

Big data, språkteknologi och digitala metoder

Lars Ahrenberg, Arne Jönsson

Översikt

- Språkteknologi, text mining och andra begrepp
- Att utveckla korpusar
- Att bygga statistiska modeller
- Några projekt med LiU-forskare

Praktiska detaljer

Laborationer

- Utförs i grupper om 2 studenter. Enklast är det att de befintliga uppsatsparen labbar ihop.
- Det är viktigt att ni kommer väl förbredda till de 2 laborationstillfällena. Läs igenom anvisningarna och försök gärna göra laborationerna på egen hand.
- Handledning i **TEAMS** under de två laborationspassen och där har varje grupp en egen kanal för att ställa frågor.

Redovisning

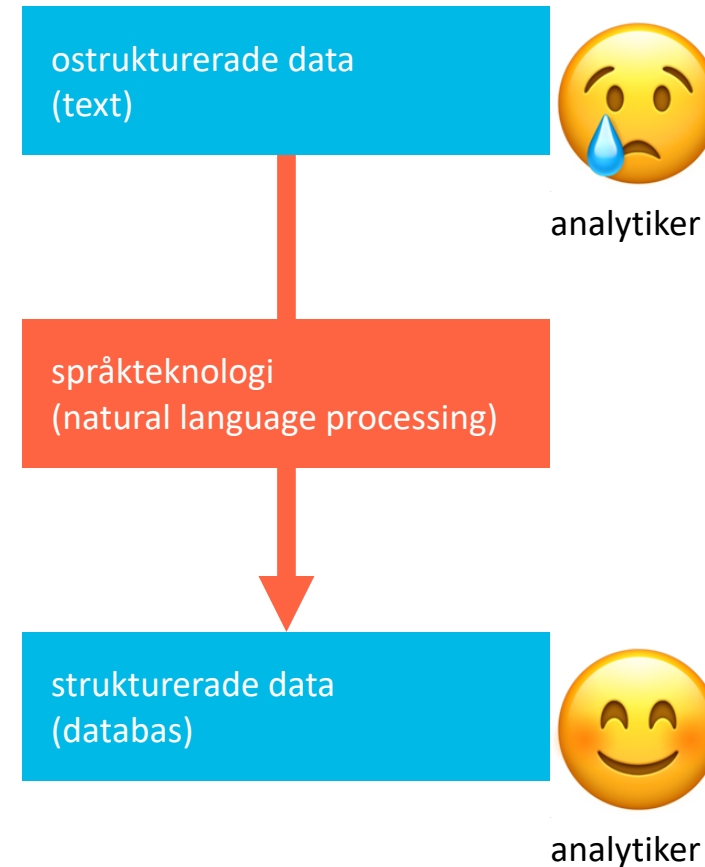
- Momentet är obligatoriskt och redovisas genom att, ladda upp en fil med svar på de frågor som finns under respektive labb.
- Omfattning ungefär en halv till en A4-sida per laboration är tillräckligt.

Vad är språkteknologi?

- **Språkteknologi** är teknologi för att analysera och generera naturligt (mänskligt) språk.
 - inte programmeringsspråk!
- Språkteknologi är ett tvärvetenskapligt forskningsområde som berör datavetenskap, lingvistik, kognitionsvetenskap, matematisk statistik, m.fl.
 - andra namn: datorlingvistik, text mining, natural-language processing, ...
- Två inriktningar:
 - Språkvetenskap / lingvistik
 - Tillämpningar

Kunskapsglappet

‘We are drowning in
information
but starved for knowledge.’
John Naisbitt (1982)



Informationsutvinning

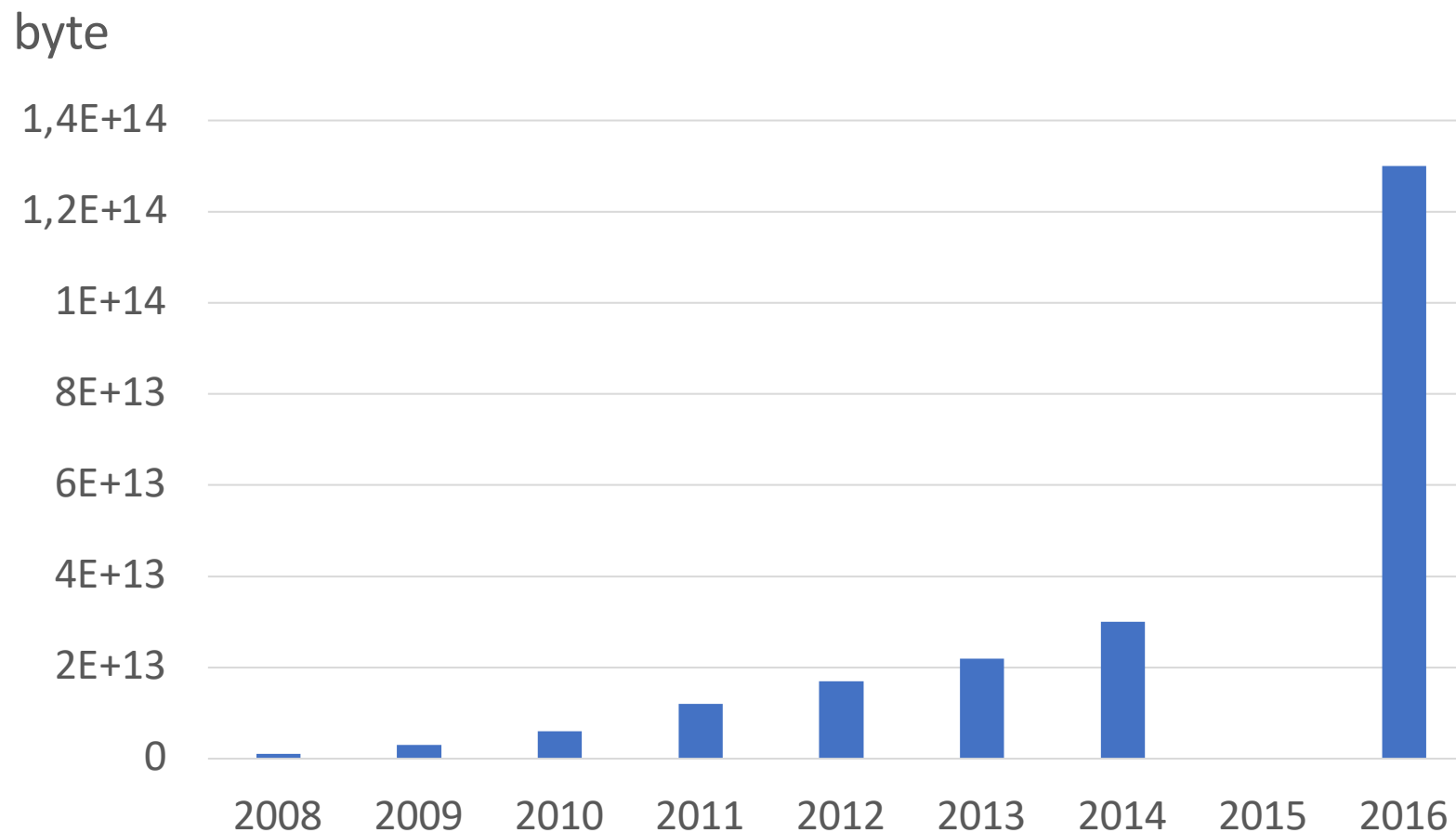
- As of 15 Mar 2002, Hawaii state health officials reported one additional recent case of dengue fever and 6 cases that occurred last year but were not confirmed by laboratory testing until 2002.

