Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 826–834
Marseille, 11–16 May 2020

SLäNDa: An Annotated Corpus of Narrative and Dialogue in Swedish Literary Fiction

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

Abstract

We describe a new corpus, SLäNDa, the Swedish Literary corpus of Narrative and Dialogue. It contains Swedish literary fiction, which has been manually annotated for cited materials, with a focus on dialogue. The annotation covers excerpts from eight Swedish novels written between 1879–1940, a period of modernization of the Swedish language. SLäNDa contains annotations for all cited materials that are separate from the main narrative, like quotations and signs. The main focus is on dialogue, for which we annotate speech segments, speech tags, and speakers. In this paper we describe the annotation protocol and procedure and show that we can reach a high inter-annotator agreement. In total, SLäNDa contains annotations of 44 chapters with over 220K tokens. The annotation identified 4,733 instances of cited material and 1,143 named speaker–speech mappings. The corpus is useful for developing computational tools for different types of analysis of literary narrative and speech. We perform a small pilot study where we show how our annotation can help in analyzing language change in Swedish. We find that a number of common function words have their modern version appear earlier in speech than in narrative.

Keywords: literary corpora, direct speech, narrative, dialogue, annotation

1. Introduction

In literary studies, as in many other research fields, there has been a trend towards using computational methods. Traditionally, most research was based on close reading of books by researchers, whereas distant reading (Moretti, 2000), the computational treatment of potentially large amounts of literary text, has become an important complement. While unsupervised methods, like topic modelling has been important in such studies, supervised models relying on annotated data are also often required. Annotated data is also useful for evaluation of proposed models, regardless of their type.

In this paper, we describe version 1.0 of SLäNDa, the Swedish Literary corpus of Narrative and Dialogue (Stymne and Östman, 2020). It is made up of eight Swedish literary novels from the late 19th and early 20th centuries, annotated mainly for different aspects of dialogue. The full annotation also contains other cited materials, like thoughts, signs and letters. The main motivation for including these categories as well, is to be able to identify the main narrative. Such a corpus can be useful for training and evaluating computational models, which in turn allows larger scale studies of cited materials, narrative, and dialogue in literature. Annotated literary corpora are quite rare, especially for Swedish, and as such we believe that SLäNDa could become a useful resource.

We consider cited materials to be all parts that stands out from the main narrative of the novels. The most common type is direct speech from the different characters, forming dialogues, but it is not uncommon to have other types of cited material, such as thoughts of characters, signs, letters, and quotations. These parts of a novel are often different from the main narrative. The language can have differences to the main narrative, for instance, characters may have different speaking styles and dialects. It is important to be able to separate these parts for many types of studies, for instance focusing on the narrative when studying the overall plot of a novel, or for comparing the literary traditions of designing the characters' direct speech during different periods. In a linguistic perspective the corpus will enable a comparison between the narrative and the direct speech regarding for example oral style and colloquial expressions. A specific theory of interest to us is that language change happens earlier in dialogue than in the narrative (Engdahl, 1962). The design of the lines and the speech tags is also closely related to the narrative technique in the novel — it may among other things throw light on the role of the narrator, see for example Allison (2018), who studied Dickens' use of speech tags in a narrative perspective.

Our main focus is on dialogues, i.e. sequences of direct speech.[1] For each speech segment, we also mark its speech tag, if any, and the speaker. The speech tag is a description of the speaker, often in the form "someone said", but it can also contain further description e.g. of how something is said, or of events happening during the dialogue. Speech tags have also been called "narrative constructions" in characters' discourse (Ek and Wirén, 2019) and "narrator's report of speech" (Semino and Short, 2004). Speech tags form a part of the narrative; they are not part of characters' speech. An example speech segment from Sandel (p. 41) is shown in (1),[2] where there is a speech tag, containing an indirect reference to the speaker ('said a voice' where the context reveals that the voice belongs to the character Alice) as well as a description of its context ('between a couple of coughs'). The remaining part of this segment is

[1] Indirect speech is not in focus in the current version of SLäNDa, but is also interesting, and could be added in a future extension of our corpus.

[2] We use our translations from Swedish to English in all examples.

the speech by Alice.

(1) – Järnet vill inte bli varmt, sade en röst mellan ett par hostningar. Vi har så litet ved. Men jag skyndar mig. '– The iron will not be warm, said a voice between a couple of coughs. We have so little wood. But I will hurry up.'

Even identifying speech segments might not be trivial, since different authors have different ways of marking them, see Section 3.2. Further, it can also be hard to distinguish where speech tags starts and ends.

In this paper we will describe the design and creation of SLäNDa. In total, it contains annotations of 44 chapters with over 220K tokens. The annotation identified 4,733 instances of cited material and 1,143 named speaker–speech mappings. We also present a small pilot study where we investigate language change in literature, comparing speech segments to narrative, focusing on common function words, which have old-fashioned and modern variants.

