘Come on! he said, and so they went.’

In this paper we present SLäNDa version 2.0 (Stymne and Östman, 2022), which is an extended and improved version of SLäNDa version 1.0 (Stymne and Östman, 2020). SLäNDa, the Swedish Literary corpus of Narrative and Dialogue, has the goal of separating the main narrative of literary fiction from any other material, mainly speech segments, but also other elements outside of the main narrative, such as thoughts, signs, and letters. For speech segments, also speakers and, when present, speech tags are annotated. Speech tags are the narrator’s presentation of speech, also known as reporting clauses, as ’he said’ in example (2). In SLäNDa v1.0, there is a selection of chapters from 8 Swedish novels from the period 1879–1944. There is a dedicated test set, however, the test set consists of one chapter each from the novels also present in the training set. This test set also contains a mix of different types of marking of dialogue.

In SLäNDA v2.0, we extend the time period to start at 1809, which has been considered as the start of modern Swedish literature (Tigerstedt, 1956). In addition, we focus on creating more informative test sets

[1] All translations from Swedish into English are our own. See Table 1 for details of the works included in SLäNDa, from which all quotations are taken.

by separating the old test set into three parts with different speech markings: quotation marks, dashes and no consistent markings. We also added two new challenging test sets only containing works that are not present in the training data, one without any marking of speech segments, and one with dashes. We have also added three new works to the training set, made minor changes to the categories used in SLäNDa version 1.0, and performed quality control of the previous annotations in SLäNDa version 1.0. SLäNDa version 2.0 is publicly released under a Creative Commons licence.[2]

In a pilot experiment we investigate the performance of a standard BERT-based model on the identification of speech segments and speech tags. Our new dedicated test sets allow us to do this separately for texts with different types of speech marking. We also investigate the usefulness of typographical marking, by training and testing our models both on the original texts, and on a version of the text with quotation marks and dashes stripped. The unmarked and stripped versions require the model to learn textual clues of speech rather than to rely on typographical clues. We find that not surprisingly we get the best results using the original training data, for test data with quotation marks and also with dashes. However, using training data stripped of quotation marks and dashes considerably improves the identification of speech segments as well as speech tags in data without any typographical marking of speech in most cases.

2. Related Work

The distinction between speech and narrative has attracted some attention before. Ek and Wirén (2019) described an effort at separating speech from narrative, using a logistic regression classifier. They annotated excerpts from four Swedish novels, 1,620 lines, partly overlapping the selection in SLäNDa, and also used in Ek et al. (2018).[3] Their classifier reached a token-level F1-score of 80.8, considerably beating a rule-based baseline based on punctuation marks. These scores show that there is plenty of room for improvements, and also a need for additional data.

Kurfalı and Wirén (2020) explored zero-shot cross-lingual identification of direct speech. They automatically created a training corpus with English data, based on identifying speech marked by quotation marks. The system, fine-tuned multilingual BERT, was tested on annotated corpora for four languages: English (Papay and Padó, 2020), Swedish (Stymne and Östman, 2020), German (Brunner, 2013), and Portuguese (Quintão, 2014). While the cross-lingual results do not reach a monolingual supervised baseline, the results are respectable for three literary corpora, with F1-scores for speech identification of .64–.85, with the highest score for English. In Portuguese a news corpus was used, and the domain shift led to lower scores with an F1-score of .33. While this study focused on identification without typographic markers, no comparisons were made between using them and ignoring them.

There have also been some attempts of classifying speech in literature for other languages. Jannidis et al. (2018) attempted both sentence-level and word-level classification for German literature, using neural networks, based on LSTMs for token-level classification. Also for German literature, Brunner et al. (2020) investigated the identification of different types of speech using a binary classifier for each type. None of the above studies attempted to identify speech tags, they only focused on distinguishing speech segments.

There are a number of corpora available for languages other than Swedish, with annotation types overlapping with SLäNDa. Papay and Padó (2020) presented a corpus of English literary texts, with annotations of both direct and indirect speech, containing information about the speaker, addressee and speech verb, but not about full speech tags. Semino and Short (2004) annotated English texts from several genres, including literature, for speech, thoughts and writing, distinguishing between direct, indirect, free-indirect and reported speech. For German, Brunner (2013) described a corpus using the same scheme as Semino and Short (2004), with the addition of speakers, for both fiction and non-fiction. Elson and McKeown (2010) described an English literary corpus with annotation of direct speech, including speakers.

