



A SHORT INTRODUCTION TO



SPRÅKBANKEN **TEXT**



Overview

BLOGGMIX 2005 (stödjer ej utökad kontext)

lemien, Kungliga biblioteket och **Språkbanken** i Göteborg.

BLOGGMIX 2006 (stödjer ej utökad kontext)

För svenska har vi **språkbanken** – en stor gratis tjänst som måste vara gudarnas största gåva till r
kräftas empiriskt så jag kollade i **Språkbanken** och om man ska tro på deras data så var det faktiskt så att forme
Här är en länk till **Språkbanken** .

BLOGGMIX 2007 (stödjer ej utökad kontext)

Detta bekräftas om man går till **Språkbanken** : utan dess like finns överhuvudtaget inte vare sig i de tidigare tid

BLOGGMIX 2010 (stödjer ej utökad kontext)

De röda symbolerna är **Språkbanken** , de blå träffar i Dagens Nyheter på Mediearkivet.
skorpusen Press65, som finns på **Språkbanken** , så finns där 46 förekomster av äta men ingen alls av kaka, och ta
atar och talar i Mediearkivet och **Språkbanken** , resultaten syns i diagrammet nedan.

BLOGGMIX 2011 (stödjer ej utökad kontext)

har därför skapat en gemensam **språkbank** med frivilliga medlemmar och förtroendevalda.

- What is Språkbanken?
- Our resources
- Our tools
- Some example applications
- To do together



WHAT IS SPRÅKBANKEN?



What is Språkbanken?

- A bank generates profit
- Språkbanken generates knowledge



Who needs Språkbanken?

Scientists, organisations, companies who are working with human communication, e.g.

- computer scientists
- philologists
- gender scientists
- historians
- health scientists
- cultural scientists
- linguists
- literary scholars
- media scientists
- pedagogues
- psychologists
- social anthropologists
- language technologists
- political scientists
- etc.

The different flavours of Språkbanken

- **Nationella språkbanken** is a national research infrastructure that is financed by Vetenskapsrådet and 10 universities and governmental agencies
- It consists of three divisions: **Språkbanken Sam**, **Språkbanken Tal** and **Språkbanken Text**
- This presentation is about **Språkbanken Text**, which is working with written language data



GÖTEBORGS
UNIVERSITET

SBTEXT

OUR RESOURCES

600+ corpora

57 lexica



What is a corpus?

- A corpus is a large digital collection of texts that can be used for, e.g., searching or training AI models
- To improve the usefulness, corpora are usually annotated, e.g.
 - morphological information (*boken* is a noun in singular definite form, and is the same word as *bok*, *böcker* and *böckerna*)
 - syntactic (*bok* is the object in *jag läser en bok*)
 - semantic information (what is the meaning of the word *bok* in this sentence – a written text, or a tree?)
 - metadata (what kind of text is it? who wrote the text? when?)
- The annotation is most often made automatically by computer algorithms (and therefore contains errors)

Some of Språkbanken's corpora

- Novels
- Newspapers
- Journals
- Sociala media
 - blogs
 - discussion forums
 - Twitter
- Governmental texts
- Wikipedia
- Finland Swedish
- Historical texts
 - Old Swedish
 - Modern Swedish
- Parallel texts
 - the same texts in 25+ languages
- Gigaword
 - a large balanced corpus of various genres from 1950 onwards

Some of Språkbanken's lexica

- SALDO: a semantic and morphological dictionary
- Hellquist's Swedish etymological dictionary
- Blissymbolics dictionary
- Schlyter's and Söderwall's dictionaries of medieval Swedish
- Sentiment lexicon
- Dictionary of loan words
- Swesaurus: a Swedish WordNet, a Swedish concept lexicon
- SweFN++: a Swedish FrameNet, a database over semantic frames