Attribut	Värde
docno	ProMed.20020322.11
doc_date	2002.03.22
disease_name	dengue fever
norm_stime	2002.03.15
norm_etime	2002.03.15
victim_types	—
location	Hawaii

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,

Antalet sidor indexerade av Google



Kommersiellt intresse

Diamond Level



ASAPP



Microsoft

Bloomberg
Engineering



facebook

Google

Platinum Level



amazon

IBM Research AI

Tencent 腾讯



Baidu 百度

Huvudsponsorer av konferensen ACL 2019

Korpuslingvistik

Korpuslingvistik

- Korpuslingvistik handlar om att undersöka språkvetenskapliga frågor utifrån språkliga data (korpusar).
- Ett grundantagande är att språkvetenskapliga undersökningar bör baseras på texter från fältet, i deras naturliga kontexter.

Vad är en korpus?

- **korpus** s. *-en -ar*, (språkv.) mängd av undersökt textmaterial
 - Svenska Akademiens ordlista
- A body of texts, utterances or other specimens considered more or less representative of language and usually stored electronically.
 - The Oxford Companion to the English Language

Några viktiga språkvetenskapliga korpusar

Namn	År	Storlek	Språk och sort
British National Corpus (BNC)	1994	100 miljoner ord	brittisk engelska, talad och skriven
American National Corpus (ANC)	2003	22 miljoner ord	amerikansk engelska, talad och skriven
Corpus of Contemporary American English (COCA)	2008	425 miljoner ord	amerikansk engelska, talad och skriven
Stockholm–Umeå Corpus (SUC)	2006	1 miljon ord	svenska, skriven

Länkar till korpusar och korpusverktyg

- [Stockholm–Umeå Corpus](#) (Språkbanken)
- [Korp \(sökvertyg\)](#) (Språkbanken)
- [Google Books Ngram Viewer](#)
- [Universal Dependencies Project](#)

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

pråk
NKEN



LYSSNA | IN ENGLISH | A-Ö

Sök innehåll

SÖK



FRÅGELÅDA RESURSER FORSKNING PUBLIKATIONER PHD PROGRAM PERSONAL

urser

ur

Resurser

Filter KubHist

Resurs	Storlek	Typ
Aftonbladet 1830-talet En korpus med texter från Aftonbladet på 1830-talet; del av Kubhist-korpusen	29872419	Korpus
Aftonbladet 1840-talet En korpus med texter från Aftonbladet på 1840-talet; del av Kubhist-korpusen	62750360	Korpus
Aftonbladet 1850-talet En korpus med texter från Aftonbladet på 1850-talet; del av Kubhist-korpusen	82080114	Korpus
Aftonbladet 1860-talet En korpus med texter från Aftonbladet på 1860-talet; del av Kubhist-korpusen	25625374	Korpus
Blekingeposten 1850-talet En korpus med texter från Blekingeposten på 1850-talet; del av Kubhist-korpusen	6872622	Korpus
Blekingeposten 1860-talet En korpus med texter från Blekingeposten på 1860-talet; del av Kubhist-korpusen	11604635	Korpus

Historisk bakgrund

- Sedan mitten av 1900-talet är språkvetenskap i huvudsak en deskriptiv vetenskap.
 - tidigare ofta preskriptiv
- Precis som i andra humaniora kan språkvetenskapliga analyser grundas i data och kvantitativa metoder.
 - Bloomfields fältstudier, Shannons kodningsteori

Historisk bakgrund

Utvecklingen av kvantitativa och statistiska metoder inom språkvetenskapen avstannade under 1950-talet:

- Det fanns inte tillräckligt mycket språkdata för att stödja intressanta statistiska generaliseringar.
- Humaniora rörde sig bort från data-orienterade beskrivningar mot introspektiva modeller av kognition.
 - Chomsky

Kvantitativ vetenskap

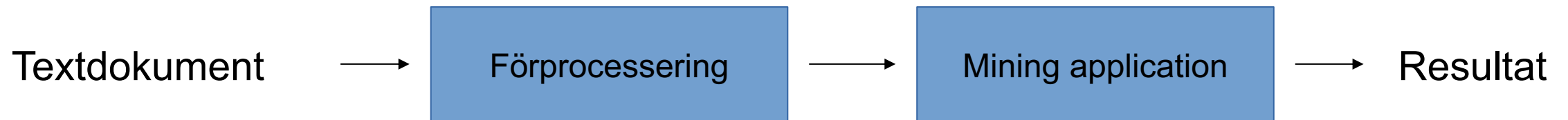
‘When measurement departs from theory, it is likely to yield mere numbers, and their very neutrality makes them particularly sterile as a source of remedial suggestions.

