Colloquialization in the Narrative and Dialogue of Swedish Fictional Prose

Sara Stymne[1], **Carin Östman**[2] **and David Håkansson**[2]
[1]Department of Linguistics and Philology, Uppsala University, Sweden
[2]Department of Scandinavian Languages, Uppsala University, Sweden
sara.stymne@lingfil.uu.se
{carin.ostman,david.hakansson}@nordiska.uu.se

Abstract

We describe an ongoing cross-disciplinary project with the goal of exploring the role of literary fiction in the development of the Swedish written language, in the period 1830–1930. We describe the overall goals of the project, including the automatic separation of narrative and dialogue, a challenging task since typographic marking is inconsistent. We present the SLäNDa corpus, where dialogue and narrative are annotated. This corpus will allow training models to separate dialogue and narrative, which will in turn allow us to further explore differences in the language of these parts. We use SLäNDa in two pilot experiments, on the role of typographical markers for automatic identification of characters' speech, and of the occurrences of old-fashioned and modern function words in dialogue and narrative.

1 Introduction

The modern Swedish novel had its breakthrough in the early 19th century (Tigerstedt, 1956). The growing number of novels by Swedish auhtors during the 19th and early 20th centuries has been characterized as an important source for the renewal of the Swedish language (Engdahl, 1962, p. 169). It has previously been established that fiction was important for the renewal of the Swedish language around the turn of the 19th century, but the focus has been on a few famous authors and works, such as *Röda rummet* by August Strindberg, which has been said to mark boundaries between different periods in the history of the Swedish language (Thelander, 1988). The role of fiction in general has, however, mainly been overlooked and little is known about the general impact of literature on language development during this period.

We present a cross-disciplinary project, with the overarching goal to perform a large-scale investigation of Swedish fictional prose with the focus on giving a more complete view of the Swedish language and communication history. One sub goal is to investigate the language in literary dialogue, as opposed to literary narrative, and compare the development of new linguistic patterns to those in other types of contemporary texts. The time period in focus is 1830–1930, which is a period of much change in the Swedish written language, largely driven by the goal to modernize the written language, and move it closer to the spoken language.

In this paper, we give an overview of the project, and then zone in on one contribution, the Swedish literary corpus of narrative and dialogue, SläNDa. The main goal of SLäNDa is to provide both training and test data for automatic separation of dialogue and narrative, which will allow an exploration of the language development in speech versus narrative. We also present two small pilot studies based on SLäNDa, where we investigate the usage of function words, and the impact of typographic markers on speech identification. SLäNDa has previously been presented in Stymne and Östman (2020, 2022), where more details can be found.

2 Project: How Fiction Modernized the Swedish Language

The work presented in this paper stems from the project *Fictional prose and language change. The role of colloquialization in the history of Swedish 1830–1930*, funded by the Swedish Research Council, 2021–2023. The overarching goal for the project is to put the spotlight on the role of fiction and fictional dialogue as a catalyst in processes of language change, by investigations into Swedish fictional prose. Researchers from Scandinavian languages, literary studies and computational linguistics take part in the project in order to explore the theme from different perspectives.

It has been stated that fiction is a rich source of language renewal (Engdahl, 1962), and we know that a few authors and works have been pinpointed as good representatives of renewing the Swedish language (Lundevall, 1953, p. 103). It is also known that there were collaborations between individual authors, and progressive language scholars during this period (Östman, 2014). However, there is a lack both of broad and of deep studies on the theme. Our knowledge on the language development in the 19th and early 20th century is to a large extent based on only a few authors, and on studies performed in the early 20th century (e.g. Lindstedt, 1922; Von Hofsten, 1935). In our project, we focus on 1830–1930, a period of large changes in language and literature, as well as in the Swedish society. The main questions addressed in the project are:

a) How does the language differ in literary narrative and literary dialogue?
b) How is language variation used stylistically in literary dialogue?
c) How does language innovations spread from dialogue to narrative?
d) How are new language norms and innovations established and spread?
e) How can computational linguistics methods be used to separate narrative and dialogue?
f) How are language colloquialization ideas expressed in individual authorships?

In this paper we focus mainly on **e**, how narrative and dialogue can be automatically separated.

3 Marking of Speech

Unlike many other languages, where quotation marks is the predominant means of marking speech, there is much more variety in Swedish, especially during the period of interest. Besides quotation marks, the main mean of marking speech is through the use of dashes, but there are also many works where no marking is used, or where the usage is inconsistent. Examples (1–3) show examples of quotation marks, dashes and no marking. The structure in the three examples is the same, the speech segment starts, followed by a speech tag, in turn followed by the continuation of the speech segment; the difference only lies in the usage of typographical marking.