2. Related Work

Ek et al. (2018) and Ek and Wirén (2019) performed studies of dialogues in Swedish novels, where they tried to identify the speaker and addressee (Ek et al., 2018)[3] and speech tags (Ek and Wirén, 2019). In these papers they create classifiers to automatically identify these aspects. In order to do this, they annotated excerpts from four Swedish novels, partially overlapping with our selection. However, the main focus in these papers was on the technical aspects of the classifier for these tasks rather than on the annotations and the annotation process.

Ek and Wirén (2019) investigated how the distinction between speech tags and speech can be automatically predicted. They found that a reasonable baseline, that speech tags start with a speech verb and end with punctuation, does not perform well, with an F-score of only 47.9. They proposed a more advanced method based on logistic regression, which performed considerably better with an F-score of 80.8. These results show that distinguishing speech from speech tags in Swedish is far from trivial, and further motivates the annotation of more material that covers this distinction.

The above articles do not specify any specific guidelines for annotation. However, they are related to the guidelines for English specified in Wirén et al. (2020). These guidelines stem from a larger initiative called Systematic Analysis of Narrative Texts through Annotation (Reiter et al., 2019). As part of this initiative, eight teams proposed a set of guidelines for narrative levels in literary texts, which were than evaluated (Willand et al., 2019). The focus here 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).

Håkansson and Östman (2019), focus on the speech tag, however, from the view of literary studies rather than from a computational linguistics perspective. They studied the structure and position of the speech tag in Swedish literary fiction, in a diachronic perspective. They compared Swedish novels from 1800-1900 to newer novels from 1976-1999. The speech tags were identified automatically using simple pattern matching based on quotation marks, possibly followed by a comma, followed by a verb in the present or past tense. This selection was noisy, and after a manual check, approximately 20% were excluded. The main result of the study was that the speech tag in the modern material is much less varied, the verb 'say' dominates, the tag is shorter and the final position dominates to a greater extent. A limitation of this study was that it only captured speech tags accompanying direct speech marked with quotation marks, which excluded a large number of works that marked speech in other ways.

Another study where the lack of reliable tools for identifying narrative, in contrast to speech, and the speaker of an utterance, as well as the speech tag, was problematic, is Stymne et al. (2018), where we tried to analyse style breaks in the novel *Kallocain* by Karin Boye. Due to the lack of such tools, we had to focus most of the study on the main narrative, which could be identified relatively reliable based on paragraph breaks, but which excluded more than necessary of the text.

Dialogue in English literature has previously been analysed on a large scale, especially the attribution of speakers to speech. Elson et al. (2010) studied the automatic identification of social networks in text, through extracting speakers engaged in dialogue. They could identify speech easily, since quotation marks were used as speech markers, even though, as they note, quotes are also used for other purposes, so the extraction was noisy. They also annotated the speakers of over 3,000 utterances, using Mechanical Turk, and used this to perform automatic quote attribution. He et al. (2013) and Muzny et al. (2017) both focus on quote attribution, and create a corpus for this task. QouteLi3 (Muzny et al., 2017) is a corpus of over 6,000 literary quotes from three novels by two authors, linked to both speakers and mentions. Muzny et al. (2017) also discussed and compared these three corpora.

Also for English, Semino and Short (2004) describe an annotated corpus of speech, writing, and thought in prose fiction, newspapers, and autobiographies, covering 250,000 words. In contrast to many other studies, they annotated speech tags in addition to speech segments. The corpus is manually annotated, and is based on a scale of speech and thought presentation categories, presented by Leech and Short (1981, p. 10). The purpose of creating the corpus was to compare the variation in discourse structure between the three narrative genres. In the process of creating the corpus they used two main types of criteria, formal and structural as graphology on one hand, and more pragmatic, contextual criteria on the other hand. They also discussed embedded speech. The only discourse category they could capture solely with formal criteria was direct speech, as it was marked by citation marks. Unfortunately this is often not the case with direct speech in Swedish fiction, see Section 3.2.

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

Author		Novel	Year	Marker	Total	Annotated
August Strindberg	AS	*Röda rummet*	1879	–	29	2
Victoria Benedictsson	VB	*Fru Marianne*	1887	» »	17	10
Oscar Levertin	OL	*Magistrarne i Österås*	1900	– (–)	12	2
Hjalmar Söderberg	HS	*Martin Bircks ungdom*	1901	» »	32	10
Selma Lagerlöf	SL	*Körkarlen*	1912	» »	12	4
Maria Sandel	MS	*Hexdansen*	1919	–	9	2
Hjalmar Bergman	HB	*Chefen fru Ingeborg*	1924	Mixed (:)	30	10
Karin Boye	KB	*Kallocain*	1940	–	20	4

Table 1: Authors and novels in the corpus, with the publication year, preferred speech marker, and number of chapters in the full novels, as well as in SLäNDa.

