Big data, Text mining och digitala metoder

Lars Ahrenberg, Arne Jönsson

13 maj 2020

SWE-CLARIN

Text som (big) data

- Större mängder än vad en person kan förväntas läsa
  - Ofta varierad och/eller brusig
  - Ofta expanderande över tid
  - Ofta analyserad reduktionistiskt, t.ex. som
    bag-of-words, *context windows*
  - Ofta analyserad med statistiska metoder, t.ex.
    topic modelling,

SWE-CLARIN

Utdrag ur SOU från 1930 (jmf. Jarlebrink et al. 2016)

```
<s>
Allenast den , som av fullt vederhäftiga och trovärdiga personer intygades v a r a fullt pålitlig vid bruket av
 rusdrycker , borde tilldelas körkort . </s>
<s>
De sakDen ordning , som sålunda kom till stånd , h a r icke visat sig tillf redskunniga . </s>
<s>
ställande . </s>
<s>
Intyg om nykterhet , ordentlighet och hänsynsfullt uppträdande lärer så gott som vem som helst kunna prestera ,
 och detta av personer , vilkas trovärdighet länsstyrelsen saknar all anledning att betvivla . </s>
<s>
F å äro de , vilka vilja utsätta sig för det obehag , som g i v e t v i s ä r förenat med att neka en v ä n ell
er granne ett dylikt intyg . </s>
```

SWE-CLARIN

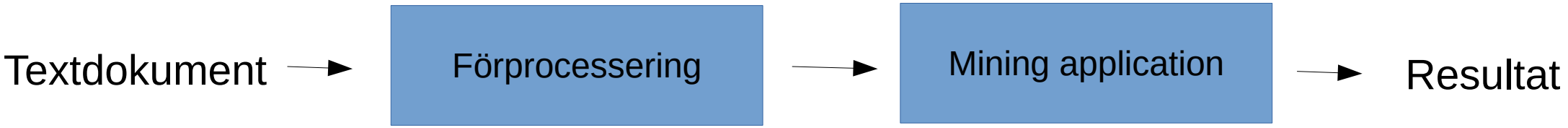
Text som språklig kommunikation

- Segmentering (i ord, meningar)
- Ordklassbestämning (”taggning”)
- Betydelsebestämning (”word sense disambiguation”)
- Syntaktisk analys (”parsning”)
- Namnigenkänning (personer, platser, företag, …)
- Rollbestämning (”vem gjorde vad när var hur?”)

SWE-CLARIN

Text mining

- Oprecis term som omfattar både statistisk analys och språkteknologi, och som används på olika sätt av olika författare.



SWE-CLARIN

Text mining vs. Text retrieval
(källa: Zhai & Massung, 2016, sid. 11)