Also focusing on annotation of literature, the Systematic Analysis of Narrative Texts through Annotation (Reiter et al., 2019), provided a shared task where eight teams proposed a set of guidelines for narrative levels in literary texts, which were then evaluated (Willand et al., 2019). The focus in this initiative was not mainly on narrative versus cited materials, though, but rather on narrative levels. Some of the guidelines did discuss characters' speech, though, like Wirén et al. (2020).

Literary texts can be analysed for a large range of purposes. Jannidis et al. (2018) use their automatic annotation of speech to analyse the distribution of speech in German 19th century literature. Elson et al. (2010) use the identification of direct speech and of speakers and addressees as a step towards analysing the social networks in literary novels. In Stymne and Östman (2020) we described a small pilot study, where we used the annotations in the training data of SLäNDa version 1.0 to investigate whether modern versions of a set of function words appear earlier in literary dialogue than in narrative, which overall was the case.

3. Corpus Description

In this section we describe the texts and annotations in SLäNDa. SLäNDa v1.0 is described in detail in Stymne and Östman (2020), so in this paper we summarize the most important details, and focus mainly on

[2] LINDAT/CLARIAH-CZ repository, `http://hdl.handle.net/11372/LRT-4739`

[3] Available from `https://github.com/adamlek/dialogue-fiction`

Author		Novel	Year	Marker	Train	Test	Status
Victor Rydberg	VR	*Den siste Athenaren*	1859	Dash	3	–	New
August Strindberg	AS	*Röda rummet*	1879	Dash	1	1	QC
Victoria Benedictsson	VB	*Fru Marianne*	1887	QM	9	1	QC
Verner Von Heidenstam	VH	*Endymion*	1889	Dash	3	–	New
Mathilda Malling	MM	*En roman om förste konsuln*	1894	Mixed	1	–	New
Oscar Levertin	OL	*Magistrarne i Österås*	1900	Dash	3	1	QC+
Hjalmar Söderberg	HS	*Martin Bircks ungdom*	1901	QM	8	1	QC
Selma Lagerlöf	SL	*Körkarlen*	1912	QM	3	1	QC
Maria Sandel	MS	*Hexdansen*	1919	Dash	1	1	QC
Hjalmar Bergman	HB	*Chefen fru Ingeborg*	1924	Unmarked	9	1	QC
Karin Boye	KB	*Kallocain*	1940	Dash	3	1	QC
Fredrik Cederborgh	FC	*Uno von Trasenberg 1*	1809	Unmarked	–	2	New
Vilhelm Fredrik Palmblad	VP	*Noveller I. Kärlek och politik*	1840	Unmarked	–	2	New
Carl Johan Love Almqvist	CA	*Syster och bror*	1847	Unmarked	–	1	New
Ludvig Nordström	LN	*Borgare*	1909	Unmarked	–	2	New
Edvard Flygare	EF	*Borta och hemma*	1860	Dash	–	1	New
Sophie Elkan	SE	*Dur och moll*	1889	Dash	–	1	New
Mathilda Roos	MR	*Hvit ljung*	1907	Dash	–	2	New
Agnes von Krusenstjerna	AK	*Tonys läroår*	1924	Dash	–	5	New

Table 1: Authors and novels in SLäNDa version 2.0, with the publication year, preferred speech marker (QM for quotation mark, and unmarked for works that mainly lack any markings), number of chapters in test and training parts, and the status compared to SLäNDa 1.0, where 'QC' means quality control, '+' that new chapters have been added from that novel, and New that the material is new to SLäNDa v2.0.

the differences between SLäNDa version 1.0 and version 2.0.

3.1. Text Selection

The period of interest for SLäNDa is the 19th and early 20th centuries. In the description of diachronic linguistic variation in the Swedish language, the turn of the 19th century has been pointed out as a period of extensive change and modernization (Engdahl, 1962). This shift can mainly be described as colloquialisation of the written language, i.e. a drift towards a more oral style (Hundt and Mair, 1999). It seems as this transition initially was quite genre-specific, as it was first observed in fiction. Previous research has proposed that one reason for the leading role of fiction might be the occurrence of direct speech (Teleman, 2003). Direct speech can be expected to be influenced by spoken language to a higher extent than the narrative. But this genre-internal variation has never been thoroughly investigated. A starting point for such an investigation is a separation of direct speech from the narrative, motivating the need of a literary annotated corpus covering this time period. This period is also of interest from a literary perspective, as it is regarded as a period with a narratological shift from "telling" towards "showing" (Allison, 2018). According to previous research one way to follow this shift, where the narrator becomes more and more invisible, is to analyze the speech tag. It has been shown that both the position — before, in the middle of, or after the speech — and the length of the speech tag is of importance (Allison, 2018; Håkansson and Östman, 2019).

