

## Två Swe-Clarín-pilotprojekt i Språkbanken

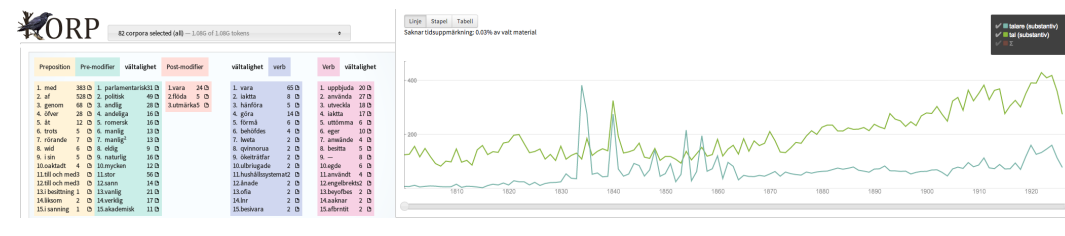
### Retorikhistoria studerad med dataintensiva metoder

Av tradition har retorikhistorien studerats via retorikhandböcker eller enskilda retoriska praktiker, till exempel stora talares produktion. Detta är utan tvivel värdefulla kunskapskällor, men de besvarar långt ifrån alla våra frågor. I själva verket kan fokusering på några få kanoniska arbeten eller historiska nyckelgestalter få oss att projicera kulturella självbilder på historien eller leda oss till falska generaliseringar. Tack vare en ökande tillgång på relevanta digitala textmaterial är vi nu vid en punkt där vi istället kan se hur nya forskningsfrågor kan ställas – och gamla forskningsfrågor belysas på nya sätt eller etablerade resultat verifieras – med hjälp av uttömmande datamängder och dataintensiva automatiska processer snarare än manuell genomgång av små materialurval, men där en metodologi ännu inte har utvecklats.

I det här pilotprojektet – ett samarbete mellan Jon Viklund, retorikhistoriker vid Uppsala universitet och Swe-Clarín-gruppen vid Språkbanken – studerar vi hur attityderna till retorik har utvecklats historiskt genom deras uttryck i det offentliga samtalet över de senaste 200 åren. Databasen för denna preliminära studie är en stor textsamling: en miljard ord digitaliserade svenska dagstidningar från 1800-talet och början av 1900-talet.

Vi tog oss an forskningsfrågorna med hjälp av Språkbankens befintliga korpusinfrastruktur Korp, vilket gjorde det möjligt för oss att direkt börja utforska materialet och dessutom notera vilken typ av ytterligare funktioner som skulle behövas för den här typen av digital humanioraforskning (eftersom Korp huvudsakligen har utformats med språkvetenskaplig forskning i åtanke).

Korps *ordbild* och *trenddiagram* visade sig vara mycket användbara för våra ändamål. Ordbilden (återgiven till vänster i bilden nedan) användes för att undersöka signifikanta attribut till substantivet *vältalighet*, för att därigenom belysa allmänna attityder till retorik och även framträdande konceptuella metaforer förknippade med den. Trenddiagrammet (till höger i bilden nedan) visade att (1) *tal* och *talare* stod i fokus i en treårscykel som sammanföll med den gamla ståndsriksdagens sammankomster och (2) intresset för politik ökade markant över slutet av 1800-talet.



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 framsprunget ur ett öfver synden förkrossadt hjerta , så hade manne**

- a. **Bible Context:** se det I sen, och fingo dock icke set; och höra det I hören, och fingo dock icke hört. Och si, en **lagklok stod upp, och frestade honom, sägandes: Mästar, hvad skall jag göra, att jag må få evinnerligt lif?** Då sade han till honom: Hvad är skrifvet i lagen? Huru läs du? Svarade han, och s

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

- b. **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

Source Context [lb1595413](#) (1407 - 1430): ... r skrifvet i lagen ? Huru läs 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 "

- c. , 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

### Bibliska citat i 1800–1900-tals svensk prosafiktion

Textåteranvändning definieras som en avsiktlig eller en oavsiktlig användning av redan existerande text för att skapa en ny text. Textåteranvändning är ett begrepp besläktat med *intertextualitet*, som innebär att ingen text står för sig själv, utan ingår i ett nät av rela-

tioner till andra texter. Textåteranvändning är ett teoretiskt och praktiskt forskningsproblem med många tillämpningsområden (som t.ex. plagiatkontroll) som har fått mycket uppmärksamhet under de senaste åren inom humaniora och samhällsvetenskap på grund av det massiva digitala tillgängliggörandet av kulturarvsmaterial och de nya möjligheter till storskaliga undersökningar som kan utföras med språkteknologisk metodik.