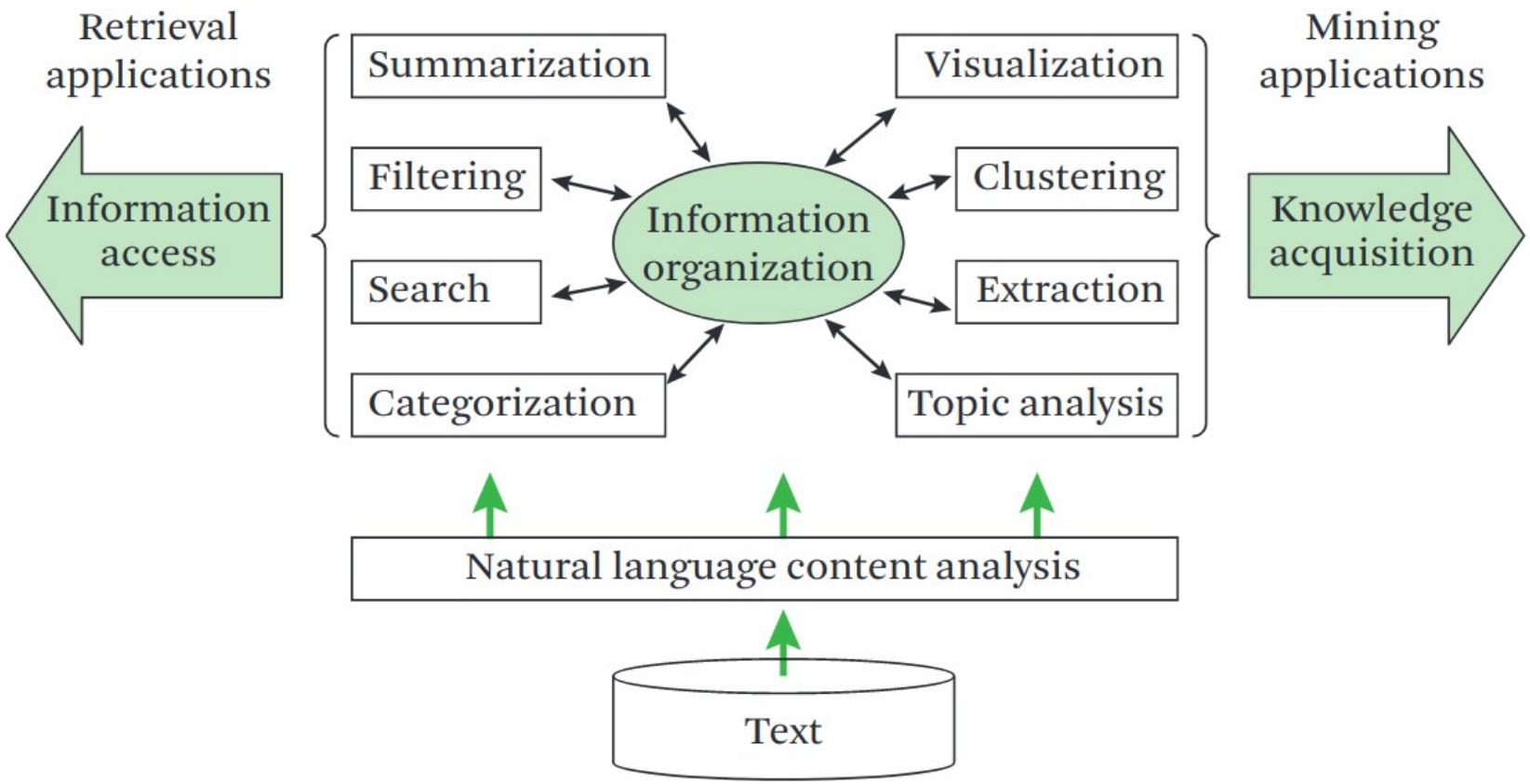


Figure 1.3 Conceptual framework of text information systems.

Text mining i humanistisk forskning
(källa: Tahmasebi & Hengshen, 2019, sid. 199)

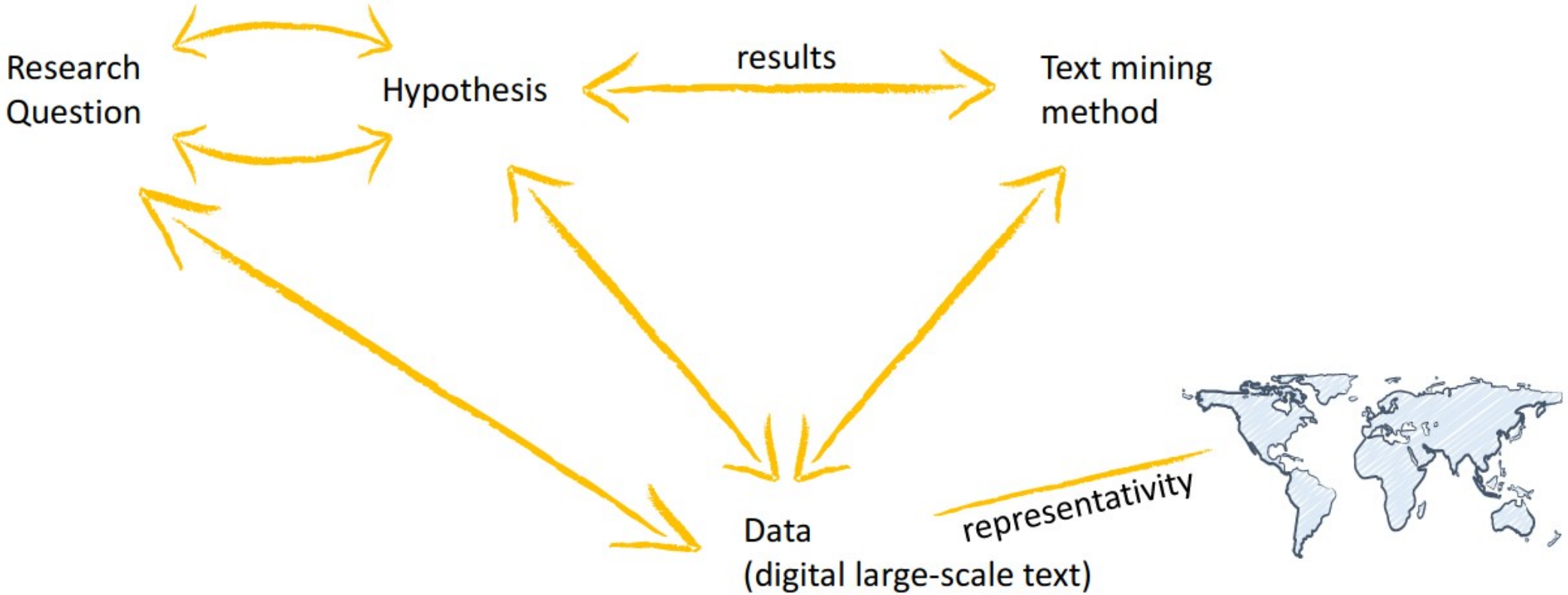
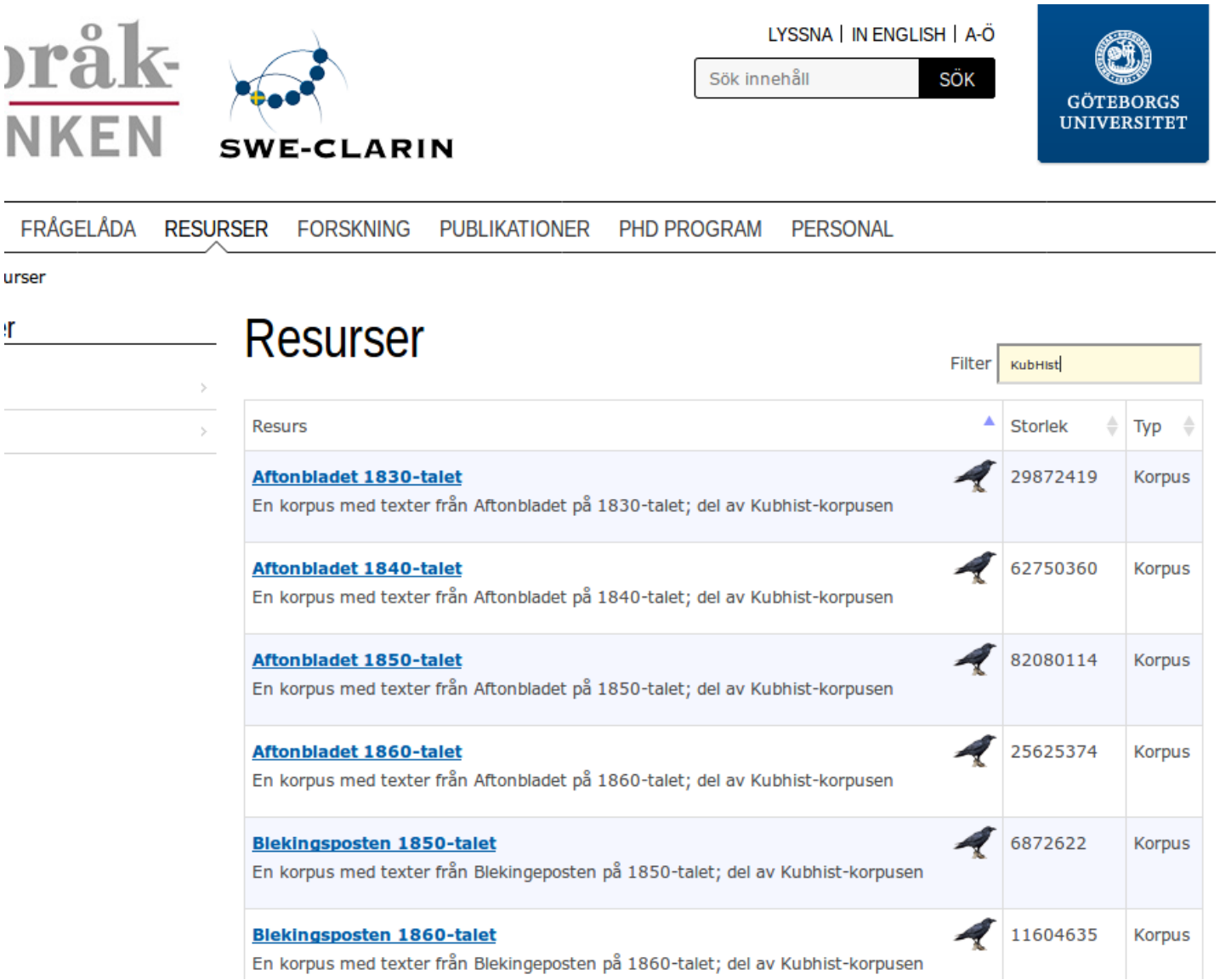