(1) »Tala så lågt du vill!» säger han. »Jag skall ändå höra dig.»
S. Lagerlöf, *Körkarlen*, p. 121

(2) – En fin karl, upplyste brodern i förtroende. Kontorist. Låter kosingen springa snällt ...
M. Sandel, *Hexdansen*, p. 82

(3) Pappa! sade svågern. Har pappa rest?
L. Nordström, *Borgare*, p 89

Example 2 shows the most common way of using dashes, just at the start of a speech segment, but not at the end of it, or in any way identifying the restart after the speech tag, as there is with quotation marks. But there were quite some variation between authors during this period, and there are cases where authors also use dashes at the end of speech segments and/or at restarting speech segments.

4 The SLäNDa Corpus

SLäNDa, the Swedish Literary corpus of Narrative and Dialogue, is a manually annotated corpus of Swedish literature, with a focus on annotating speech segments. It has previously been described in Stymne and Östman (2020, 2022). It was developed with the main goal of enabling the training and evaluation of models for separating dialogue and narrative, but can also enable speaker identification. Some similar corpora exist for other languages (e.g. Papay and Padó, 2020; Semino and Short, 2004; Brunner, 2013), but for Swedish, we are only aware of a small corpus containing four works, all with dashes (Ek and Wirén, 2019). SLäNDa version 2.0 is publicly released in the LINDAT/CLARIAH-CZ repository,[1] under the Creative Commons licence CC BY-NC-SA.

4.1 Annotation

SLäNDa contains annotations of all aspects that are not part of the main narrative. That means that all unannotated parts form the narrative. The annotated categories are:

- Speech
- Speech tag
- Other type of speech tag
- Embedded Speech
- Thought
- Sign
- Letter
- Quotation
- Other

For a standard speech tag, our criterion, based on Semino and Short (2004) and Teleman (2003),

[1] http://hdl.handle.net/11372/LRT-4739

is that it should always contain a verb signaling speech, whereas other type of speech tag is used for more indirect passages introducing speech. Since the main purpose of the corpus is on speech and dialogue, we in addition also annotate the speaker of utterances, including both the mention in the speech tag, if present, and a resolved speaker. Note that we focus on direct speech. While also interesting, the current version of SLäNDa does not contain annotations of indirect speech.

The guidelines for SLäNDa were originally agreed upon within a cross-lingual group of researchers, where we iterated between updating the guidelines and rounds of pilot annotation. The annotations in SLäNDa version 1.0 were performed by three annotators, after an initial training session and a round of pilot annotation, which led to minor updates of the guidelines. The guidelines were further updated for version 2.0, and the additional data there was annotated by a single expert annotator. For the annotation we used the WebAnno tool (Yimam et al., 2013). The inter-annotator agreement for the main distinctions in SLäNDa version 1.0, had kappa values of 0.72 and 0.83 between two pairs of annotators. The speaker identification agreement was lower, but mainly due to one annotator not resolving pronouns, thus leaving many identifications out.

4.2 Texts

SLäNDa contains excerpts from 19 works. The criteria for selecting texts are:

- Fictional prose: novels or collections of short stories
- Released 1809–1940
- Available in a proofread XML-format
- Creative Commons license

In addition we aimed at having a mix of works with different author genders, and a mix of strategies for marking speech. For the full list of included works, see Stymne and Östman (2022). All texts are retrieved from Litteraturbanken, *The bank of literature*, a large collection of Swedish literary works.[2]

SLäNDa is split into a training set, with texts from 11 authors with a mix of speech marking, and five test sets with different types of speech marking. From SLäNDa v1.0, there are three small test sets, with quotation marks, dashes, and no marking, where the works in the test sets are also present in the training data. From SLäNDa v2.0 there are two larger test sets with dashes and no marking of speech, only containing works of authors not in the training data. The overall size of SLäNDa v2.0 is 274,704 tokens, with 2051 speech segments in the training data and 1790 in the test data. Speech tags are also common, with 930 instances in the test data, and 826 in test data, whereas the other classes are less common.

In order to further be able to investigate the effect of typographical marking of speech segments, we provide two version of the data, the original data, and a stripped version where all quotation marks and dashes marking speech have been removed, to create artificially unmarked speech segments. Such data can be used to explore the effect of these markers in both training and test data.

[2] https://litteraturbanken.se/

	Original		Stripped	
	Speech	Tags	Speech	Tags
Dash	90.3	80.3	17.3	73.0
Dash-strip	63.0	68.0	74.0	74.7
Unmarked	72.7	70.0	75.3	73.7

Table 1: F1-scores for the prediction of speech segments and speech tags in the SLäNDa v2.0 test sets, when training with the original data, and with stripped data.

5 Pilot Experiments

In this section we describe two pilot experiments based on the SLäNDa corpus.

