Two Swe-Clarin pilot projects in Språkbanken

A big-data approach to the study of rhetorical history

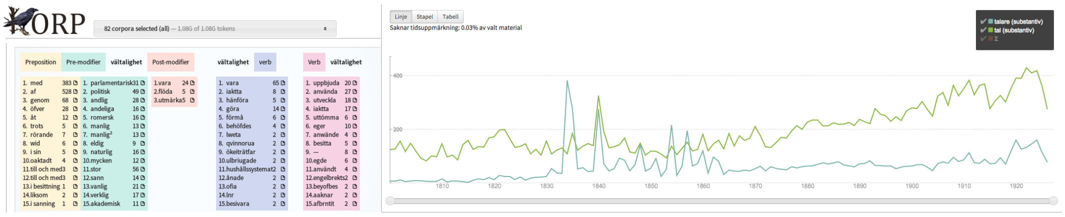
Rhetorical history is traditionally studied through rhetorical treatises or selected rhetorical practices, for example the speeches of major orators. Although valuable sources, these do not give us the answers to all our questions. Indeed, focus on a few canonical works or the major historical key figures might even lead us to reproduce cultural self-identifications and false generalizations. However, thanks to increasing availability of relevant digitized texts, we are now at a point where it is possible to see how new research questions can be formulated – and how old research questions can be addressed from a new angle or established results verified – on the basis of exhaustive collections of data, rather than small samples, but where a methodology has not yet established itself.

In this pilot project – a collaboration between Jon Viklund, rhetorical historian at Uppsala University and the Språkbanken group of Swe-Clarin – we study the historical development of attitudes towards rhetoric as expressed in public discourse over the last 200 years. The data for the preliminary investigation conducted in the pilot project is a massive collection of texts: one billion words of digitized Swedish newspapers from the 19th and early 20th century.

The research questions were addressed using Språkbanken's existing corpus infrastructure Korp. This allowed us both to start to explore the material quickly and to note where added functionality would be needed to that provided by Korp, which was designed primarily with linguistic research in mind.

Nevertheless, Korp functions such as the *word picture* and *trend graph* proved to be useful tools for this kind of investigation.

Thus, the word picture (shown to the left in the picture below) was used to investigate significant modifiers of the noun *vältalighet* 'eloquence', illustrating general attitudes to rhetoric as well as salient conceptual metaphors employed in characterizing it, and the trend graph (to the right in the picture below) revealed (1) a cyclically varying preoccupation with 'speech' and 'speakers' following the three-year intervals between the sessions of the old-style parliament (abolished in 1866), and (2) a markedly increasing interest in political matters towards the end of the 19th century.



82 corpora selected (all) — 1.08G of 1.08G tokens

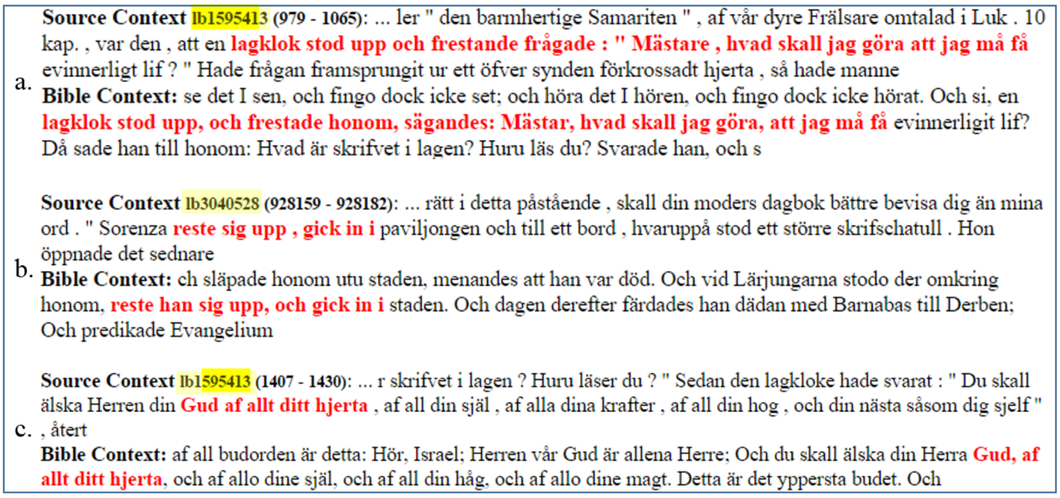
Preposition		Pre-modifier		vältalighet	Post-modifier	
1. med	383	1. parlamentarisk	31		1.vara	24
2. af	528	2. politisk	49		2.flöda	5
3. genom	68	3. andlig	28		3.utmärka	5
4. öfver	28	4. andeliga	16			
5. åt	12	5. romersk	16			
6. trots	5	6. manlig	13			
7. rörande	7	7. manlig²	13			
8. wid	6	8. eldig	9			
9. i sin	5	9. naturlig	16			
10.oaktadt	4	10.mycken	12			
11.till och med	3	11.stor	56			
12.till och med	3	12.sann	14			
13.i besittning	1	13.vanlig	21			
14.liksom	2	14.verklig	17			
15.i sanning	1	15.akademisk	11			

vältalighet	verb	
	1. vara	65
	2. iaktta	8
	3. hänföra	5
	4. göra	14
	5. förmå	6
	6. behöfdes	4
	7. lweta	2
	8. qvinnorua	2
	9. ökeiträtfar	2
	10.ulbriugade	2
	11.hushållssystemat	2
	12.ånade	2
	13.ofia	2
	14.lnr	2
	15.besivara	2