Figure 1: A schematic model of the research process in data-intensive humanities.

SWE-CL

https://spraakbanken.gu.se/swe/resurser



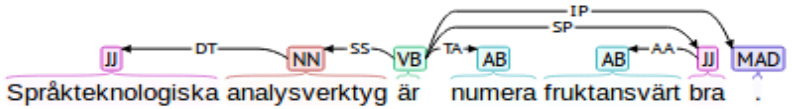
SWE-CLARIN

Att tillföra språklig information

SWE-CLARIN

Verktygen ger många olika slags språklig information

<sentence id="8f74-86ba"> [Visa XML]



token	msd	lemma	lex	sense
Språkteknologiska	JJ. POS. UTR+NEU. PLU. IND+DEF. NOM			
analysverktyg	NN. NEU. PLU. IND. NOM			
är	VB. PRS. AKT	vara	vara..vb.1	vara..1
numera	AB	numera	numera..ab.1	numera..1
fruktansvärt	AB. POS	fruktansvärd, fruktansvärt	fruktansvärd..av.1, fruktansvärt..ab.1	fruktansvärt..1 (0.796), fruktansvärd..1 (0.204)
bra	JJ. POS. UTR+NEU. SIN+PLU. IND+DEF. NOM	bra	bra..av.2	bra..4
.	MAD			

</sentence>

SWE-CLARIN

Pågående projekt vid LiU

- Handikappbegreppets utveckling
- Verktyg för textkomplexitetsmätningar
- Analys av hur standarder används
- Analys av betydelseskiften på Pinterest
- …