All texts are retrieved from Litteraturbanken,[4] *The bank of literature*, a collection of Swedish literary works, with the goal of representing all Swedish literature. Their main focus is on works no longer under copyright, meaning that there is a high number of works for our period of interest from the 19th and early 20th century. The main criteria for including a text in SLäNDa version 1.0 was:

1. Well-known novels
2. Time period: 1870–1940
3. Available in a proofread XML-format
4. Creative Commons license

In SLäNDA version 2.0 we have slightly modified criteria 1 and 2. We now also include collections of short stories (VP, LN, EF, SE) and lesser-known authors, and we have extended the time period to start at 1809, which has been considered the birth of the modern Swedish novel (Tigerstedt, 1956). We also selected new works to create two test sets without overlap with the works in the training data set, where speech is either unmarked or marked with dashes. We did keep criteria 3 and 4, though, which limits the number of available works, especially since many works in Litteraturbanken are only available as images or as raw OCR:ed text, without any proofreading, and not all works are released under a free license.

Table 1 summarizes the contents in SLäNDa version 2.0, and shows how it is related to SLäNDa v1.0, either by additional quality control, or adding new texts. In the case of Levertin, we annotated two additional chapters, since these chapters are quite short compared to

[4] `https://litteraturbanken.se/`

many other novels. Three 19th-century authors, Rydberg, Heidenstam, and Malling, using dashes or mixed marking, were added to the training data. We also added two new challenging test sets with four authors each.

The selected texts from Litteraturbanken were available in an XML-format describing the page layout. We converted this to a text-based format with light XML markup, mainly about chapter and paragraph breaks, see Stymne and Östman (2020) for details.

3.2. Marking of Speech

The two main ways of marking speech in Swedish literature is by using either quotation marks, as in example (4) from Benedictsson, p. 309, or dashes, as in example (3) from Sandel, p. 41. These two variants exist both in older literature and in modern works. However, there are also variants of these markings. The most common way to use dashes is as in example (1), where a dash marks the start of the speech segment, but not its end, and not where the speech restarts after the speech tag. However, there are exceptions to this. In Rydberg, the dash is also used after the speech tag, to mark where the speech restarts, as in example (5), from Rydberg, p. 238. Levertin uses another variant, where he also adds a dash at the end of the first part of the speech, before the speech tag, as in example (6), from Levertin, p. 134. This makes it close to the quotation marks style, but without a final mark at the end of the speech segment. There are also many cases where speech segments are not marked at all, as in example (7) from Cederborgh, page 19. There are also a few cases where there are errors in the typographical markings, as in example (8) from Benedictsson, p. 284, who normally uses standard quotation marks, but where the final quotation mark, after 'in there' is missing in this case.

(3) – Järnet vill inte bli varmt, sade en röst mellan ett par hostningar. Vi har så litet ved. ...
’– The iron will not be warm, said a voice between a couple of coughs. We have so little wood. ...’

(4) ≫Du borde gifta dig≫, sa Börje.
‘≫You should get married≫ , Börje said.’

(5) – Karmides, sade hon mildt, – prisade vare gudarne! Jag har återfunnit dig.
‘– Karmides, she said softly, – praised be the lords! I have found you again.’

(6) – Stackars gosse – sade Roos tyst till Stråle. – Han gör hvad han kan ...
‘– Poor boy – Roos said quitely to Stråle. – He does what he can ...’

(7) Å! jag är bestulen på allt hvad jag äger och har, ropade Uno.
‘Oh! everything that I own has been stolen, Uno shouted.’

(8) ≫Nej, det fins ingen derinne, svarade Marianne.
‘≫No, there is no one in there, Marianne answered.’

It is also possible to mix different styles in the same work. One example is Cederborgh, who mostly has no marking of speech at all, as in example (7), but in some cases use a drama-like style, as in example (9), from Cederborgh, p. 70. There are also a few instances where Cederborgh uses a dash, typically in the middle of a paragraph, which is unusual in other works, as in example (10), from Cederborgh, p. 16. However, dashes are commonly used in this novel for a variety of purposes, and this usage could be viewed as marking a break rather than as marking speech.

(9) *Kattzaun*. Å, du skall väl ha resource.
Uno. Jag spelar alldrig.
‘*Kattzaun*. Oh, you should have resources.
Uno. I never play’

(10) Natten var långt liden och alla trötta af resan. – Nu tillagar jag en pålsk canapée åt mig framför brasan, sade den resande Herrn, ...
‘The night had nearly passed and everyone was tired from the journey. – Now I will cook a Polish canapé for me in front of the fires, the travelling man said, ...’