5.1 Dialogue Identification

In a first pilot, we investigated the effect of typographical markers in the training and test data. The task was identification of speech segments and speech tags; the other categories of SLäNDa were ignored in this experiment. We set the experiment up as a token classification task, where we converted the original annotations in SLäNDa to an IOB-format, where all speech segments, speech tags, and other annotations were marked as spans. We used a toolkit originally developed for named entity recognition, the T-NER toolkit (Ushio and Camacho-Collados, 2021), fine-tuning the Swedish BERT model KB-BERT (Malmsten et al., 2020). T-NER uses a linear layer on top of the last BERT layer and a cross-entropy loss. We use its default

parameters.[3]

Table 1 shows the F1-scores for the identification of speech tags and speech segments in the SLäNDa v2.0 test sets, including a stripped version of the test set with dashes, when trained with the original training data, and a stripped version. We can see that dashes seem to help identification, when present, since the highest scores are achieved for the dash test set with original training data. The big drop with stripped data is mainly due to the dashes not being matched; the token-level identification is affected to a lower extent. For the test data without any marking, it is in both cases better to train with stripped training data, than with the original training data. The performance is much worse for the stripped dash set than for the originally unmarked set when trained with the original data, suggesting that there might be more textual clues in the originally unmarked data, than what is needed with dashes present, but with stripped training data, the performance is on par for these two test sets.

5.2 Function Words

In another small experiment, we wanted to explore how modern the language is in dialogue versus narrative for a small set of authors, by investigating the use of a small set of function words, which have modern and old-fashioned variants. This is explored for four contrastive pairs of words, in the training data from SLäNDa v1.0 covering eight authors 1879–1940, with the full results available in Stymne and Östman (2020). Here we will focus on two contrasts: *skall* versus *ska* and *icke* versus *inte*.

The old-fashioned *skall* is completely dominant in the narrative for the six oldest authors, up until 1919, with no instances of *ska*. In the dialogue *ska* does occur earlier, with examples from 1887 and 1901 onwards, but *ska* does not occur in majority until the three last works, 1919–1940. In the newest work, Boye 1940, only *ska* is used both in the narrative and dialogue. The old fashioned *icke* predominantly occurs for the four earliest authors, 1879–1901, the later authors only use it sporadically, and never in speech. The modern variant *inte* occurs in all authors except Levertin. For the four early authors it predominantly occurs in speech, rarely in the narrative. A clear example of mixed usage is Söderberg from 1912, where both variants occur, but *icke* predominantly in the narrative and *inte* predominantly in speech. This usage can be compared to a study by Engdahl (1962) of Swedish magazines 1878–1950, where he shows that *icke* is predominant until 1925, and *inte* from 1930; which means that the shift seems to be earlier in literature, and especially in literary dialogue.

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

6 Conclusion and Future Work

In this paper we introduced an ongoing project with the goal of exploring the role of literary fiction in the development of the Swedish written language in the period 1830–1930. A special emphasis was put on the SLäNDa corpus, where speech segments, including speakers, are annotated. This corpus will allow us to train and evaluate models to separate dialogue and narrative, discussed in connection to a pilot experiment, which will in turn allow us to further explore differences in language in these parts.

We are currently exploring how a range of linguistic features change over time in the narrative and dialogue of a high number of novels and collection of short stories. We have identified an inventory of indicators, which has previously been proposed as markers of modernization, covering word choice, morphology, and syntax. In a first step we explore only works that use quotation marks, where dialogue and narrative is easy to separate, but in the next step we plan to improve our current classifiers for text with dashes and no marking so that we can include those works in the study as well. The automatic analysis can give a picture on a large scale, but it can also guide where the efforts in manually studying a selection of works can be focused, for instance by identifying outliers.

Acknowledgments

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*. We thank Johan Svedjedal, Karl Berglund, Mats Dahllöf, and Joakim Nivre 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.

References

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

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

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

Torvald Lindstedt. 1922. Studier över stilen i Gösta Berlings saga. *Nysvenska studier*, 2:31–77.

Karl-Erik Lundevall. 1953. *Från åttital till nittital. Om åttiotalslitteraturen och Heidenstams debut och program*. Ph.D. thesis, Stockholms högskola.

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

Carin Östman. 2014. Selma Lagerlöf och språkråden. språkvård och skönlitteratur i tidigt 1900-tal. In *Svenskans beskrivning 33*, pages 532–540, Helsinki, Finland.

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

Elena Semino and Mick Short. 2004. *Corpus Stylistics. Speech, writing and thought presentation in a corpus of English writing*. Routledge, London.

Sara Stymne and Carin Östman. 2020. SLäNDa: An annotated corpus of narrative and dialogue in Swedish literary fiction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 826–834, Marseille, France. European Language Resources Association.

Sara Stymne and Carin Östman. 2022. SLäNDa version 2.0: Improved and extended annotation of narrative and dialogue in Swedish literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5324–5333, Marseille, France. European Language Resources Association.

Ulf Teleman. 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.

Asahi Ushio and Jose Camacho-Collados. 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. Association for Computational Linguistics.

Louise Von Hofsten. 1935. Några stildrag hos Selma Lagerlöf med utgångspunkt från Charlotte Löwenskiöld. *Nysvenska studier*, 15:150–183.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 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. Association for Computational Linguistics.