SWE-CLARIN

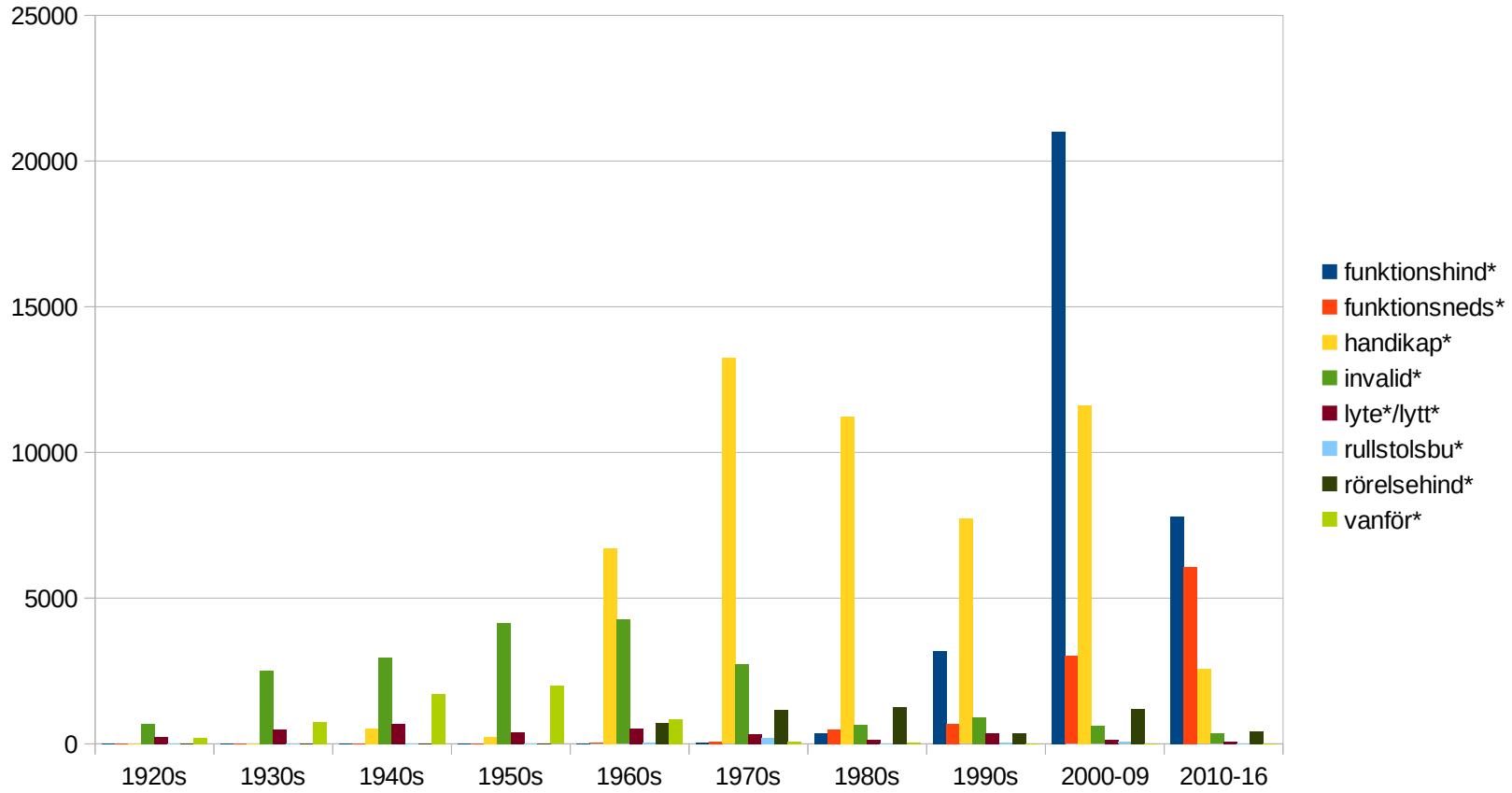
100 år av handikappbegrepp

- Projekt med IBL (Henrik Danielsson, Lotta Holme)
- Initailt analyserar vi SOU:er från 1922 och framåt
- Data hämtat via Språkbankens XML-version

Ahrenberg et al. (forthc.).