3. Texts

SLäNDa contains texts from eight novels by eight different Swedish authors, published between 1879 and 1940, see Table 1 for an overview. This period is regarded as a period of modernization for the Swedish language in syntax and vocabulary; thus we wanted a selection of texts where these changes might be illuminated. It has been suggested, but not thoroughly investigated, that literature, and especially literary dialogue, was driving this change (Engdahl, 1962). We also wanted both male and female authors, as well as authors representing different literary traditions and using a mix of speech markers.

For each novel we have selected between 2–10 chapters for annotation. We choose to annotate full chapters because we wanted clearly defined units of contiguous text. The chapters are sampled throughout the novels, and are not in contiguous order. Out of these, one annotated chapter per novel are held out as the test set of the corpus, to be used in future computational studies. For novels with at least eleven chapters, we choose chapter eleven for the test corpus, for the remaining novel by Sandel, we chose chapter two. The remaining chapters form the training part. We choose to annotate a higher number of chapters for three authors, one early author, one in the middle time span, and one of the later authors. The number of chapters annotated for each novel is shown in Table 1. All together, SLäNDa contains 220,941 tokens, which is much larger than the 40,623 tokens used in Ek and Wirén (2019) and nearly as large as Semino and Short (2004). While we believe this version of SLäNDa already to be useful, we plan to release later versions containing a more balanced sample from each author, as well as more authors.

3.1. Formats and Licence

All texts were downloaded from Litteraturbanken (The Swedish Literature Bank),[4] a collection of Swedish literary texts from the 19th and 20th centuries. The texts in our sample are released under the CC-BY-NC-SA licence[5]. Their texts come in an XML format that describes the printed version of the books, including features such as page breaks and hyphenation of words. We are mainly interested in the running text, and thus preprocess the text to a custom xml-format, containing only minimal markup. The markup left shows titles, chapters, minimal markup such as italics, and divides the texts into paragraphs. In most cases, speech segments, and other cited materials, are kept within single paragraphs.

SLäNDa is also released under the CC-BY-NC-SA licence. SLäNDa version 1.0, which is described in this paper, is available at `http://hdl.handle.net/11372/LRT-3169`.

3.2. Markers of Speech

In the selected novels, speech are marked in different ways. The two dominating options is to use citations marks surrounding the speech, as in (2) from Benedictsson (p.286), used by three authors, or to use a dash at the beginning of a speech segment, as in (1), used by three authors. These two variants of marking speech are common in Swedish, both historically and in modern texts. Two of the authors use more unusual variants. Levertin also uses a dash, but sometimes, adds it also before and after speech tags, which the other authors usually do not, see (3) for an example (Levertin, p. 134). Finally, Bergman mostly uses no marker at all, as in (4), Bergman p. 94. Speech tags can also precede the speech followed by a colon. He also sometimes, but far from always, marks the end of the speech segment with a dash, instead of the beginning, as in (5), Bergman (p. 248). The different options for speech markers are summarized in Table 1.

(2) »Och så vill jag be er resa», tillade hon utan att lyfta sina ögon.
'»And so I would like to ask you to travel», she added without lifting her eyes'

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

(4) Det är inte sant! utbrast fru Ingeborg upprörd. Du sitter och ljuger på din egen mor!
'It is not true! Mrs Ingeborg exclaimed angrily. You are lying on your own mother!'

(5) Fru de Lorche klappade henne på handen och svarade: Vad du gör för Louis, det gör du mig! –
'Mrs de Lorche tapped her hand and answered: What you do for Louis, you do for me! –'