Verb		vältalighet
1. uppbjuda	20	
2. använda	27	
3. utveckla	18	
4. iaktta	17	
5. uttömma	6	
6. eger	10	
7. anwände	4	
8. besitta	5	
9. —	8	
10.egde	6	
11.användt	4	
12.engelbrekts	2	
13.beyofbes	2	
14.aaknar	2	
15.afbrntit	2	

Saknar tidsuppmärkning: 0.03% av valt material

KORP
Linje
Stapel
Tabell
talare (substantiv)
tal (substantiv)
Σ
400
200
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920



a. **Source Context** lb1595413 **(979 - 1065)**: ... ler " den barmhertige Samariten " , af vår dyre Frälsare omtalad i Luk . 10 kap. , var den , att en **lagklok stod upp och frestande frågade : " Mästare , hvad skall jag göra att jag må få** evinnerligt lif ? " Hade frågan framsprungit ur ett öfver synden förkrossadt hjerta , så hade manne
Bible Context: se det I sen, och fingo dock icke set; och höra det I hören, och fingo dock icke hörat. Och si, en **lagklok stod upp, och frestade honom, sägandes: Mästar, hvad skall jag göra, att jag må få** evinnerligit lif? Då sade han till honom: Hvad är skrifvet i lagen? Huru läs du? Svarade han, och s

b. **Source Context** lb3040528 **(928159 - 928182)**: ... rätt i detta påstående , skall din moders dagbok bättre bevisa dig än mina ord . " Sorenza **reste sig upp , gick in i** paviljongen och till ett bord , hvaruppå stod ett större skrifschatull . Hon öppnade det sednare
Bible Context: ch släpade honom utu staden, menandes att han var död. Och vid Lärjungarna stodo der omkring honom, **reste han sig upp, och gick in i** staden. Och dagen derefter färdades han dädan med Barnabas till Derben; Och predikade Evangelium

c. **Source Context** lb1595413 **(1407 - 1430)**: ... r skrifvet i lagen ? Huru läser du ? " Sedan den lagkloke hade svarat : " Du skall älska Herren din **Gud af allt ditt hjerta** , af all din själ , af alla dina krafter , af all din hog , och din nästa såsom dig sjelf " , återt
Bible Context: af all budorden är detta: Hör, Israel; Herren vår Gud är allena Herre; Och du skall älska din Herra **Gud, af allt ditt hjerta**, och af allo dine själ, och af all din håg, och af allo dine magt. Detta är det yppersta budet. Och

Biblical quotes in Swedish 18th–19th century prose fiction

Text reuse, a form of text repetition, recycling or borrowing, is a theoretically and practically interesting problem that has attracted considerable attention during the last years e.g., in the cultural heritage context and in measuring of journalistic reuse. Text reuse is closely relate to the notion of *intertextuality*, i.e. the "shaping of a text's meaning by another text", and has become an exploration playground for scholars in digital humanities worldwide due to the growing availability of large electronic historical corpora of various kinds that has open new possibilities for intertextual-based language data explorations.

In this pilot project – a collaboration between prof. Mats Malm, a scholar in comparative literature at the University of Gothenburg and the Språkbanken group of Swe-Clarin – we study the multifaceted relations between non-contemporary Swedish texts by identifying pairs of text passages likely to contain substantial overlap, with the aim to empirically support new interpretations of the historical texts. The data for the investigation conducted in this pilot project is the Charles XII Bible translation into Swedish, completed in 1703, against the content of the Swedish prose fiction 1800–1900.

The research questions were addressed by applying existing software. This allowed us both to start to explore the text sources quickly and to note what added functionality would be needed in order to increase both precision and most importantly recall. The problems we encountered are numerous and challenging. Recycled text chunks cover often only a small portion of a document and may be significantly modified. Algorithms must take into consideration various alternations that may have transformed a text segment to a completely new counterpart. Not only spelling and orthographic variations, but also synonym usage as well as OCR-errors can be problematic. The three examples shown above illustrate both the strengths [a] but also the limitations ([b] trivial associations and [c] need for better matching techniques) of simple, yet powerful techniques for historical text exploration, here sequence alignment.

The Swe-Clarin steering committee is finally in place!

The Swe-Clarin steering committee, consisting of 8 distinguished members of the Swedish and Norwegian research community, has been chosen to represent a wide range of fields in digital humanities as well as language technology. The steering committee's first meeting was held in Gothenburg on February 19th and there are plans for another meeting already this spring. In the coming newsletters we will present each member of the steering committee as well as the members of the national coordination team.

Our next physical consortium meeting will be held in Uppsala on April 4-5. We encourage any researchers interested in Swe-Clarin to submit proposals for discussion topics, future work or collaborations. The proposals can be submitted via info@sweclarin.se latest on March 24th.

Calendar

14–15 Mars: NCN workshop in Oslo on Big language data
15–17 Mars: Nordic DH conferens `http://dig-hum-nord.eu`
4–5 April: Consortium meeting in Uppsala.
Spring: Swe-Clarin-on-tour stops in Stockholm. The theme is the Nordic Museum questionnaires.
November: Another Swe-Clarin-on-tour stop in conjunction with SLTC, the Swedish Language Technology Conference, in Umeå.

Future News

If you have information for the newsletter, please send it to the Coordination Team (info@sweclarin.se). If you do not get this newsletter automatically, you can sign up for the news list here:
`http://lists.sweclarin.se/mailman/listinfo/news_lists.sweclarin.se`