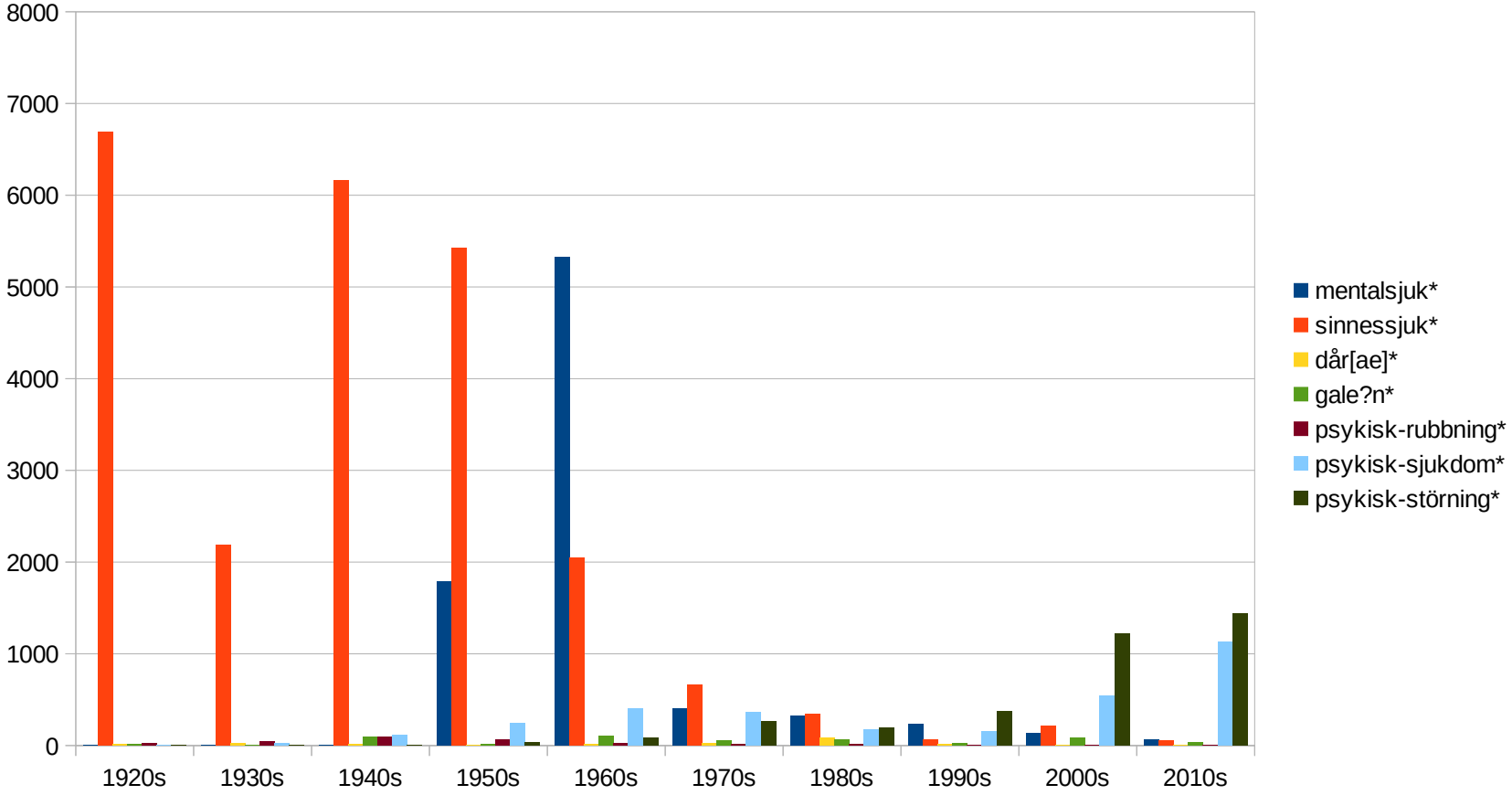
SWE-CLARIN

Fysiskt handikapp



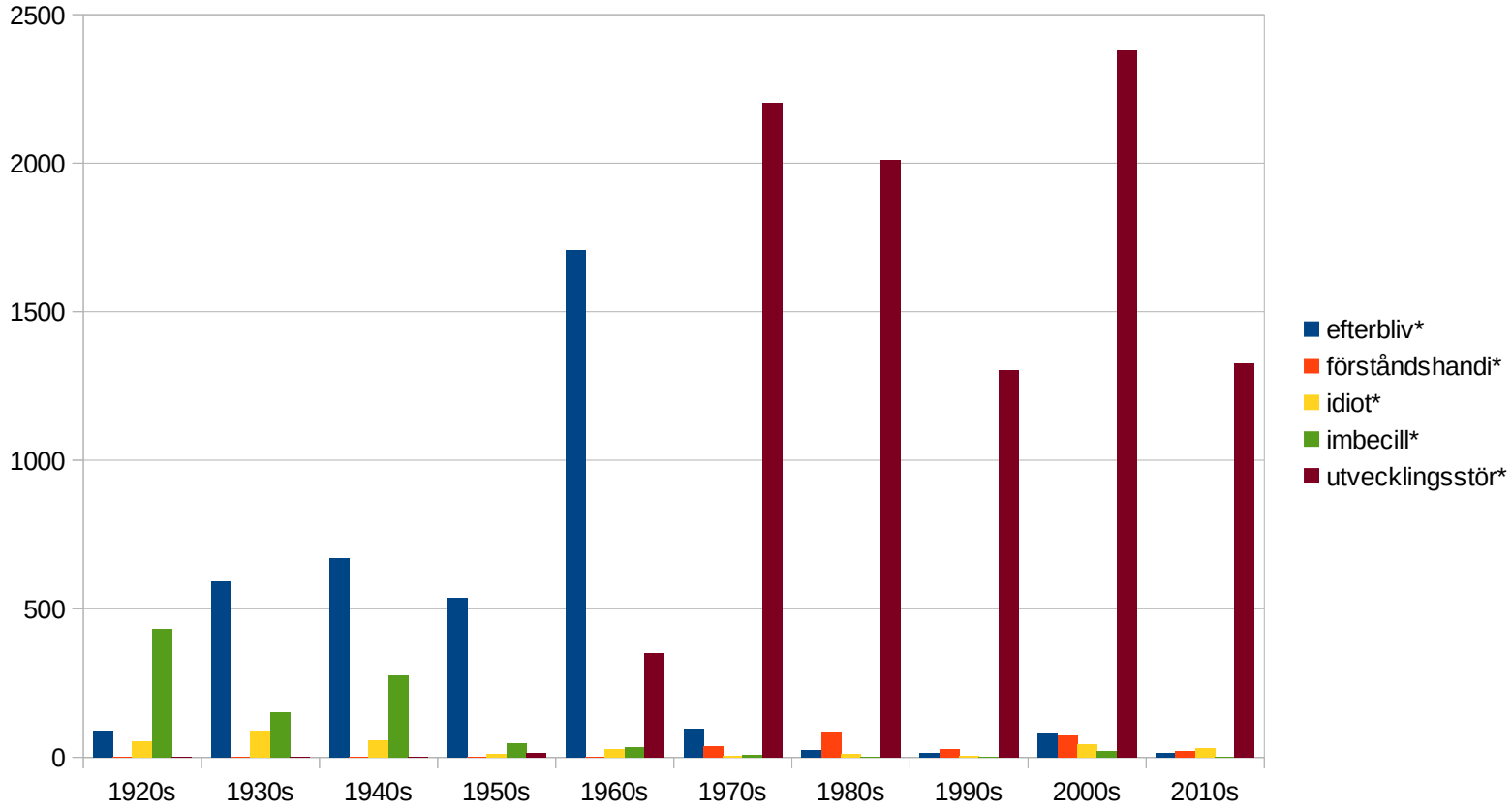
SWE-CLARIN

Psykiskt handikapp



SWE-CLARIN

Förståndshandikapp



SWE-CLARIN

Ordlikheter, mha vektorlikhet

SOU-dokument 1922-29:

- >>> wrd = '*imbecilla*'
- [('sinnesslöa', 0.8798)
- 'asociala', 0.8652)
- 'idioter', 0.8225)
- 'fallandesjuka', 0.7944)
- 'obildbara', 0.7874)
- 'straffriförklarade', 0.7550)
- 'bildbara', 0.7478)
- 'sinnessjuka', 0.7309)
- 'vanartade', 0.7046)
- 'alkoholist', 0.6863)]

- SOU-dokument 1950-59:
- >>> wd = '*imbecilla*'
- [('gravt', 0.8649)
- 'debila', 0.8018)
- 'konvalescenter', 0.7833)
- 'obildbara', 0.7811)
- 'tuberkulossjuka', 0.7750)
- 'utvecklingshämmade', 0.7739)
- 'missanpassad', 0.7715)
- 'uppegående', 0.7683)
- 'defekta', 0.7635)
- 'synsvaga', 0.7590)]

Motsvarigheter över tid (temporala analogier)

70-tal	80-tal	90-tal	00-tal	10-tal
handikapp	*handikapp*	*handikapp*	*funktionsned-sättningar*	*funktionsned-sättningar*
funktionshinder	*sjukdomstillstånd*	*sjukdomar*	*sjukdomar*	*funktionsned-sättning/ar*
	funktionshinder	*handikapp*	*funktionsned-sättningar*	*funktionsned-sättningar*

SWE-CLARIN

Verktyg för textkomplexitetsmätningar

- Ordnivå
  - Svåra, tvetydliga, långa ord och begrepp
  - Andra språk
  - Förkortningar
- Meningsnivå
  - Långa/Svåra meningar
  - Andel bisatser
- Övergripande
  - Längd, innehåll, variation

SWE-CLARIN