3.3. Test Sets

In SläNDa v1.0 there was no separation between different types of speech marking in neither of the data sets, even though the data set was distributed by author and chapter, so it would be possible to make such splits based on author information. In addition, all works in the test set were also present in the training set, meaning that results could be over-estimated since any classifier would likely be better at classifying works and authors seen in the training data, than any other works and authors. We also believe it is more important to test the performance on challenging works. To address these issues, we made the following changes in SLäNDa v2.0:

1. We separated the original test data from SLäNDa v1.0 into three separate sets, based on the marking of speech, *QM* for works using quotation marks, *Dash-v1* for works marking speech with a dash, and *Unmarked-v1* for texts mainly without any speech markers, which is the case for Bergman.
2. We annotated two completely new test sets, without overlap with the training data in authors or works.
    - *Unmarked-v2* contains works that do not have any standard typographical marking of speech, and in most cases use no markings at all.
    - *Dash-v2* contains works that use standard dashes, i.e. use them only at the beginning of speech segments.

Test set	S-segments	S-tags	Authors
Quotation marks	161	72	3
Dash-v1	140	77	4
Unmarked-v1	26	25	1
Dash-v2	886	323	4
Unmarked-v2	581	331	4

Table 2: Test sets in SLäNDa version, with number of speech segments and speech tags, and the number of authors in each set.

Table 2 summarizes the size of these test sets. Note that Unmarked-v1 is very small, and only contains texts from one author, Bergman, which makes it a bit limited. We did want to keep it, though, in order to keep the original test data from SLäNDa v1.0. However, we think that the new challenging unmarked-v2 test set is much more important, since it does not overlap with the authors in the training data, and also is considerably larger.

3.4. Annotation Scheme

The main goal of SLäNDa is to annotate direct speech and narrative. Only direct speech is annotated, we do not annotate indirect speech.[5] We also distinguish the main narrative by marking everything which do not belong to it, such as signs or letters. For speech we do a more detailed annotation, also including speech tags, and speaker information. If the speaker is explicitly mentioned in the speech tag, the mention is annotated. In case the mention is a pronoun, or other non-exact descriptions, such as 'a voice' in example (3), the annotators also resolve the reference to a known character of the novel (*Alice*, in (3)), or mark it as unknown if it is not possible to resolve it. In case there is not a speech tag with a specific mention of the speaker, the speaker annotation is added as additional information on the speech segment. We also mark the order of speech tags with respect to the speech segment(s). Here we added a new category *medial*, to be used for cases where the speech tag is in between two parts of a speech segment (as in examples (5)–(6)). In SLäNDa v1.0, we only used tags for before and after the first part of a speech segment. This conversion could be done automatically, and was checked during quality control.

For speech tags we had one category in SLäNDa v1.0. However, we noted that there were a lot of cases where speech was introduced, but not by a traditional speech tag. For those cases we added a new category, *Other type of speech tag*. Our formal criterion of a standard speech tag, based on Teleman et al. (1999) and Semino and Short (2004), is that it should contain a verb signalling speech, as 'said' or 'shouted', and that the speech tag should consists of a single clause, not more. The reported clause related to the verb should syntactically be a direct object, as in example (11), from Benedictsson, p. 284, where the speech verb 'answered' is used.

(11) ≫Han får aldrig säga mamma, han ska säga *mor*≫, svarade Marianne.
'He may never say mum, he should say *mother*, Marianne answered.'

However, we do have quite a few cases where the formal criteria for speech tags are not satisfied, but where the function is the same, as in example (12), from Heidenstam, p. 139. According to a strict syntactic criterion there is no verb here directly signalling speech, only 'sent', and the speech is not a direct object to the verb. But as a reader we perceive 'sent him the ... question' as parallel to 'asked him'. For these cases we introduced the category *Other type of speech tag*. The reported speech is not affected in these cases, it is still considered a standard speech segment. Semino and Short (2004, p. 39) also discussed this type of border line case, and they use the tag "functions as NRS" in their corpus for examples as: *She turns to me, 'What do you think?'*.

(12) Scheik Ibrahim skickade honom brådskande och halft hviskande följande ängsliga fråga: – Har du räknat pengarna?
'Scheik Ibrahim hastily and almost whispering sent him the following nervous question: – Have you counted the money?'

The full list of first level annotation categories is shown in the first column of Table 3.[6] *Other* was intended for cases not covered by the named categories, and the annotators were asked to define the type if used. However, it was only used by mistake twice in SLäNDa v1.0, and removed during quality control for SLäNDa version 2.0.