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

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

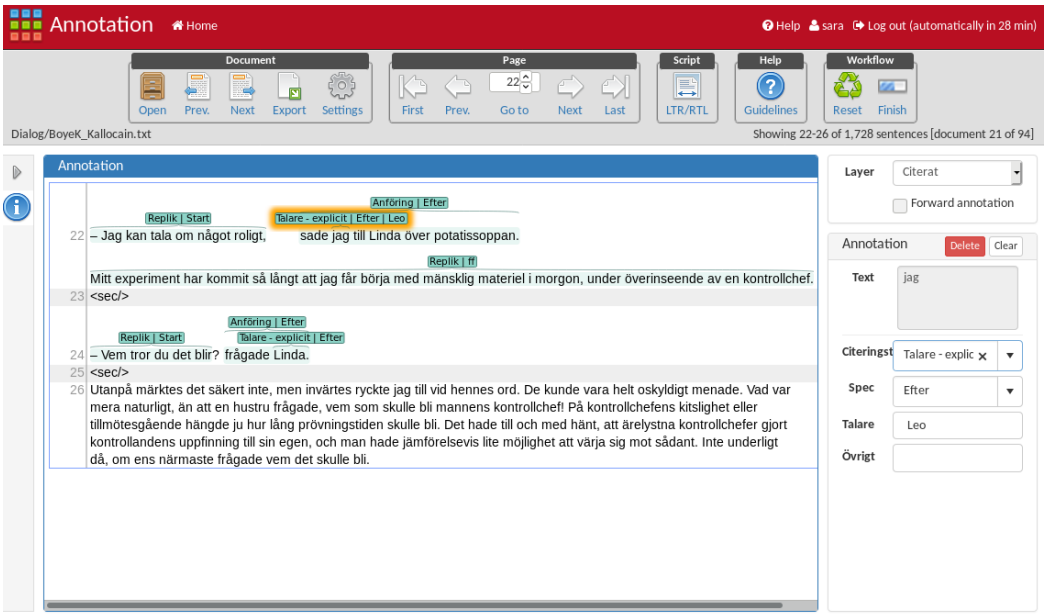

Figure 1: The WebAnno annotation interface, for annotation of Boye (p. 13).

4. Annotation

In this section we will describe in more detail what we annotate, and how we label our annotations, the annotation process, the format of the annotated corpus, and inter-annotator agreement.

4.1. Annotation Protocol

The main purpose of the annotation is to annotate all segments of cited material, which are not part of the main narrative of the text, as well as speech tags, identifying speakers, and sometimes providing further information about them. In addition we mark speakers when applicable. Note, that this means that we also get a reliable annotation of narrative, since we annotate everything that is not part of the narrative.

At the highest level, we mark all words in a text segment containing cited material, or being a speech tag. The annotation is then given one of the following labels:

- Speech
- Speech tag
- Embedded Speech
- Thought
- Sign
- Letter
- Quotation
- Other

Note that *quotation* is used for references to text outside of the novel, not synonymously with utterances from characters in the novel, which often are referred to as quotes in computational linguistics papers (Muzny et al., 2017, among others), which we call speech (segments) in this paper. The remaining types include text from signs of different types, letters, and thoughts. Embedded speech is speech that is quoted within a normal speech segment. The *Other* tag is used when the annotator did not find any of the other tags suitable. They were instructed to describe the type in free text, if they found any such instances. This tag was rarely used, though.

Speech, and to some extent thoughts, are often accompanied by a speech tag. The speech tag could occur after the speech (2), in the middle of a speech segment (3) or before the speech (5). To account for all these cases, speech tags have a further annotation describing if they occur before or after the start of the speech section it refers to. Speech segments, and thoughts, are marked with a second layer to account for segments split by speech tags, with *start* specifying the start of a speech segment (or an unsplit segment), and *cont* specifying the continuation of a previous segment.

Further, we also add annotations for the speaker of each utterance. If this is explicit in the speech tag, the speaker is marked there, identifying the mention. If the speaker is referred to by a pronoun or other indirect description, this should be resolved to a specific speaker. In cases where the speaker is not explicitly annotated, the speaker is marked on the start of the speech segment, by name, or by *unknown* if it is impossible to figure out who the speaker is, which is sometimes the case.

4.2. Annotation Tool and Formats

When choosing an annotation tool for this project we had a number of desiderata. We wanted a graphical user interface, that was easy to use for experts on literature without any advanced technical knowledge. We needed a tool that could handle our initial XML-format with very light markup, and which easily would let us customize our own tag-set, where we could annotate long spans of words, and where we could have nested annotations. And we needed to be able to export the texts in a usable format. We also preferred a web-based tool, so that the annotators could work from any computer.

Using these criteria we identified the tool WebAnno (Eckart de Castilho et al., 2018) to be a good fit (Yimam et al., 2013; Yimam et al., 2014), see Figure 1. While WebAnno could not easily handle our XML format, we could

```
#Text=≫Nej, det fins ingen derinne, svarade Marianne.
50-1    3277-3278   ≫          Speech[4]        Start[4]    *[4]
50-2    3278-3281   Nej        Speech[4]        Start[4]    *[4]
50-3    3281-3282   ,          Speech[4]        Start[4]    *[4]
50-4    3283-3286   det        Speech[4]        Start[4]    *[4]
50-5    3287-3291   fins       Speech[4]        Start[4]    *[4]
50-6    3292-3297   ingen      Speech[4]        Start[4]    *[4]
50-7    3298-3305   derinne    Speech[4]        Start[4]    *[4]
50-8    3305-3306   ,          Speech[4]        Start[4]    *[4]
50-9    3307-3314   svarade    Speech tag[5]    After[5]    *[5]
50-10   3315-3323   Marianne   Speech tag[5]|Speaker[6]    After[5]|*[6]   *[5]|Marianne[6]
50-11   3323-3324   .          _                _           _
```

Figure 2: Example of the TSV3 output format, for the sentence translating as '≫No, there is nobody in there, Marianne answered'. From Benedictsson, p. 284. Categories are translated to English from the original Swedish.

use its plain text line-oriented format. While this did not treat our XML-markup as XML, it was sufficient for our purpose, since the XML tags were not important for the annotation. Using this scheme, each paragraph was treated as a line, and could be annotated as a full chunk, if needed.

