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

UPPSALA UNIVERSITET

Sara Stymne and Carin Östman

Introduction

- The Swedish Literary corpus of Narrative and Dialogue
- Distinguishes narrative from other parts, such as speech, thoughts, letters
- Extended and improved compared to SLäNDa version 1.0
- Main focus on separating speech from non-speech
- Analysis of the importance of tyopgraphic speech marking

Marking Speech

Quotation marks (QM)

»Du borde gifta dig», sa Börje.
'»You should get married» , Börje said.'

Benedictsson, p. 309

Dashes

– 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. …'

Sandel, p. 41

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

Rydberg p. 248

Unmarked

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

Cederborgh p. 19

Data Sets

	Training	Test
Speech segment	2051	1790
Speech tag	930	826
Other type of speech tag	33	16
Embedded speech	5	0
Thought	46	39
Quotation	11	8
Letter	8	12
Sign	1	0

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

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

Test sets in SLäNDa version, with number of speech segments and speech tags, and the number of authors.

Contents

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

Authors and novels in SLäNDa version 2.0, with the publication year, preferred speech marker, 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.

Pilot Experiments

Separating speech, speech tags and narrative

- Task: separate narrative from speech segments and speech tags
- Focus: What is the impact of typographic marking of speech?
  - Use of dedicated test sets with different marking
  - Stripped version of data, with quotation marks and dashes removed
- System:
  - Token-level classification (IOB-format)
  - Fine-tuning of Swedish BERT (KB-BERT)
  - T-NER toolkit
- Evaluation: Segment-level F1-score

Results

	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

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

Summary

- Extended corpus with annotations of narrative and other elements
- Focus on test sets with different types of speech marking
- Final project goal: analyse language change in literary dialogue and narrative
  - Small early studies suggest that modernization of Swedish happened earlier in dialogue than in narrative
  - SLäNDa allows the development of tools for large-scale studies of this hypothesis!
- Supported by Swedish Research Council: *Fictional prose and language change. The role of colloquialization in the history of Swedish 1830–1930*