But numbers register the departure from theory with an authority and finesse that no qualitative technique can duplicate, and that departure is often enough to start a search.’

Thomas Kuhn (1961)

Text mining

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



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

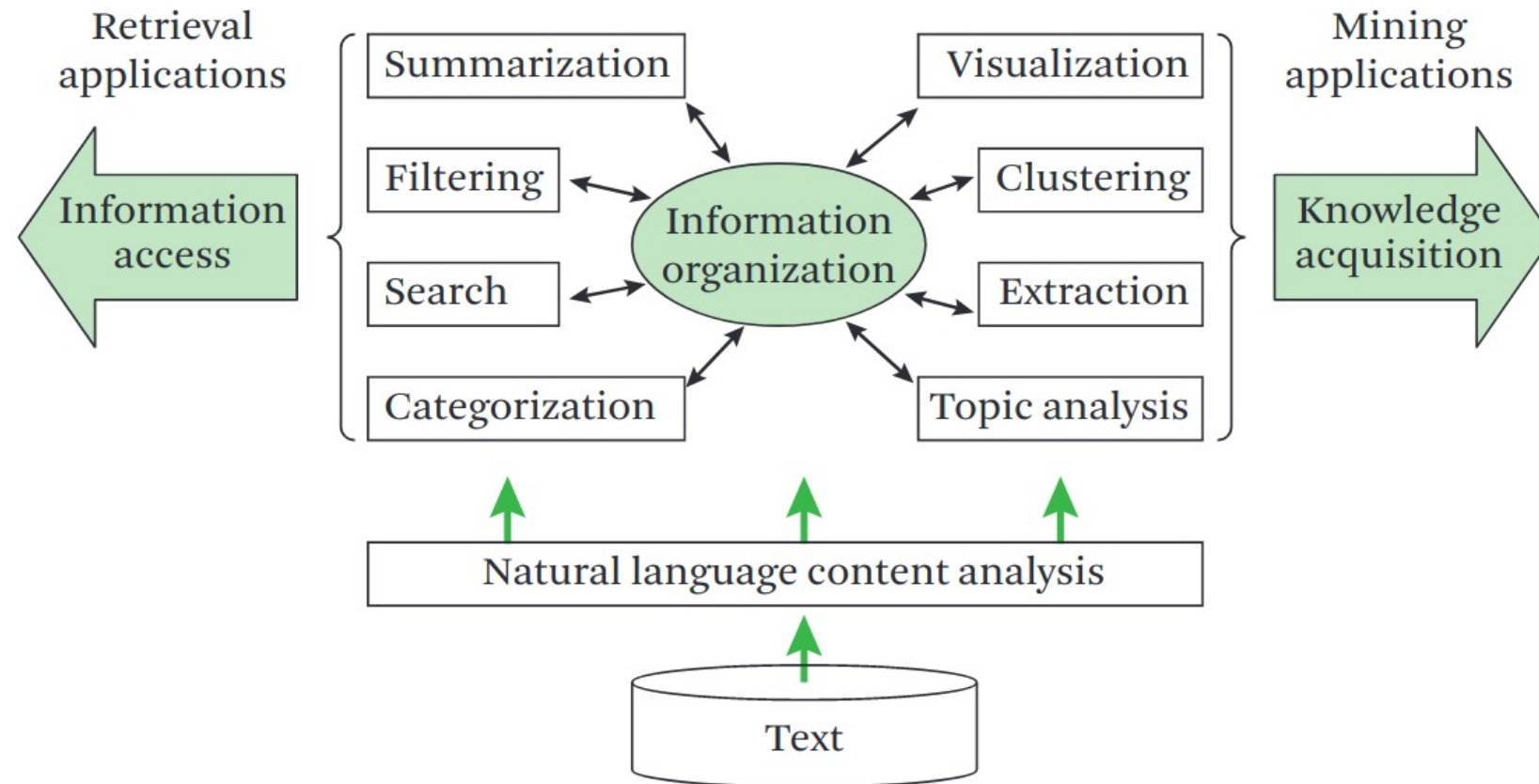


Figure 1.3 Conceptual framework of text information systems.