For the output format, we use the native WebAnno format TSV 3.[6] It gives the full paragraphs in comments, followed by one word per line, containing the annotation in tab separated columns for different annotation levels, and information on sentence number, word number, and character offsets. Nested annotations are given in the same column, separated by a vertical bar. Each annotation is identified by a unique number. One issue with the TSV format is that it tokenizes XML markup as well as ellipses (. . .). We retokenize this in a post-processing step. Figure 2 shows an example of our final annotation files, which can then be used for different types of further processing. Note the nested annotation, where the word *Marianne* is annotated both as a speaker and as part of the speech tag.

4.3. Annotation Process

The annotation was carried out in two steps. First there was a pilot phase which served to finalize the annotation scheme and guidelines. This was followed by the main annotation phase, which produced our final corpus.

In the pilot phase, the overall goal was discussed in a group of researchers from literary studies, Scandinavian studies, and computational linguistics. When the goals were set, the authors of this paper defined an initial annotation scheme, and found a suitable tool for the annotation. At this stage, six persons with varying background tried out the annotation on small parts of the texts, and discussed the outcome. Based on this discussion, the guidelines were revised to make the task clearer, and one new category, embedded speech, was added, since we had found a need for that. None of the materials annotated during this phase was included in the final corpus, since it was deemed too noisy.

The final annotation was performed by three persons, one researcher in Scandinavian Languages who is a native Swedish speaker, one assistant, and one student in Scandinavian Languages and Literature, at advanced level. Both the assistant and the student were non-native speakers but with a very good command of Swedish. All texts in the final corpus are annotated by one person.

The annotators were given written guidelines, describing the above annotation scheme, and containing a simple flow chart, as well as an oral introduction. They were also instructed to bring any issues up with the annotation group.

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

	$=$	$\approx$	$\neq$	Miss
Speech	124	3	0	28
	16	1	0	3
Speech tag	28	9	0	16
	16	1	0	3
Other types	0	0	0	11
	–	–	–	–
Speakers	101	–	10	6
	3	–	6	6

Table 2: Inter-annotator agreement over the main categories, and for speakers. $=$ is exact match, $\approx$ is matching annotation, but the number of words slightly mismatch, Miss is when an annotation is missing from either annotator, and $\neq$ is when both annotators have annotated a span, but with different labels/speakers. The top row for each category is A1 vs A2 and the bottom row is A1 vs A3.

4.4. Inter-Annotator Agreement

To estimate how hard the annotation task is for the annotators, and thus how reliable our corpus is, we performed a study on the inter-annotator agreement. To do this, annotator 1 (A1) annotated two chapters in parallel with annotator 2 (A2), and two chapters in parallel with annotator 3 (A3). All chapters were from different authors, covering Benedictsson, Bergman, Boye and Söderberg. From this annotation we compare the top-level annotation between our main eight categories, and having no annotation, as shown in Table 2. In addition to exact matches, we also count cases where the words in the annotations overlap, but do not match exactly. This typically happens because it is ar-

Type	Unit	AS	VB	OL	HS	SL	MS	HB	KB	**Total**
	Chapters	1	9	1	9	3	1	9	3	36
Training	Words	2819	52696	2099	9771	9959	9740	19177	9175	155998
	Chapters	1	1	1	1	1	1	1	1	8
Test	Words	3306	5731	1908	353	2742	4240	5540	3189	27009

Table 3: Total number of chapters and words in the test and training parts of SLäNDa. Author given by initials, see Table 1 for full list

bitrary in the guidelines if punctuation marks are included or not. We also compare if the identified speakers match between the annotators.

Because many categories were rare in this material, we collapsed all categories except speech and speech tags into the other group. The only such cases were A1 having annotated 9 instances of embedded speech, and 2 thoughts, that A2 had not identified. Overall, though, there are no mismatches between the different categories. The only errors are cases where one annotator has identified the annotation, and the other annotator has not. In the overwhelming majority of cases, the annotators have marked the same spans, there are only a few cases of overlapping spans. We also calculated the Kappa statistic of overlap between the two pairs of annotators (Carletta, 1996), for the identification of main category. The values were 0.72 for A1 vs A2 and 0.83 for A1 vs A3, which is normally interpreted as substantial and near perfect agreement.

Studying the annotations made only by one annotator, it seems that they are mostly due to mistakes. There are, however, also a few difficult borderline cases. One example is (6) from Söderberg (p. 191), where it is hard to say if the speech is direct or indirect, and if it should be excluded, since we are not concerned with indirect speech. In most cases the difference between direct speech (DS) and indirect speech (IS) have been obvious but for the few borderline cases we have based our classification on a description presented in (Leech and Short, 1981, p. 256). The primary difference between DS and IS is that DS is expected to report faithfully what was uttered and to report with the exact forms of the words. Thus IS is signalled by a lack of speech marks, by the form of a subordinate clause, by the change from first- and second pronouns to third-person, by change of the tense of the verb, from present to past tense, and by the change of deictic adverbs as *here* to *there*. The fact that there are no speech markers used, instead of Söderberg's standard citation marks, points to indirect speech while the remaining deictic adverb points to direct speech. The line contains no pronoun, and the tense of the verb would be possible in direct speech but prototypically changed in an indirect version: *Henrik Rissler had said that there the time had stopped.*

(6) Här har tiden stått still, hade Henrik Rissler sagt.
'Here the time has stopped, Henrik Rissler had said.'