STePS / Analys / Text - Mozilla Firefox

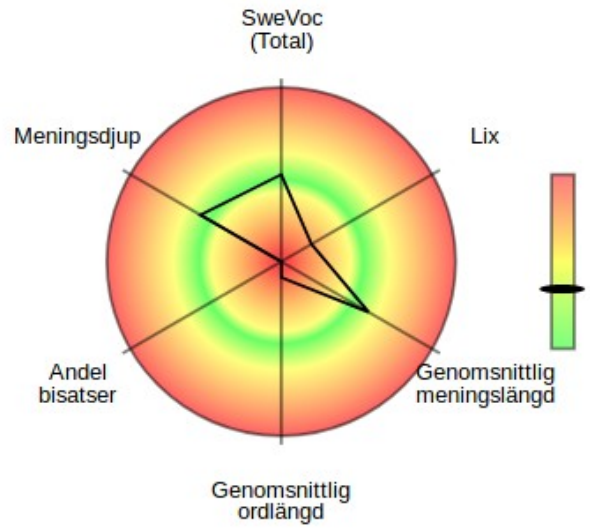
STePS / Analys / Text

localhost/analys/text/

STePS Hem Analys Historik Om

Analys

Visualisering



Ytliga mått

Ladda ner resultat

Original text

STePS: Swedish Text Profiling System
Vilket är förhållandet mellan text och språk? Med språkteknologiska resurser av olika slag kan vi fånga diverse språkliga egenskaper hos en text framför allt utifrån ord och meningar i termer av frekvenser och automatiska analyser av ordklasser, ordförråd och grammatiska relationer. I projektet STePS vill vi utveckla ett system som ger textforskare av olika slag möjlighet att använda och anpassa olika automatgenererade textmått för egna behov. STePS utnyttjar resultat från projektet Diginclude men utökar och anpassar dem för andra målgrupper, specifikt textforskare. Idag finns ett

SWE-CLARIN

Korpusinsamling

- En samling av parallellställda meningar från svenska kommun- och myndighetshemsidor, där lättlästa meningar parats ihop med meningar skrivna på standardsvenska
- Syfte: Textanalys avseende bl. a. textkomplexitet, samt utveckling av metoder för automatisk textförenkling
- 59 513 meningspar

SWE-CLARIN

Textanalys av standards

- Projekt med IEI
- Analys av The Economist
  - Vilka betydelser har standards för företag, industrier och samhället?
  - Hur har detta förändrats sedan 1990?
  - Vad föranledde dessa förändringar?