Our lexica are created to be used for automatic text analysis



GÖTEBORGS
UNIVERSITET

SBTEXT

OUR TOOLS

**Sparv**
Språkbankens annoteringsverktyg



Analyspråk: svenska 

Ladda exempel:  Drama  Ätta sidor  Talbanken  Läsbart  Ikea  Exempelkorpus

☒ Ren text ☐ XML

1 Vad har vi för olika analyser?



☒ Lexikalanalys ☐ Sammansättningsanalys ☒ Dependensanalys ☐ Attitydanalys ☐ Namntaggare ☐ Läsbarhetsvärden

Our most important tools



Korp: a search engine that provides access to about 15 billion words of Språkbanken's corpora



Karp: a tool for working with Språkbanken's lexica



Sparv: an tool that annotates texts with morphological, syntactic and semantic information



Strix: a document-oriented search engine that can take a document's semantic content in account (in contrast to Korp which is word-oriented)



Lärka: a platform for second language learning and second language research



Korp

- Korp is a search engine
- intended for investigating linguistic phenomena

<https://spraakbanken.gu.se/korp>



Korp – word search (KWIC = keywords in context)

- gives information about the frequency of the word, showing its occurrences in context, here's an example for the word *kyssa*

KWIC Statistik Ordbild

Antal träffar: **12 843**

« < 1 2 3 4 5 6 7 8 9 10 11 »

Gå till sida av 514 [Visa kontext](#)

Frequency

Name of corpus

Examples

Hennes genombrott kom 1957 med en låt om "hur den svenska flickan **kysser**".

ÅBO UNDERRÄTTELSE 2012

ÅBO UNDERRÄTTELSE 2013

Redan vid .du kan **kyssa** bruden.

Jag brukar säga att jag tar inget ansvar om man **kysser**

ASTRA 1960–1979 (stödjer ej utökad konte

Han lyfte den ur ett förgyllt gotiskt stativ, korsade sig och **kysste** relikens vördnadsfullt.

Vill ni också **kyssa** den signora?

Fortida furstar har bugat för miraklet, **kysst** ampullerna och skänkt f

ASTRA NOVA 2008–2010

tomater och luktärter, spelar vi rock och jazz, blir vi fortfarande stående mitt på en gata och ett nöjesfält för att **kyssas** ?

Info about corpus

Korpus

Åbo Underrättelser 2012

Info about word

Textatt

datum:

Ordattribut

ordklass: verb

sammansatta lemmgram:

- kyss (substantiv) + se (verb)

Visa fler (1)

dependensrelation: Prepositions

komplement

förlädd:

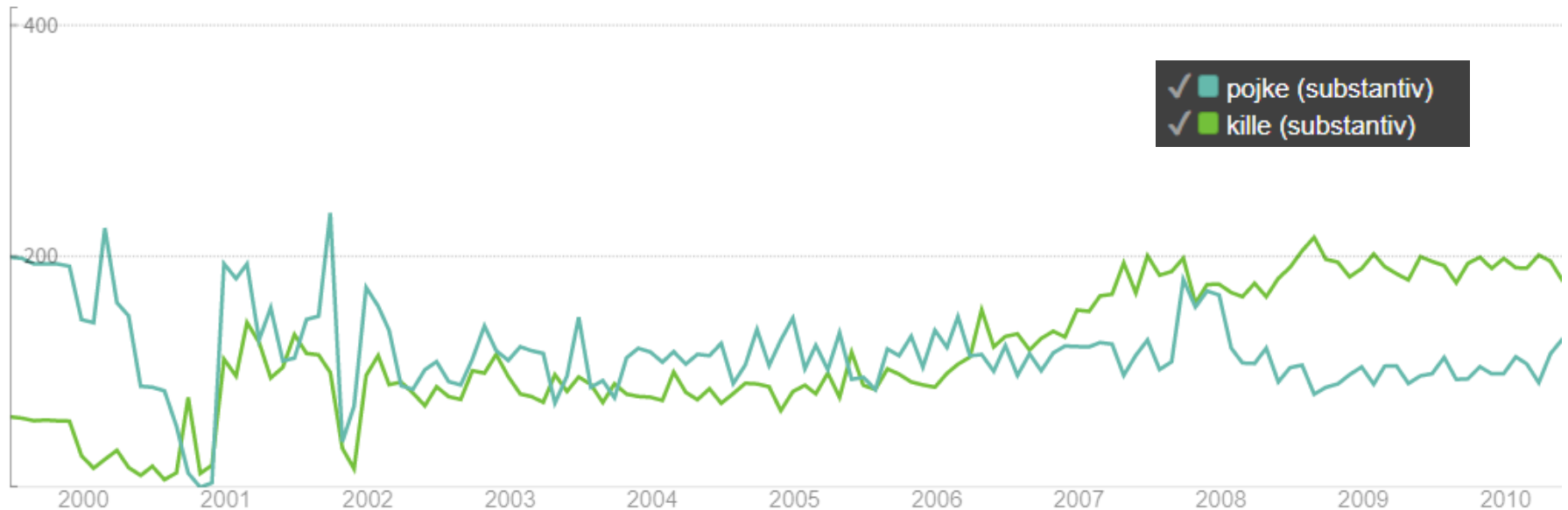
WRONG! Stupid computer!