The identification of speakers were more difficult, however. A1 and A2 agreed in the majority of cases, whereas A1 and A3 had more disagreements. The missing speakers is likely mainly due to inattention by the annotators. The mismatches seem to have many reasons. For A1 vs A3, all mismatches are due to that one of the annotators has not resolved a pronoun, which the other has done, which happens also for A1 vs A2 in 4 cases. A2 also has a larger tendency than A1 to mark the speaker as unknown. In only 4 cases, however, A1 and A2 disagree on the identity of the speaker, and in no case A1 and A3 disagree. Note that the number of speakers is smaller than the number of speech segments, because that contains both the start and continuation of speech segments. It seems like a common issue that all speakers are not identified; for instance, Elson et al. (2010) had the same issue when collecting their corpus. Even so, we would like to improve this part of the annotation in future work. In our final corpus, we only kept the annotations from A1, for the doubly annotated texts.

	Training	Test
Total annotations	3950	783
Speech	1356	244
Speech (cont)	297	82
Speech tags	783	171
Thoughts	38	4
Other types	31	4
Speakers	1356	244
Named	1001	142
Pronouns	31	21
Unknown	208	33
Unmarked	106	48
Unique speakers	68	36

Table 4: Basic statistics on the number of annotations in the training and test portions of SLäNDa.

5. Statistics

Table 3 shows the size of the training and test parts of SLäNDa, in total and for each author. We think the size of the corpus is reasonably large, with over 155K words in the training part and 27K words in the test part. Words are defined as all tokens except punctuation marks. The total size of the corpus is over 183K words and 220K tokens.

Table 4 shows the number of annotations across all authors in the training and test parts and Table 5 shows the number of annotations per author in the training corpus. Overall it can be seen that we have a reasonable number of annotations for the categories related to speech that is our main interest, both in total and per author. Embedded speech is relatively rare, though, and only occur in Bergman.[7] The other types of cited materials are quite rare,

[7]We found embedded speech also in Strindberg in the pilot phase, but it does not occur in the chapters from Strindberg currently in SLäNDa.

	AS		VB		OL		HS		SL		MS		HB		KB	
Type	#	W	#	W	#	W	#	W	#	W	#	W	#	W	#	W
Speech	74	1486	891	17640	12	356	37	688	69	5138	74	2446	167	5417	33	1969
Speech tag	24	169	432	2216	12	65	41	232	53	373	46	316	148	756	27	164
Embedded	0	0	0	0	0	0	0	0	0	0	0	0	9	151	0	0
Thought	2	14	4	17	0	0	7	737	15	164	3	21	7	93	0	0
Sign	0	0	0	0	0	0	0	0	0	0	1	40	0	0	1	3
Letter	1	9	1	62	0	0	2	658	0	0	3	88	0	0	0	0
Quotation	6	101	2	31	0	0	0	0	0	0	2	67	1	9	0	0
Other	0	0	1	4	0	0	0	0	1	2	0	0	0	0	0	0

Table 5: Statistics of how many annotations (#) and words (W) there are in the training part of the corpus for each category and author. Author given by initials, see Table 1 for full list

often occurring in only one or a few authors. The possible exception is thoughts, which occurs for the majority of authors, but even they are rare compared to speech. As important as the annotated part is the unannotated parts, which consists only of the main narrative, since all cited materials is marked by annotations. The narrative is by far the largest part of SLäNDa.

Table 4 also gives some additional details of speakers. As indicated by the inter-annotator agreement study, annotators have not always identified a speaker of a speech segment (*Unmarked*). They have also failed to resolve in total 52 cases of pronoun reference, which they were also supposed to do. There is also quite a high number of speech segments where the annotators were not able to resolve who the speaker was (*Unknown*). We still believe that the number of instances where the speaker is identified would be enough for training a classifier.[8] There are 68 unique identified speakers in our training data, and 36 in the test data. The speakers are identified by proper names, e.g. *Börje, Martin Birck*, title and name, e.g. *Fru Landén* ('Mrs Landén), occupation e.g. *skådespelerskan* ('the actress'), relationships e.g. *modern* ('the mother'), or other descriptions, typically noun phrases, e.g. *den grekiske musikern* ('the Greek musician').