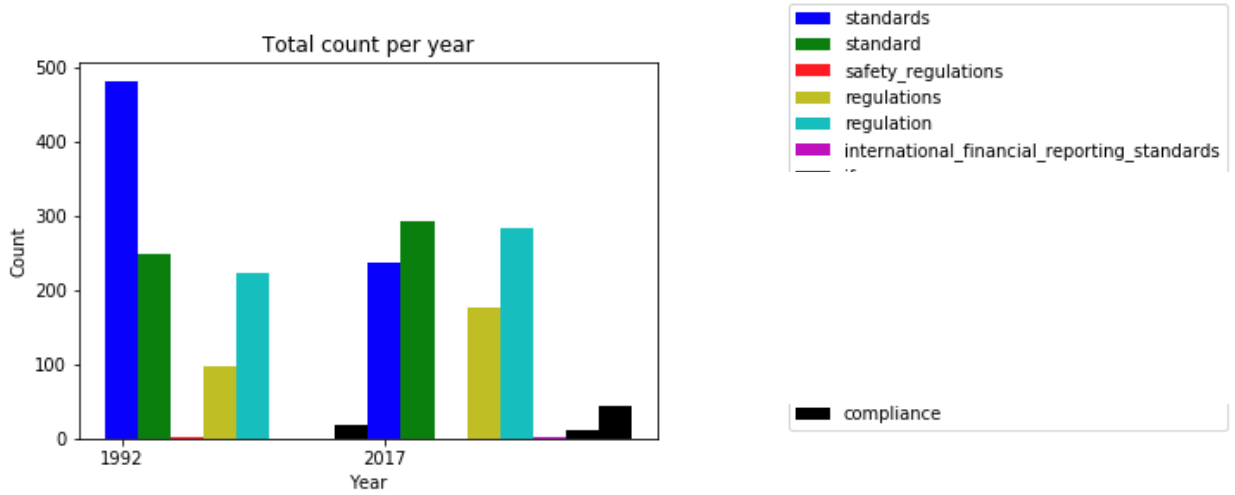
SWE-CLARIN

1992

standard	standards	regulation	regulations	compliance	safety_regulations
249	481	223	98	18	3

2017

standard	standards	regulation	regulations	compliance	ifrs	international_financial_reporting_standards
292	237	284	175	43	12	2



Article date: Nov 21, 1992

Title: Homosexuals and the Military: Out of the Locker

Words: standards: 1, regulations: 1

Abstract: The case involving Keith Meinhold, a USN petty officer discharged from the service for admitting he was gay, is discussed. Pres-elect Bill Clinton has suggested that barring homosexuals from the service should be overturned.

Sentences:
He acknowledges that the present policy discriminates unfairly, but argues, with soldierly tact, that "civilian **standards** of fairness and equality don't apply down where the body bags are filled.
Mood: indicative Modality: 0.75 Sentiment: (-0.25, 0.5)

There is no reason why homosexuals, any more than heterosexuals, should roam bases like ravening wolves; existing military **regulations** prohibit all sexual relations, regardless of slant.
Mood: conditional Modality: 0.35 Sentiment: (0.3, 0.4777777777777778)

SWE-CLARIN

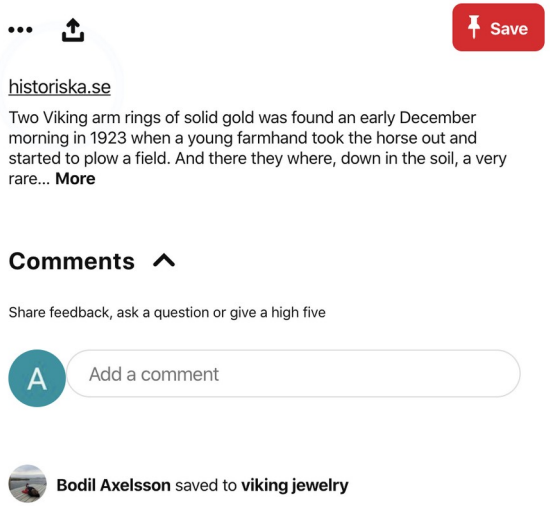
Korpusinsamling

- Rapporter från svenska företag
  - 629 195
  1) ISO-certifierade
  2) Nämner standarden i sina rapporter utan att vara certifierade
  3) Nämner inte
- Skiljer sig språket mellan dem
  - Vokabulär
  - Syntax
  - Attityd (sentiment)
  - Språklig profil?