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

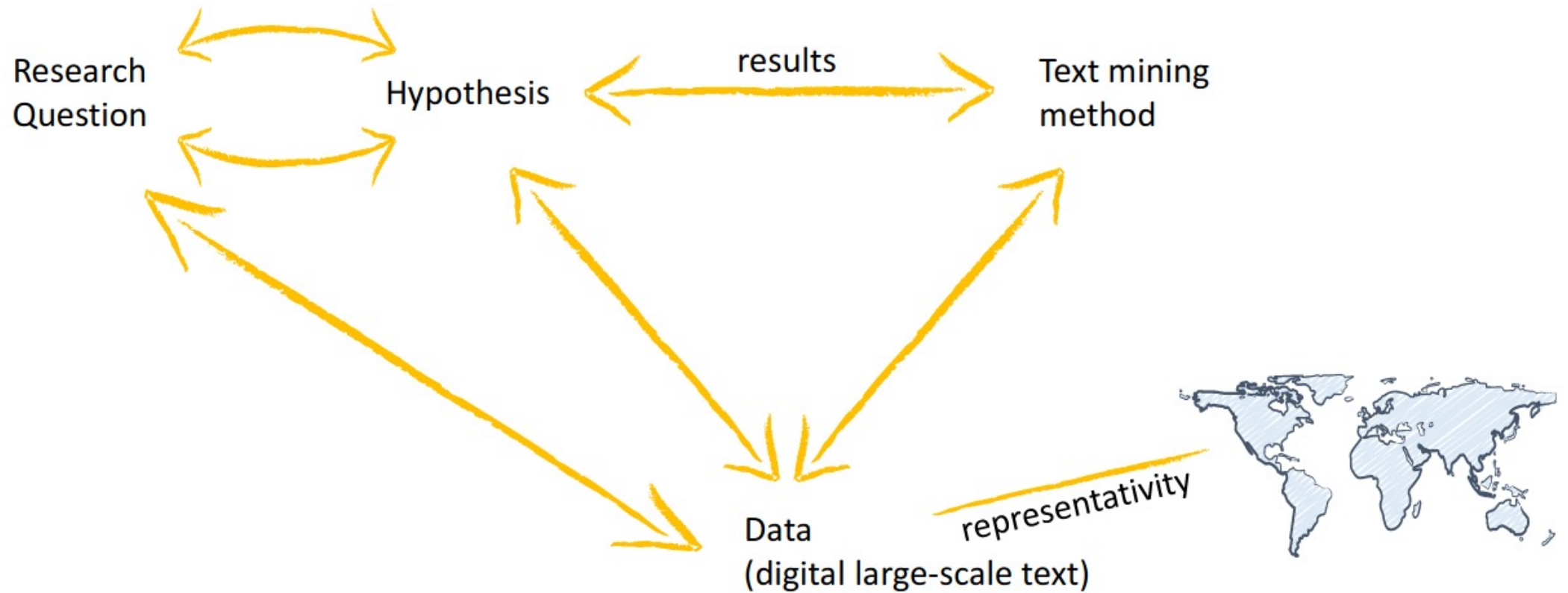


Figure 1: A schematic model of the research process in data-intensive humanities.

Att bygga korpusar

Från råmaterial till annoterad text

Aktivitet	Beskrivning
Urval	Välja ut de texter som ska vara med i korpusen.
Insamling	Samla in texterna, t.ex. genom att ”spindla” webben.
Avformatering	Ta bort strukturell uppmärkning, t.ex. XML.
Segmentering	Dela upp textmaterialet i relevanta enheter.
Annotering	Lägg till relevant information, t.ex. ordklasser.

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?")



WIKIPEDIA
Den fria encyklopedin

[Skriv en ny artikel](#)

Bybrunnen

Slumpartikel

Stöd Wikipedia

[Hjälp](#)

Skriv ut/exportera

Ladda ner som PDF

Sidor som länkar hit

Specialsidor

Sidinformation

Citera denna sida

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99

 Bilder & media

Visa

Redigera

Redigera wikitext

Visa historik

Sök

Q

Denna artikel avhandlar arten trädkrypare. För familjen Certhiidae se [Trädkrypare \(familj\)](#)

Trädkrypare (*Certhia familiaris*) är en fågelart inom familjen trädkrypare (Certhiidae).

Den liknar andra arter inom familjen med böjd näbb, mönstrad brun ovalsida, vitaktig undersida och långa styva stjärtpenor som den använder för kunna balansera upprätt på trädstammar och grenar. Den kan lättast urskiljas från den likartade [trädgårdsträdskryparen](#), som delar en stor del av dess utbredning i [Europa](#), genom skillnaden i sång.

Trädskryparen har nio eller fler **underarter** som häckar i olika delar av dess utbredningsområde i de **tempererade** delarna av **Eurasien**. Arten återfinns i alla sorters skogsmarker, men där den överlappar med trädgårdsträdskryparen i Västeuropa återfinns den med större sannolikhet i **barrskogar** eller på högre höjder. Den häckar i trädhålor eller bakom barkflagor, och föredrar introducerade **mammuträd** som häckningsplatser där sådana finns tillgängliga. Honan lägger vanligen fem eller sex vita ägg med rosa prickar i det fodrade boet, men ägg och ungar är sårbara för angrepp från hackspettar och däggdjur, däribland ekorrar.

Trädskryparen är insektsätare och klättrar uppför trädstammar som en mus, för att leta efter insekter som den plockar från skrevor i barken med sin fina böjda näbb. Den flyger sedan till foten av ett annat träd med en distinkt oregelbunden flykt. Fågeln lever ensam på vintern, men kan bilda gemensamma sovplatser

Status i världen: Livskraftig (lc)^[1]

Status i Sverige: Livskraftig



Domän

Eukaryoten

Trädkrypare - Wikipedia

sv.wikipedia.org/wiki/Trädkrypare

Reader

WIKIPEDIA

Den fria encyklopedin

Huvudsida

Skriv en ny artikel

Deltagarportalen

Bybrunnen

Senaste ändringarna

Slumpartikel

Ladda upp filer

Stöd Wikipedia

Kontakta Wikipedia

Hjäl

Artikel

Diskussion

Visa

Redigera

Redigera wikitext

Visa historik

Sök

Trädkrypare ★

Denna artikel avhandlar arten trädkrypare. För familjen Certhiidae se [Trädkrypare \(familj\)](#)

Trädkrypare (*Certhia familiaris*) är en fågelart inom familjen trädkrypare (Certhiidae).

Den liknar andra arter inom familjen med böjd näbb, mönstrad brun ovsida, vitaktig undersida och långa styva stjärtpennor som den använder för kunna balansera upprätt på trädstammar och grenar. Den kan lättast urskiljas från den likartade [trädgårdsträdkryparen](#), som delar en stor del av dess utbredning i [Europa](#), genom skillnaden i sång.