3.5. Annotation Process

The annotation was performed using the WebAnno graphical web-based annotation tool (Eckart de Castilho et al., 2018). WebAnno is a freely available tool, which supports a number of formats both for input and output, and allows customization of annotation schemes (Yimam et al., 2013; Yimam et al., 2014). For input we used the WebAnno text-based input format, which treated our light XML-format as text, which worked well for our purpose. The text was segmented into paragraphs, and displayed as such to the annotators, since speech segments are typically contained within single paragraphs. We exported into the WebAnno TSV 3 format, see details in Section 3.6.

[5] While indirect speech is not in focus in the current version of SLäNDa, it is also interesting, and could be added in a future extension.

[6] Swedish terms are used in the actual SLäNDa annotation. We use translated terms in this paper. There is a mapping of terms in the SLäNDa 2.0 documentation.

≫	B-SPE	Speech[4]	1
Du	I-SPE	Speech[4]	0
borde	I-SPE	Speech[4]	0
gifta	I-SPE	Speech[4]	0
dig	I-SPE	Speech[4]	0
≫	B-SPE	Speech[4]	1
,	O	_	0
sade	B-TAG	Speech tag[5]	0
Börje	I-TAG	Speech tag[5]||Speaker[6]	0
.	O	_	0

Figure 1: An example of the IOB-format used for identifying speech segments and speech tags. Text from Benedictsson, p. 309: '≫You should get married≫ , Börje said.'

The SLäNDa v1.0 annotation scheme was developed after discussions in a cross-disciplinary team of researchers from Computational Linguistics, Literary Studies and Scandinavian Studies, followed by a round of pilot annotations, not used in the final corpus, followed by additional discussions. After this process, the final guidelines were produced, and three annotators were trained. The guidelines for version 2.0, were further discussed in the cross-disciplinary team, and slightly adapted as described in section 3.4. For version 2.0, there was a single annotator, who had been involved in all the steps above, and thus knew the annotation scheme very well. The annotator, the second author of this paper, is a researcher in Scandinavian Languages specializing in the analysis of literary texts, and a native Swedish speaker. After the automatic conversion of the placement of speech tags, this annotator went through the texts in SLäNDa v1.0, to control for mistakes and inconsistencies, as well as to adapt to the minor changes in annotation scheme. All new texts were also annotated by this annotator.

In Stymne and Östman (2020) we showed the inter-annotator agreement for SLäNDa v1.0. Overall, the annotators agreed to a high extent on the main classification, with Kappa values of 0.72 and 0.83 for two pairs of annotators. There were no mismatches between the categories, only a few cases where one annotator had missed a segment annotated by the other annotator. The speaker identification had a few more issues, mainly because one annotator did not resolve some pronouns and had a higher tendency to mark speakers as unknown. This is a known issue from other similar annotation projects; Elson et al. (2010) also noted that all speakers were not identified in there corpus. As SLäNDa v2.0 was annotated by a single annotator, we do not show any further agreement numbers. We do believe that the quality has been improved compared to version 1.0, though, thanks to quality control and to a well-trained expert annotator.

3.6. Formats and Licence

The format used in SLäNDa v1.0 is the WebAnno TSV format,[7] which is a tab-based format with information about each token in columns, native to the WebAnno tool we used for annotation, see Stymne and Östman (2020) for details and an example. The XML-annotations were tokenized in the TSV output, as were ellipsis, and we performed post-processing on these, to treat them as single tokens.

We keep the TSV format as the main format for SLäNDa v2.0 as well, without any changes. We provide the annotations for each chapter, as well as concatenated into a training set, and the five test sets described in section 3.3. In addition, we also convert SLäNDa into a new format, an IOB-based format for the token-level identification of speech segments and speech tags. The IOB-based format, exemplified in Figure 1, is tab-based, where on each line we present the token, an IOB tag, the original TSV tag, and an indication if the token is a quotation mark or dash which should be removed in a stripped version. Note that quotation marks and dashes also occur in other cases than for speech marking, so we have identified those cases where they occur at the beginning or end of a speech segment by a set of simple rules. Since the focus in this case is on speech identification, we keep the IOB-tags rather sparse, and group all annotations other than speech segments and speech tags into a single other category. This means that we have the following three categories as main tags:

1. SPE: speech segment
2. TAG: speech tag
3. OTH: other annotation type.

We mark the first tokens of an occurrence with "B", as "B-SPE" for the beginning of speech, and all following tokens of the occurrence with "I", as "I-SPE" for inside speech. The "O" tag is used for other, in our case the narrative, which is the part of the text without any annotations. Note that marking the quotation marks (with 1), allows easy filtering into stripped data.