As noted before, the size of the corpus varies between the authors. Especially Söderberg stands out as having quite short chapters, thus being the smallest part of the test data. The training data from Söderberg has a reasonable size, however, since we have annotated many chapters from this novel. Levertin, on the other hand stands out as having a lower number of annotated segments than the other authors, having less dialogue than the other authors.

6. Pilot Experiments

In a small pilot study we wanted to study the contrast between narrative and speech, with a specific focus on language change. One aspect of the changes in the Swedish language about 100 years ago is that the modernization in syntax, morphology, and lexicon is regarded to first occur in literary fiction (Engdahl, 1962). These changes in fiction are also presumed to first occur in the dialogues, with lines designed in a more colloquial style. As a second step in the transition these orally influenced patterns made their way into the narrative. But so far this process has not been thoroughly investigated. In our pilot study we wanted firstly to see if there is a difference between narrative and speech in our material and secondly to see if we could catch any linguistic changes over time in the material.

In order to do this we extracted all text in the training corpus annotated as speech. We also extracted all unannotated text, which is the remaining narrative, after we had annotated all cited materials. We then compared basic statistics, and the prevalence of old-fashioned and modern variants of some common function words in this material. In calculating phrase length, we followed previous work (Holm, 2015) on the analysis of Swedish literature, and defined a graphical phrase as a stretch of words between punctuation marks, such as comma, period and colon. While this study was performed on our annotated data, we envision that a similar study could be made based on automatically classified data in future work.

The result of the pilot study is shown in Tables 6 and 7. The overall tendency is that there is a difference between the design of the narrative and of the direct speech: words, phrases and sentences are shorter in dialogue. The use of the function words shows the same tendency; the modern variants *ska*, *inte*, *bara*, *också* are more frequent in speech than in narrative. We can also see how the modern, colloquial forms as *ska* (instead of *skall*) initially are introduced in speech - in the earlier texts the authors either choose only *skall* (AS and OL) or use *ska* only in speech (VB, HS, SL and MS). The modern *ska* doesn't occur in narrative until our two latest texts, HB from 1924 and KB from 1940. In KB *ska* is the only variant - no *skall*. The variation between the two variants of the negation, the modern *inte* and the old-fashioned *icke*, is extra interesting as we can compare our results with a previous contemporary study on non fiction, Engdahl (1962). He studied the linguistic changes in two Swedish magazines during 1878-1950 and showed that the older *icke* is in majority in all texts until 1925. From 1930 and on *inte* is the most frequent variant. The transition to the more modern variant is much earlier in our corpus: *inte* occurs in all texts except OL and gets majority quite early, from HS (1901) and on. The presumption that literary texts are in lead of the linguistic progression in this period is supported in our pilot study.

[8] We have 1,001 instance of identified speakers in our training set and 142 in our test set, which can be compared to Ek et al. (2018) who have 822 lines of speech for training and test and 75 for development.

Type	Unit	AS 1879	VB 1887	OL 1900	HS 1901	SL 1912	MS 1919	HB 1924	KB 1940	Total
Narrative	Word	5.0	4.6	4.9	4.5	4.4	5.0	5.0	4.5	4.7
	Phrase	8.2	7.3	7.3	7.8	7.2	7.4	6.7	8.3	7.3
	Sentence	24.1	15.1	18.3	21.1	19.0	14.4	13.5	18.3	15.9
Speech	Word	4.2	4.0	4.5	4.6	4.1	4.3	4.4	4.1	4.1
	Phrase	5.3	6.2	5.7	5.8	4.6	6.3	6.7	6.5	6.2
	Sentence	12.9	11.0	16.2	12.5	15.9	14.4	10.7	13.2	12.0

Table 6: Average length of words in characters and phrases and sentences in words for different text types for the eight authors in the training set. Author given by initials, see Table 1 for full list

Word	Type	AS 1879	VB 1887	OL 1900	HS 1901	SL 1912	MS 1919	HB 1924	KB 1940	Total
bara ('only')	N	1.0	0.5	1.2	0.9	1.9	0.6	0.5	2.4	0.8
	S	0	3.9	2.8	4.4	3.3	3.7	1.7	3.5	3.3
inte ('not')	N	1.9	0.8	0	1.3	10.5	2.8	9.5	15.3	4.9
	S	10.8	19.6	0	17.4	19.6	12.3	21.1	22.7	18.8
också ('also')	N	1.9	0.8	2.4	1.5	0.7	1.0	0.6	2.3	1.0
	S	3.4	2.0	8.4	2.9	1.6	1.6	0.9	2.0	1.8
ska ('will'/'shall')	N	0	0	0	0	0	0	0.4	0.4	0.1
	S	0	0.4	0	1.5	0.8	4.9	3.5	3.5	1.7
endast ('only')	N	1.0	1.8	0.6	0.3	0.7	1.3	0.5	0	1.1
	S	0	0.8	2.8	0	0.2	0.4	0	0	0.5
icke ('not')	N	9.6	10.1	5.3	11.6	0	5.0	0.5	0.3	6.1
	S	10.1	2.6	16.9	1.5	0	0	0	0	1.9
även ('also')	N	0	0	0	0	0.5	0.7	0.4	0.3	0.2
	S	0	0	0	0	0.2	0	0	0	0.1
skall ('will'/'shall')	N	1.0	0.2	0.6	0.3	2.3	0	0.5	0	0.4
	S	7.4	6.3	2.8	7.3	4.7	2.0	0.6	0	4.9