Trädkrypare

Status i världen: Livskraftig (lc)^[1]

Status i Sverige: Livskraftig



38

1

0

57

Search Resource Content

Trädkrypare — sv.wikipedia.org

Images

Scripts

Stylesheets

Cookies — sv.wikipedia.org

Extension Scripts

Local Storage — sv.wikipedia.org

Session Storage — sv.wikipedia.org

Filter Resource List

div#mw-cont...

p

T

"Den liknar andra arter inom fam..."

<table class="infobox" cellspacing="5" style="width: 22em; text-align: left; font-size: 88%; line-height: 1.5em; font-size: 95%; width: 20.3788em; padding: .18em;"></table>

<p>...</p>

<p>

"Den liknar andra arter inom familjen med böjd näbb, mönstrad brun ovsida, vitaktig undersida och långa styva stjärtpennor som den använder för kunna balansera upprätt på trädstammar och grenar. Den kan lättast urskiljas från den likartade "

trädgårdsträdkryparen

", som delar en stor del av dess utbredning i "

Europa

", genom skillnaden i sång."

</p>

<p>...</p>

<p>...</p>

<p>...</p>

<div id="toc" class="toc"></div>

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>

Textsegmentering

- Textsegmentering är uppgiften att dela upp en text i lingvistiskt meningsfulla enheter, såsom ord, meningar och paragrafer.
- När enheterna man segmenterar i är ord eller ordliknande enheter kallas textsegmentering för tokenisering.
 - ord, tal, skiljetecken

Tokenisering

- Rå text
 - Den liknar andra arter inom familjen med böjd näbb, mönstrad brun ovansida, vitaktig undersida och långa styva stjärtpennor som den använder för att kunna balansera upprätt på trädstammar och grenar.
- Tokeniserad text
 - Den liknar andra arter inom familjen med böjd näbb , mönstrad brun ovansida , vitaktig undersida och långa styva stjärtpennor som den använder för att kunna balansera upprätt på trädstammar och grenar .

Meningssegmentering

- I vissa sammanhang vill vi inte bara identifiera ord utan även större enheter såsom meningar eller stycken.
- Meningssegmentering är uppgiften att dela upp en text i meningar.
- Meningssegmentering är svårare än att dela upp en text efter en punkt eller ett annat skiljetecken.
 - *We visited the U.S. After that, we visited Canada.*

Att tillföra språklig information



Sparv

Språkbankens annoteringsverktyg

Analysspråk:

svenska ▼

Ladda exempel:

 Drama

 Åtta sido

Mata in

Ladda upp

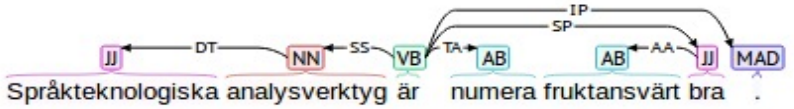
☒ Ren text

☐ XML

1 Språkteknologiska analysverktyg är numera fruktansvärt bra.

Verktögen 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>

En stor utmaning – ambiguitet

Jag	bad	om	en	kort	bit
PN	VB	PP	DT	JJ	NN
NN	NN	SN	PN	AB	VB
		PL	RG	NN	
		AB	NN		

Exempel: Joakim Nivre

En stor utmaning – ambiguitet

Jag	bad	om	en	kort	bit
PN	VB	PP	DT	JJ	NN
99,45%	80,93%	63,14%	95,26%	70,22%	100,00%
NN	NN	SN	PN	AB	VB
0,55%	19,61%	29,79%	2,37%	19,66%	0,00%
		PL	RG	NN	
		5,13%	2,28%	10,11%	
		AB	NN		
		1,92%	0,01%		

Data: Stockholm-Umeå Corpus (träningsdel)

Statistik och modeller

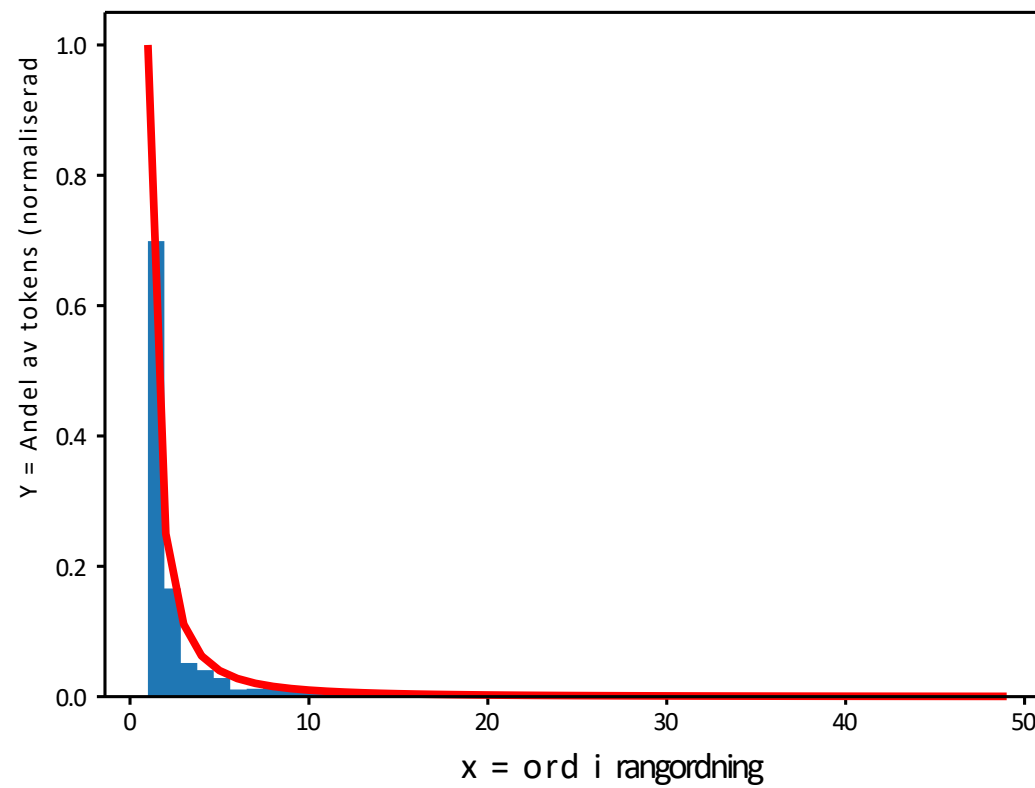
Frekvensanalys

- Ordrekvenser i t.ex. tidningskorpusar speglar en 'tidsanda' eller skribenters 'förgivettaganden'
- Relativa frekvenser är ofta mer intressanta än absoluta frekvenser
- Jämförelser över tid speglar språkutvecklingen (och samhällsutvecklingen)