SWE-CLARIN

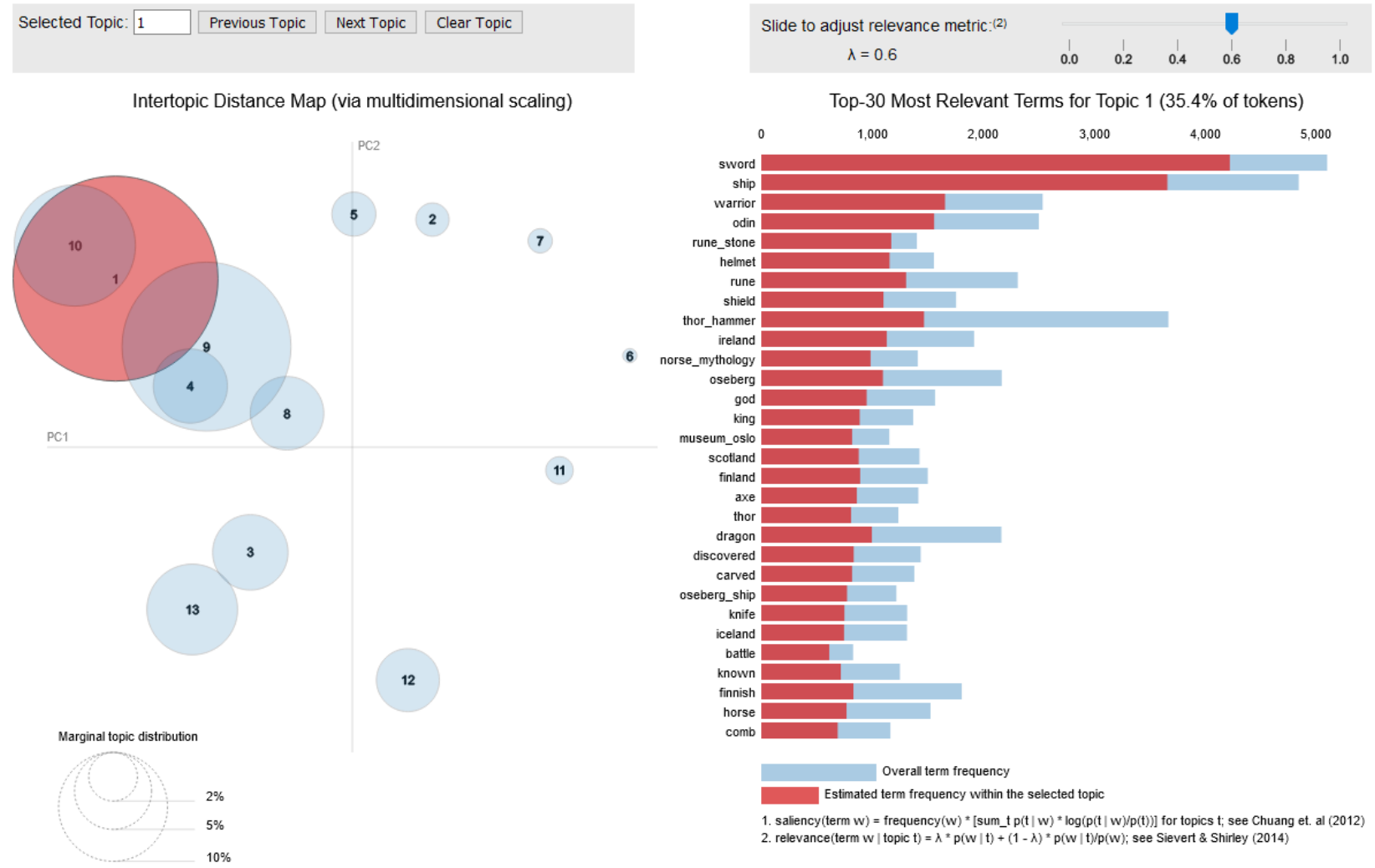
Analys av Pinterest

- Hur påverkas Historiska museets föremål betydelse när de byter kontext
- Yrkespersoner och amatörer skapar tillsammans nya grupperingar av föremål
- Topikanalys av de språkliga beskrivningarna



SWE-CLARIN

Topic i Pinterest



SWE-CLARIN

Preliminära resultat

- Identifierat 13 topics
- Ett tydligt topic kring den klassiska manliga myten
  - Svärd, tors hammare, knivar, slag, makt, yxor, hjälmar, sköldar etc
- Ett topic med mer fokus på kläder
  - Tunikor, väskor, broderi, stygn, kostymer, kläder, silke, ylle etc
- Ytterligare topic identifierade

SWE-CLARIN

Litteratur

- ChengXiang Zhai and Sean Massung, 'Text Data Management and Analysis', Morgan & Claypool, 2016: 3-13.
- Nina Tahmasebi and Simon Hengchen, 'The Strengths and Pitfalls of Large-Scale Text Mining for Literary Studies', Samlaren, 2019.
- Lars Ahrenberg, Henrik Danielsson, Hampus Arvå, Staffan Bengtsson, Lotta Holme and Arne Jönsson (forthc.) Studying Disability-Related Terms with Swe-Clarin Resources. To appear in Proceedings of the 2019 Clarin Annual Conference.
- J. Jarlbrink, P. Snickars och C. Colliander, 2016. Maskinläsning: om massdigitalisering, digitala metoder och svensk dagspress
- Eva Pettersson, Jonas Lindström, Benny Jacobsson, Rosemarie Fiebranz, 2016: HistSearch - Implementation and Evaluation of a Web-based Tool for Automatic Information Extraction from Historical Text. Proceedings of the 3rd HistoInformatics Conference, Krakow, Poland, 11 July, 2016.

SWE-CLARIN