I det här pilotprojektet – ett samarbete mellan prof. Mats Malm, litteraturvetare vid Göteborgs universitet och Swe-Clarín-gruppen vid Språkbanken – studerar vi hur textåteranvändning tar form både ortografiskt, syntaktiskt och semantiskt. Från ordagranna upprepningar av enstaka korta eller långa ordsekvenser (t.ex. meningar) som kan finnas återgivna till ett måldokument från ett källdokument, till försök att hitta enkla parafrafer och omskrivningar. Datan för denna preliminära studie är svensk prosafiktionsamlingen 1800–1900, och Karl XII:s bibel som trycktes 1703.

Vi tog oss an forskningsfrågorna med hjälp av tillgänglig mjukvara, vilket gjorde det möjligt för oss att direkt börja utforska materialet och dessutom notera vilken typ av funktionalitet som behövs för att öka korrektheten, framförallt täckningen. Ett textfragment kan vara omskrivet men fortfarande kan det finnas "spår" av källan som kan kännas igen i den nya texten. Historisk textåteranvändning är ett komplext och svårhanterligt problem, texter producerade under tidigare århundraden skiljer sig i många avseenden från moderna texter som mycket av den existerande språkteknologiska infrastrukturen har utvecklats för. Inte bara annan ortografi och stavningsvariation, men också användning av ord som under åren har blivit ersatta med synonymer, varianter och nya betydelser, omformuleringar m.m., försvårar överblick, tolkning och vidare analys. Samtidigt får man även räkna med att digitaliserade äldre texter kan innehålla brus i form av t.ex. OCR- och formateringsfel. De tre exemplen ovan illustrerar både styrkan [a] men också begränsningarna ([b] triviala associationer och [c] behov av bättre matchningstekniker) hos de enkla men kraftfulla tekniker för historisk textutforskning som finns tillgängliga idag, här sekvenslinjering.

## Swe-Clarins styrgrupp är äntligen på plats!

Swe-Clarins styrgrupp består av åtta framstående forskare från Sverige och Norge och representerar en bred grupp av forskare inom digital humaniora och språkteknologi. Styrgruppen höll sitt första möte i Göteborg den 19 februari och det finns planer för ytterligare ett möte under våren. Vi kommer att presentera styrgruppens medlemmar, samt den nationella samordningens medlemmar, i de kommande nyhetsbrev.

Vårt nästa fysiska konsortiemöte kommer att hållas i Uppsala under 4-5 april. Vi uppmanar alla forskare som är intresserade av vårt arbete inom Swe-Clarín att skicka in motioner gällande diskussionspunkter, framtida arbete och samarbeten till oss. Motioner kan skickas via [info@sweclarin.se](mailto:info@sweclarin.se) fram till och med 24 mars

### Kalender

14–15 mars: NCN workshop i Oslo om Big language data

14–15 mars: DH konferens <http://dig-hum-nord.eu>

4–5 april: Konsortiemöte i Uppsala.

Vår: Swe-Clarín på turné-stopp i Stockholm. Temat är Nordiska museets frågelistor.

November: Ett Swe-Clarín på turné-stopp planeras i anslutning till SLTC, den svenska språkteknologikonferensen, i Umeå.

### Framtida nyheter

Meddela gärna samordningsgruppen ([info@sweclarin.se](mailto:info@sweclarin.se)) om ni har information till nyhetsbrevet. Om man inte får nyhetsbrevet skickat kan man anmäla sig till e postlistan här:

[http://lists.sweclarin.se/mailman/listinfo/news\\_lists.sweclarin.se](http://lists.sweclarin.se/mailman/listinfo/news_lists.sweclarin.se)