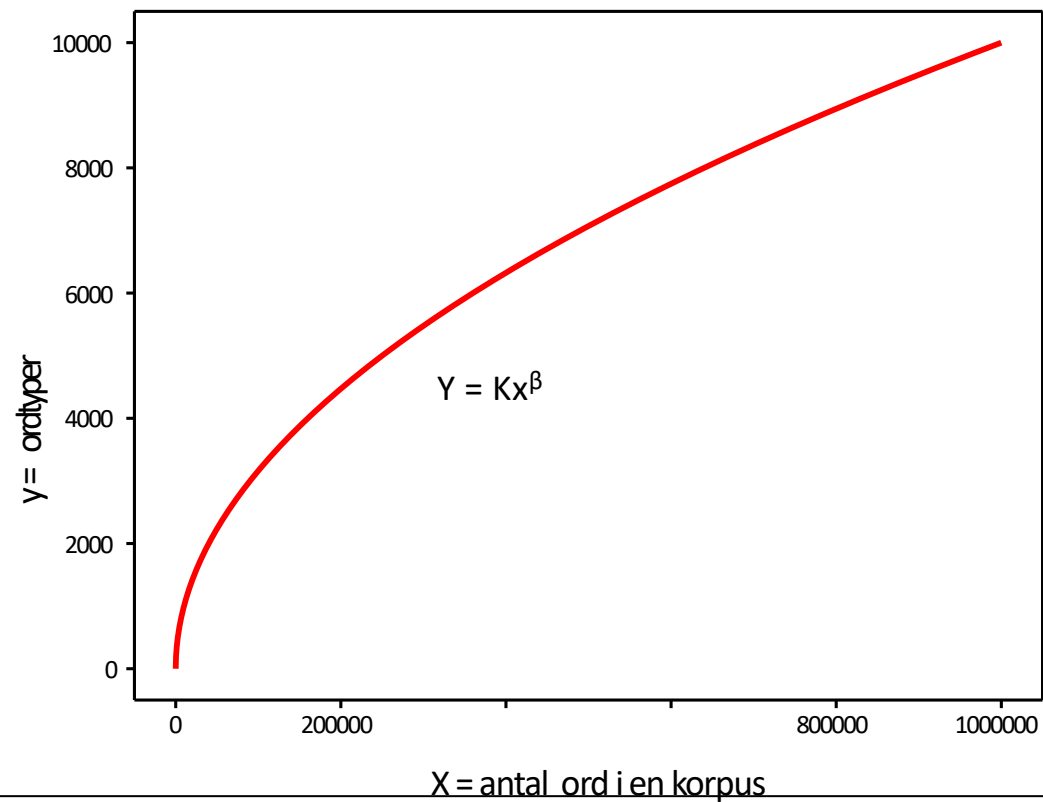
Frekvensanalys

- Man kan räkna mycket annat än ord:
 - bigram (= två ord i direkt ordningsföljd)
 - n-gram (= n ord i direkt ordningsföljd)
 - → författarbestämning, ...
 - samförekomster (= ord inom ett visst avstånd från varandra)
 - ord av semantiska kategorier
 - t.ex. positiva och negativa värdeord
 - → sentimentanalys, ...

Zipfs lag

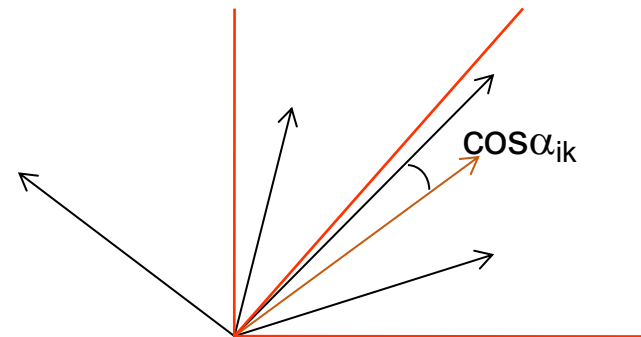


Heaps lag



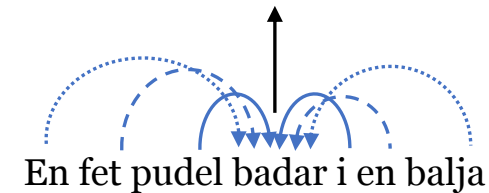
Vektorrumsmodeller

- Tekniker som reducerar den lingvistiska variationen och fångar semantiskt relaterade begrepp
- Ord representeras som vektorer (Word embeddings)
- Meningar fås genom att addera ordvektorer, dokument genom att addera meningsvektorer
- Likhet mellan meningar och dokument mäts som närhet i vektorrummet



Distribuerad semantik

- Ett ords betydelse beror på hur det används i kontext
- Varje ord representeras som en vektor
 - Några hundra dimensioner
 - Träningskorpus, några miljoner ord
- Exempel, random indexing



- Han bet hunden.
- Tilldela en gles slumpvektor till varje ord

Han	(	0	-1	0	1	0	-1	0	0	0	1)
bet	(	0	0	0	0	0	1	1	-1	-1	0)
hunden	(	-1	0	1	0	0	1	0	-1	0	0)