Table 7: Proportion per thousand words for a number of words with modern (top) and old-fashioned (bottom) variants in N(arrative) and S(peech). Author given by initials, see Table 1 for full list

7. Future Work

So far we have annotated excerpts from eight novels. However, the number of annotations differs between the authors, and we would like to have a more balanced annotation in the future. We would also like to include more authors in our corpus. Yet another issue is the quality of the annotation. In several cases the speaker of an utterance has not been marked or resolved by annotators. This is something we want to address in the next release of the corpus. We think that overall another round of quality check would be useful, to overcome other minor issues in the consistency of annotations. Especially if more annotators are brought into the project, it would be useful to extend the guidelines, for better clarity.

So far we have focused on collecting a good quality corpus. The obvious next step is to start using it. Having this corpus has helped in identifying some of the different styles of marking dialogue that Swedish authors use. This can aid future work on identifying speech segments and speech tags, both using machine learning, like Ek and Wirén (2019), or for rule-based methods.

Our pilot study showed that we can study interesting aspects of language change from literature. However, so far we only used our manually annotated training corpus to study this. In future work, though, SLäNDa will allow us to train a classifier for identifying speech, speech tags and narrative, which would in turn allow us to apply such studies on a much larger automatically annotated selection of texts. Also, in the pilot study, we only looked at common function words. In future work we want to extend this to also look at other aspects of language change, such as changes in syntax and morphology. Our corpus will also be a useful tool for studies of narrative perspective, in line with the work of Allison (2018) and Håkansson and Östman (2019).

8. Conclusion

In this paper, we describe a new corpus, SLäNDa, where Swedish literary fiction has been manually annotated for cited materials, like signs, speech, and quotations, which results in also isolating the remaining narrative text. The main focus is on dialogue, for which we annotate speech segments, speech tags, and speakers. We show that we can perform this annotation with a high inter-annotator agreement for the identification of cited materials. The agreement for speakers is lower, but mainly due to missing annotations, rather than confusions between speakers.

SLäNDa contains excerpts from eight Swedish novels written between 1879–1940, a period of modernization of the Swedish language. In total, SLäNDa contains annotations of 44 chapters with over 220K tokens. The annotation identified 4,733 instances of cited material and 1,143 named speaker–speech mappings. In addition, the unannotated part of SLäNDa consists solely of narrative.

We performed a small pilot study where we show how these types of annotations can help in analysing language change in Swedish. We focused on lexical aspects, and

showed that for a number of function words with an old-fashioned and modern variant, the change between the two versions happened earlier for speech than for the narrative. In both cases the change was earlier than has previously been seen in other text types.

Acknowledgements

This corpus was mainly created with the support from the cooperation project "From close to distant reading" funded by the Disciplinary Domain of Humanities and Social Sciences at Uppsala University. We thank the other project members Karl Berglund, Mats Dahllöf, David Håkansson, Joakim Nivre, and Johan Svedjedal for insightful discussions. We thank Rosa Cirillo and Ercan Eriksson-Aras for help with the annotation, Richard Eckart de Castilho for help with WebAnno, and Per Starbäck for technical support. We thank Litteraturbanken (The Swedish Literature Bank) for making all texts available.

Bibliographical References

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

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2).

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. European Language Resources Association (ELRA).

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.

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*, pages 261–280.

He, H., Barbosa, D., and Kondrak, G. (2013). Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria.

Holm, L. (2015). Rytm i romanprosa. en studie av rytmiska signalement i tio samtida svenska romaner. In Carin Östman, editor, *Det skönlitterära språket. Tolv texter om stil*, pages 215–235. Morfem, Stockholm.

Leech, G. N. and Short, M. (1981). *Style in fiction: a linguistic introduction to English fictional prose*. Longman, London.

Moretti, F. (2000). Conjectures on world literature. *New Left Review*, 40(1):54–68.

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

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., Svedjedal, J., and Östman, C. (2018). Språklig rytm i skönlitterär prosa. En fallstudie i Karin Boyes *Kallocain*. *Samlaren. Tidskrift för forskning om svensk och annan nordisk litteratur*, 139:128–161.

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.

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.

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.

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/`.

Muzny, G. and Fang, M. and Chang, A. and Jurafsky, D. (2017). *QuoteLi3*. `https://nlp.stanford.edu/~muzny/quoteli.html`.

Stymne, S. and Östman, C. (2020). *SLäNDa*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. `http://hdl.handle.net/11372/LRT-3169`, version 1.0.