Korp – other possibilities



Korp – trend diagram: language change





Korp – comparisons (based on probabilities)

Which parties focus on *education* and which focus on *health care* in their programs and manifestos?

Utmärkande för *utbildning*

Centerpartiet	215
Vänsterpartiet	88
Piratpartiet	6
Moderata samlingspartiet	121
Folkpartiet liberalerna	181
Sveriges liberala parti	4
Lantmanna- och borgarepartiet	3

Utmärkande för *sjukvård*

Kristdemokraterna	62
Miljöpartiet de gröna	48
Ny demokrati	6
De rödgröna	6
Sverigedemokraterna	1
Alliansen	21
Sveriges socialdemokratiska arbetareparti	45



Korp – word pictures


Who kisses whom/what in Sweden, and in which way?


Subjekt	kyssa	Objekt	Adverbial
1. pojke	41 🗲	1. tjej	181 🗲
2. främling	25 🗲	2. främling	97 🗲
3. kille	42 🗲	3. fot	136 🗲
4. man	46 🗲	4. groda	93 🗲
5. kvinna	39 🗲	5. flicka	103 🗲
6. hand	18 🗲	6. hand	107 🗲
7. kille ²	23 🗲	7. kille	97 🗲
8. tjej	26 🗲	8. klubbmärke	45 🗲
9. oskuld	10 🗲	9. kille ²	73 🗲
10. gaypar	8 🗲	10. kompis	71 🗲
11. gång ²	22 🗲	11. kind	51 🗲
12. hemlandshustru	6 🗲	12. tomte	51 🗲
13. sol	20 🗲	13. tomt	51 🗲
14. sol ²	20 🗲	14. läpp	57 🗲
		1. på kind	86 🗲
		2. på mun	72 🗲
		3. i regn	52 🗲
		4. i nacke	49 🗲
		5. senast	63 🗲
		6. på panna	33 🗲
		7. på hand	33 🗲
		8. första gången	38 🗲
		9. gång ²	72 🗲
		10. gång	55 🗲
		11. gång ³	55 🗲
		12. sen	68 🗲
		13. i arsle	18 🗲
		14. på hals	17 🗲



Karp

- Karp is a tool for working with Språkbanken's lexica

 Svenska | English Valj en resurs nedan eller sök i Karps standardurval.

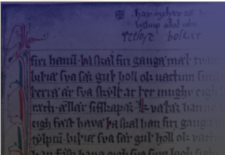


Dalin

Dalins ordbok - Ordbok över 1800-talsspråket

SALDO

Semantiskt och morfologiskt lexikon för språkteknologi.




Fornsvenska

Schlyter och Söderwall


Konstruktikon

Svenskt konstruktikon är en fritt tillgänglig konstruktionsdatabas, dvs. en samling beskrivningar av svenska konstruktionsmönster. Dessa konstruktioner kan vara allt från väldigt generella strukturer till högst specifika uttryckssätt.