- Varje gång ett ord förekommer i texten, addera kontextvektor i ett fönster runt ordet
bet: $(0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1 \ -1 \ -1 \ 0) + (0 \ -1 \ 0 \ 1 \ 0 \ -1 \ 0 \ 0 \ 0 \ 1) + (-1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ -1 \ 0 \ 0) =$
 $(-1 \ -1 \ 1 \ 1 \ 0 \ 1 \ 1 \ -2 \ -1 \ 1)$

Några projekt vid LiU

Några projekt vid LiU

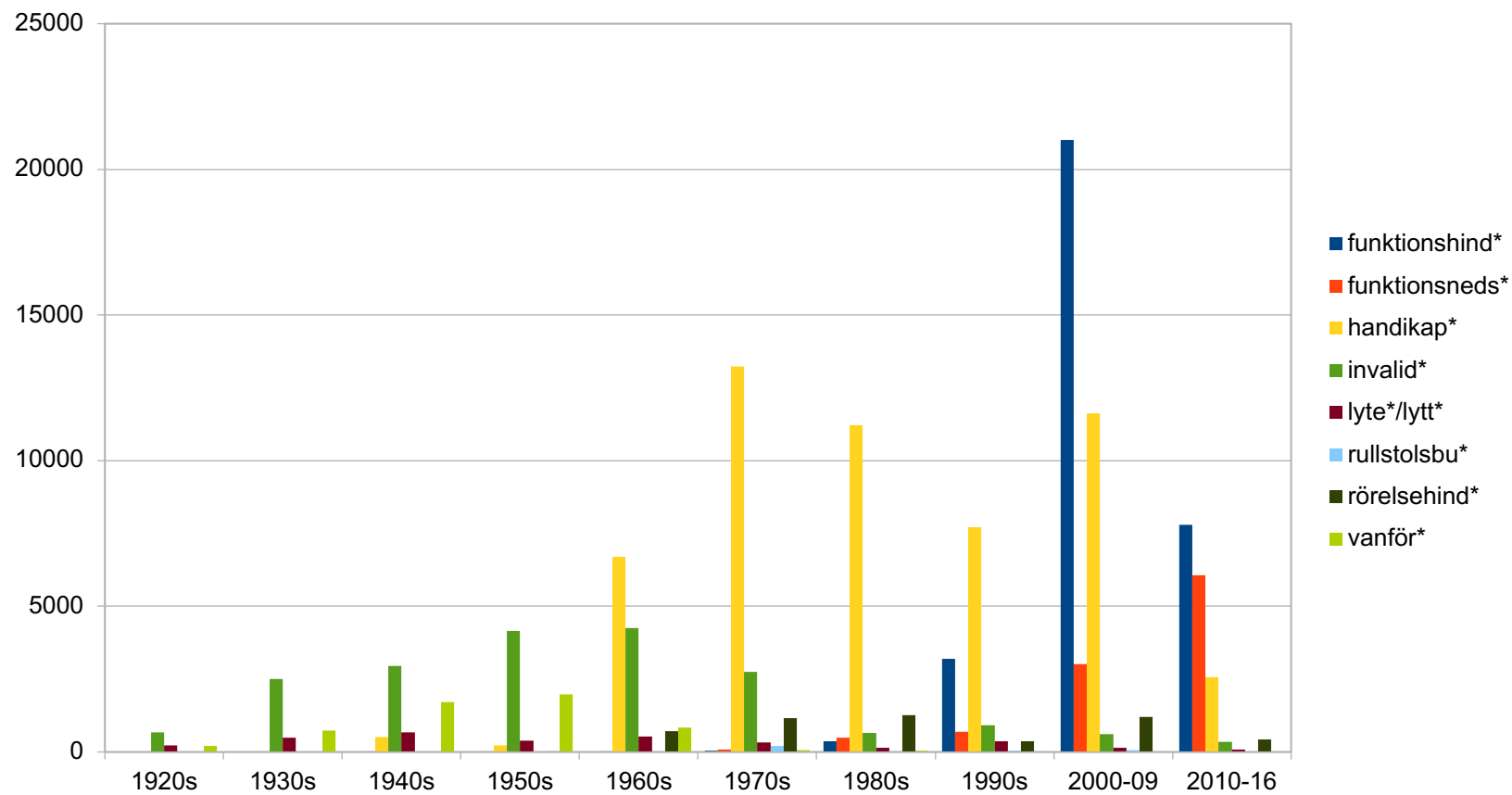
- Handikappbegreppets utveckling
- Verktyg för textkomplexitetsmätningar
- Analys av hur standarder används
- Analys av betydelseskillnaden på Pinterest

Projekten genomförs inom det nationella infrastrukturprojektet SweClarin
(<https://sweclarin.se/>)

100 år av handikappbegrepp

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

Fysiskt handikapp



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	<i>handikapp</i>	<i>handikapp</i>	<i>funktionsned-sättningar</i>	<i>funktionsned-sättningar</i>
funktionshinder	<i>sjukdomstillstånd</i>	<i>sjukdomar</i>	<i>sjukdomar</i>	<i>funktionsned-sättning/ar</i>
	funktionshinder	<i>handikapp</i>	<i>funktionsned-sättningar</i>	<i>funktionsned-sättningar</i>