[7] https://webanno.github.io/webanno/releases/3.4.5/docs/user-guide.html#sect_webannotsv

	SLäNDa v1.0		SLäNDa v2.0					
Type	Training	Test	Training	QM	Dash-v1	Unm.-v1	Dash-v2	Unm.-v2
Speech segment	1653	326	2051	160	137	26	886	581
Speech tag	783	171	930	71	76	25	323	331
Other type of speech tag	–	–	33	1	4	–	5	6
Embedded speech	5	–	5	–	–	–	–	–
Thought	38	4	46	7	1	2	14	15
Quotation	11	–	11	–	–	–	4	4
Letter	7	4	8	–	4	–	8	–
Sign	2	–	1	–	–	–	–	–
Other	2	–	–	–	–	–	–	–

Table 3: Summary of the number of annotations of different types, in the test and training sets of SLäNDa version 1.0 and version 2.0.

SLäNDa version 2.0 is publicly released in the LINDAT/CLARIAH-CZ repository (Stymne and Östman, 2022) under the under the Creative Commons licence CC BY-NC-SA.[8]

3.7. Statistics

Table 3 summarizes the number of annotations in SLäNDa version 1.0 and 2.0, splitting the data into the training and test sets of each version. We can clearly see that all annotations except speech segments and speech tags are quite rare. Of the other categories, only thoughts occur in all partitions. The new variant of speech tag that we added to SLäNDa v2.0 is considerably rarer than standard speech tags, but do occur in all splits except the small Unmarked-v1 test set. The total size of SLäNDa version 2.0 is 274,704 tokens.

4. Pilot Experiments

In our pilot experiments, we focused on the task of distinguishing speech segments and speech tags from other text, i.e. from the main narrative. Our main purpose is to investigate how useful typographical cues, like quotation marks and dashes are in identifying speech. In order to investigate this issue, we used two variants of the training data, one with the original text, and one where all dashes and quotation marks indicating speech are stripped, so that the text resembles text written without any speech markers. To balance the training data, we only kept 50% of the lines without any speech. To evaluate the effect of typographic markers, we used our test set splits, with different variants of markings. For the test sets with quotation marks and dashes, we also created a stripped variant. This enables us to investigate the effect of typographical marks both in training data and in test data. In all experiments, we stripped away 10% of the training data, to be used as development data.[9]

We use a standard model for sequence labeling based on the T-NER toolkit (Ushio and Camacho-Collados, 2021), fine-tuning the Swedish BERT model KB-BERT (Malmsten et al., 2020). T-NER follows Devlin et al. (2019) and use a linear layer on top of the last BERT layer and a cross-entropy loss, implemented using the Huggingface transformer model (Wolf et al., 2020). We use the default parameters from T-NER, with a learning rate of 1e-5, wight decay of 1e-7, batch size 32, and a total number of 5,000 steps with a warmup period of 700.[10] We run each experiment three times with different random seeds, and report the average score. For the evaluation we use exact match of each speech segment and speech tag and report the F1-score. This is a harsher metric than token-level match, used in some previous work, e.g. in Ek and Wirén (2019).

Table 4 shows the results of this study. Not surprisingly we get high results for identifying speech segments with typographical markers, when training with the original matching data, whereas the identification of speech tags seems more challenging. The most informative marking, quotation marks are more helpful than dashes on speech identification, but not on speech tags. When we train on stripped data, the performance goes down considerably, since the system has not been able to learn the typographical cues. The very low score of .3 for speech segments with quotation marks is mainly due to the metric requiring matching of the full sequence, including the quotation marks, which have not been seen during training. Token-level accuracy is also affected, but to a smaller extent. Also note that while the scores for speech tag identification drops, it does not drop dramatically. The performance on the dash-v2 test set has higher scores than dash-v1, but follows the same pattern.

For the new unmarked-v2 test set, the pattern is, as expected, that the scores are higher when using the stripped training data than with the original training data. The identification of speech segments is

[8] Creative Commons, Attribution-NonCommercial-ShareAlike 2.5 Sweden: `https://creativecommons.org/licenses/by-nc-sa/2.5/se/deed.en`

[9] We converted data from the SLäNDa IOB-format into the input format of T-NER. We also classified the *Other* class of the IOB-format, be we do not report those results here, as this class is rare, and that is not the main point in focus. All data sets used in these experiments are also available with the released SLäNDa v2.0.

[10] Note that our goal is not to reach state-of-the-art performance, but to investigate the effect of typographical markers.