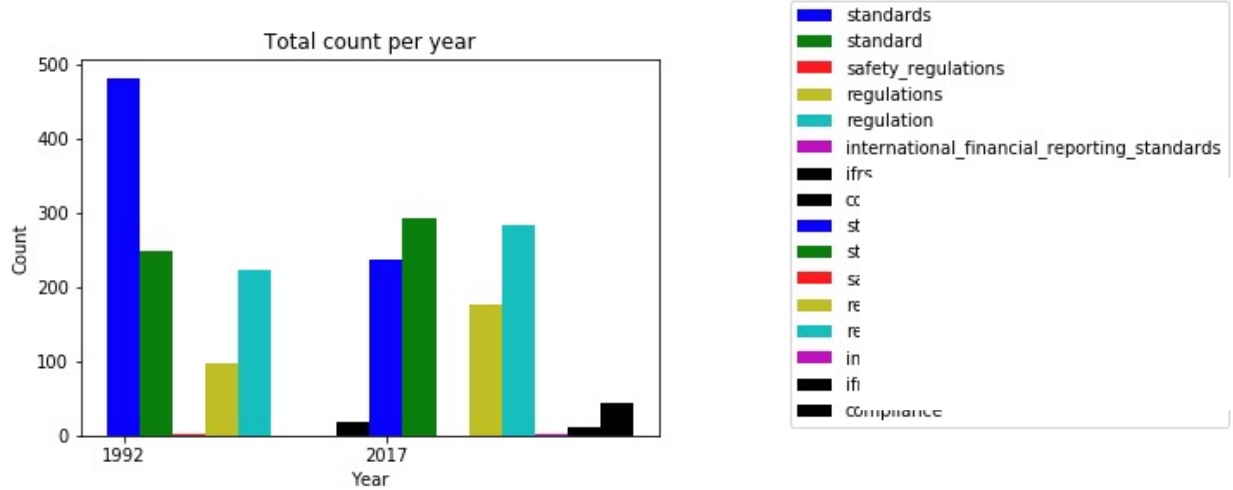
Insamling av lättläst text

- 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

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?

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)

Korpusinsamling

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

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



historiska.se

Two Viking arm rings of solid gold was found an early December morning in 1923 when a young farmhand took the horse out and started to plow a field. And there they where, down in the soil, a very rare... [More](#)

Comments ^

Share feedback, ask a question or give a high five

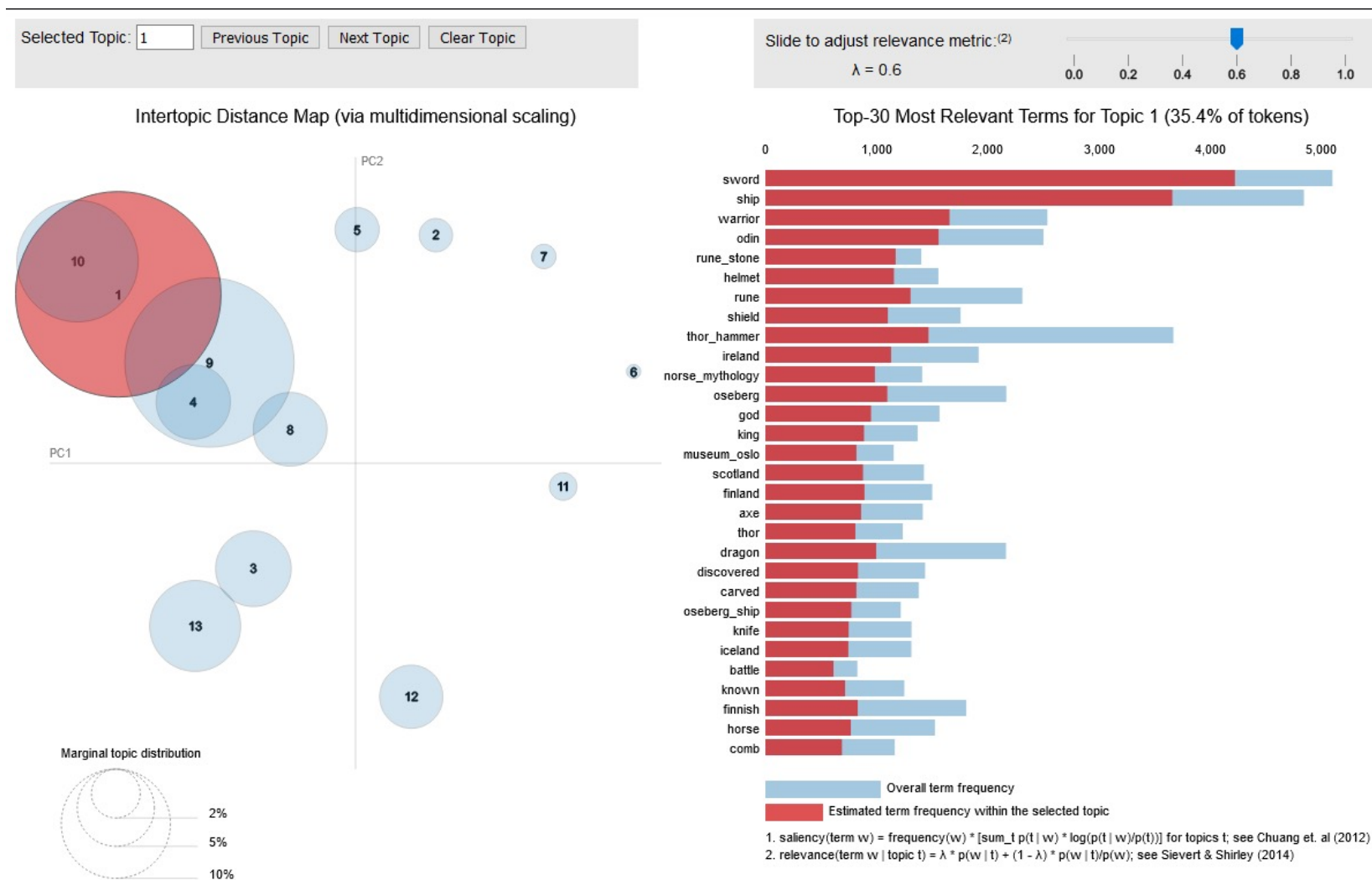


Add a comment



Bodil Axelsson saved to **viking jewelry**

Topic i Pinterest



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