	Original training data		Stripped training data	
	Speech	Tags	Speech	Tags
Quotation marks	93.3	66.3	0.7	64.0
Dash-v1	84.0	71.3	17.3	63.7
Quotation marks (stripped)	59.3	60.3	88.7	70.7
Dash-v1 (stripped)	58.7	60.3	68.0	65.7
Unmarked-v1	45.0	38.0	36.3	33.7
Dash-v2	90.3	80.3	17.3	73.0
Dash-v2 (stripped)	63.0	68.0	74.0	74.7
Unmarked-v2	72.7	70.0	75.3	73.7

Table 4: F1-scores for the prediction of speech segments and speech tags.

worse than for the test sets with typographical marking, when they are trained with matching data, indicating that identifying unmarked speech is indeed more challenging than typographically marked speech. The unmarked-v1 test set does not show the expected pattern, though, and has very low scores with both training sets, but lower with stripped data. We believe this could be due to the fact that it is both very small, and, more importantly, only contains data from Bergman, who is also present in the training data. We believe that with the original data, the classifier mainly relies on the works of Bergman, which is the most similar, and that in this case, adding artificial stripped data from other authors rather confuses the classifier. We thus think that the unmarked-v1 test set should be used with caution, and recommend mainly relying on unmarked-v2 for originally unmarked text.

The three stripped test sets follow the pattern of unmarked-v2; the performance is low with the original training data, but considerably improves with stripped training data. The difference in performance is even more striking than for unmarked-v2, with very low scores with original data, and larger improvements with stripped data. The stripped quotation marks set has the highest score for speech segment identification with stripped training data. We find this unexpected, since we would expect a lesser need to mark speech textually when the typographical marking is more informative, and a bigger need for textual marking when there are no typographic clues. The scores for the stripped dash sets do fall below unmarked-v2, though, and unmarked-v2 has the highest scores among these test sets for speech tag identification. It is also worth noting that the performance on the stripped test sets with stripped training data is overall considerably below that of the original test sets with original training data, especially for speech identification, further showing that identification without markers is more challenging than without, even when only starting dashes are used.

5. Conclusion and Future Work

We have presented SLäNDa version 2.0, a corpus of Swedish literary texts with annotations of narrative and everything not belonging to the main narrative, with a special focus on speech. We have performed quality control on SLäNDa version 1.0, updated the annotation scheme slightly, and extended it with a number of new texts. A specific focus has been on enabling fine-grained evaluation of the separation of speech, speech tags, and narrative. We have created dedicated test sets for works that mark speech with either quotation marks or dashes, as well as works that do not use any consistent marking, and mainly has no marking at all.

In a pilot experiment, we explored the impact of typographical marking of speech, on the identification of speech segments and speech tags. We explored this by using our dedicated test sets, as well as stripping typographical markers from the training data. As expected, we had the best results on data with typographical marking with matching training data. However, even for our challenging test sets, we could in most cases improve the results by using stripped training data without any typographic markers.

We presented a first attempt at separating speech segments and speech tags from the main narrative. In future work we plan to improve such classification further, especially for the challenging cases with limited typographical markers. We also want to compare the current BERT model to other architectures, including simpler baselines. Our final goal is to use these classifiers for investigating changes in the Swedish written language during the 19th and early 20th century, with the hypothesis that change happened earlier in speech than in the narrative. Automatic classifiers will allow large scale investigations, as a contrast to earlier smaller scale work on the theme, such as Engdahl (1962).

In addition to the gold data annotated in SLäNDa, we plan to automatically annotate speech segments in works that exclusively use quotation marks, as additional silver standard training data, as was done for English by Kurfalı and Wirén (2020). As shown in our pilot experiments, such data can also be stripped from quotation marks, to be used to simulate more challenging ways of marking speech. We will also use SLäNDa for training classifiers for other type of non-narrative, and for the identification of speakers.

Acknowledgements

This work is funded by the Swedish research council in project 2020-02617: *Fictional prose and language change. The role of colloquialization in the history of Swedish 1830–1930*, PI David Håkansson. We thank David Håkansson and Johan Svedjedal for insightful discussions. We thank Litteraturbanken (The Swedish Literature Bank) for making such a large and diverse collection of literary works available. Computations were enabled by resources in project UPPMAX 2020/2-2 at the Uppsala Multidisciplinary Center for Advanced Computational Science.

6. Bibliographical References

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

Boyd-Graber, J., Hu, Y., and Mimno, D. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval*, 11(2-3):143–296.

Brunner, A., Tu, N. D. T., Weimer, L., and Jannidis, F. (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.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ek, A. and Wirén, M. (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.

Ek, A., Wirén, M., Östling, R., N. Björkenstam, K., Grigonytė, G., and Gustafson Capková, S. (2018). Identifying speakers and addressees in dialogues extracted from literary fiction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Elson, D. and McKeown, K. (2010). Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, pages 1013–1019, Atlanta, Georgia, USA.

Elson, D., Dames, N., and McKeown, K. (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, July. Association for Computational Linguistics.

Engdahl, S. (1962). *Studier i nusvensk sakprosa. Några utvecklingslinjer*. Skrifter utgivna av Institutionen för nordiska språk vid Uppsala universitet, Uppsala.

Håkansson, D. and Östman, C. (2019). ”afbröt skolläraren ifrigt”. en diakron studie av anföringssatsen i svensk skönlitteratur. *Samlaren. Tidskrift för forskning om svensk och annan nordisk litteratur*, 140:261–280.

Hundt, M. and Mair, C. (1999). “Agile” and “Uptight” genres: The corpus-based approach to language change in progress. *Journal of Corpus Linguistics*, 4:221–242.

Jannidis, F., Konle, L., Zehe, A., Hotho, A., and Krug, M. (2018). Analysing direct speech in German novels. In *Abstract zur Konferenz Digital Humanities im deutschsprachigen Raum 2018*, pages 114–118, Cologne, Germany.

Kurfalı, M. and Wirén, M. (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, December. International Committee on Computational Linguistics.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with words at the National Library of Sweden - making a Swedish BERT. *CoRR*, abs/2007.01658.

Papay, S. and Padó, S. (2020). RiQuA: A corpus of rich quotation annotation for English literary text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 835–841, Marseille, France, May. European Language Resources Association.

Quintão, M. E. (2014). Quotation attribution for Portuguese news corpora. Master’s thesis, Técnico Lisboa/UTL, Portugal.

Reiter, N., Willand, M., and Gius, E. (2019). A shared task for the digital humanities chapter 1: Introduction to annotation, narrative levels and shared tasks. *Journal of Cultural Analytics*, 12.

Semino, E. and Short, M. (2004). *Corpus Stylistics. Speech, writing and thought presentation in a corpus of English writing*. Routledge, London.

Stymne, S. and Östman, C. (2020). SLäNDa: An annotated corpus of narrative and dialogue in Swedish literary fiction. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 826–

834, Marseille, France, May. European Language Resources Association.

Teleman, U., Hellberg, S., and Andersson, E. (1999). *Svenska Akademiens grammatik*. Norstedts Ordbok, Stockholm, Sweden.

Teleman, U. (2003). *Tradis och funkis. Svensk språkvård och språkpolitik efter 1800*. Norstedts Ordbok, Stockholm, Sweden.

E. N. Tigerstedt, editor. (1956). *Ny illustrerad svensk litteraturhistoria. Del 3*. Natur och kultur, Stockholm, Sweden.

Ushio, A. and Camacho-Collados, J. (2021). T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online, April. Association for Computational Linguistics.

Willand, M., Gius, E., and Reiter, N. (2019). A shared task for the digital humanities chapter 3: Description of submitted guidelines and final evaluation results. *Journal of Cultural Analytics*, 12.

Wirén, M., Ek, A., and Kasaty, A. (2020). Annotation guideline no. 7: Guidelines for annotation of narrative structure. *Journal of Cultural Analytics*, 1.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Yimam, S. M., Gurevych, I., Eckart de Castilho, R., and Biemann, C. (2013). WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria, August. Association for Computational Linguistics.

Yimam, S. M., Biemann, C., Eckart de Castilho, R., and Gurevych, I. (2014). Automatic annotation suggestions and custom annotation layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96, Baltimore, Maryland, June. Association for Computational Linguistics.

7. Language Resource References

Eckart de Castilho, R. and Banski, P. and De Boer, M. and Klie, J.-C. and Krause, T. and Nothman, J. and Pfeiffer, W. and Winchenbach, U. (2018). *WebAnno*. `https://webanno.github.io/webanno/`.

Stymne, S. and Östman, C. (2020). *SLäNDa: An Annotated Corpus of Narrative and Dialogue in Swedish Literary Fiction*. LINDAT/CLARIAH-CZ Repository, `http://hdl.handle.net/11372/LRT-3169`.

Stymne, S. and Östman, C. (2022). *SLäNDa Version 2.0: Improved and Extended Annotation of Narrative and Dialogue in Swedish Literature*. LINDAT/CLARIAH-CZ Repository, `http://hdl.handle.net/11372/LRT-4739`.