SweFN

Svenskt frasnät






Svenskt kvinnobiografiskt lexikon

<https://spraakbanken.gu.se/karp>



Karp – the Saldo lexicon

- How does the semantic neighbourhood look for the word *hus*?

Hitta ingångar där  primär deskriptor  är lika med  hus..1

The search query


SALDO ▼ 77 TRÄFFAR (VISAR 25)












	BETYDELSE	LEMGRAM	ORDKLASS	PRIMÄR	SEKUNDÄR	BARN (PRIMÄRA)	BARN (SEKUNDÄRA)
⚙	balkong	balkong (substantiv) 🐦 204656	substantiv	hus		piskbalkong	balkongdörr balkongdörr ² balkonglåda balkongräcke
⚙	bankhus	bankhus (substantiv) 🐦 3153	substantiv	hus	bank		
⚙	bod ²	bod (substantiv) 🐦 116910	substantiv	hus			bodlänga
⚙	bollhus	bollhus (substantiv) 🐦 1347	substantiv	hus	bollspel		



Karp – etymological dictionary

- From which languages have Swedish borrowed the most?

Hitta ingångar där inte etymon-språk är lika med
och där också etymon-språk finns 

fra	200	
deu	114	
lat	88	
ita	48	
eng	40	
dan	33	
ell	33	
lty	23	
nor	21	
spa	17	
medellågtyska	13	

Search query: "find all entries that have specified which language the word comes from, and that language is not Old Swedish"



Karp – diachronic pivot

- I want to search for the word *kvinna* in older texts, but I don't really know how it was spelled during the period I'm interested in

kvinna (substantiv)

🐦 6395274

Dalin/1800-tal

Exakta träffar:

qvinna (substantiv)

🐦 55743

qvinno (substantiv)

🐦 121

Breda träffar:

bondqvinna (substantiv)

🐦 80

fästeqvinna (substantiv)

🐦 167

hjelteqvinna (substantiv)

🐦 3

dagsverksqvinna (substantiv)

🐦 0

grannqvinna (substantiv)

🐦 156

Fornsvenska

Exakta träffar:

grankuna (substantiv)

🐦 0

granqvinna (substantiv)

🐦 0



Karp – Blissymbolics

- is a semantic graphical language where the where the concepts are represented by ideographic images
- Which Bliss-symbols correspond to *farfar* and *mormor*?

Hitta ingångar där  vad som helst  är lika med  farfar eller mormor

	BCI-AV#	GLOSS	SALDO	WORDNET
  	14626	grandfather granddad grandpa	farfar morf	
<hr/>   	14629	grandmother grandma granny	farmor mormor	



Korp ♥ Karp – automatic text analysis

Let's search for the word *grym* in Korp:

Different meanings!

GP 2013	
- Nja, det vet jag inte, men Sean Banan är grym	som i verkligheten.
serna om tortyr, övergrepp, skräck blir just så obegripligt grymma	Vi tappade bollen – och de hade sådan grym kvalitet i det som de gjorde.
Mika har gjort grymt	mycket poäng och Kari är en av allsvenskans bästa vänsterbackar, säge
mot Bålberget, ger sig Therése Söderlind i kast med detta grymma	och svärbegripliga kapitel i Sveriges historia.
Häcken är en grym	klubb, en skön klubb.
ånad ser arrangörerna till att ge besökarna tung bas och grym	electro.
leagles till laboratorier i Storbritannien för att användas i grymma	experiment när avelsanläggningen i Örkelljunga stänger ner, skriver ho
Vi tappade bollen – och de hade sådan grym	kvalitet i det som de gjorde.
GÖTEBORG: Grymt	gott!
Och då är hon grym	.
Det hade varit grymt	att vinna två raka guld, samtidigt måste fokus ligga på varje enskild ma
Vecka efter vecka presenterar de den ena grymma	bokningen efter den andra.
programledaren Timo Räisänen konstaterade: "Fatta vad grymma	vi är på musik i det här landet! "
Den sista löpningen är grym	och berör mig starkt.
leagles till laboratorier i Storbritannien för att användas i grymma	experiment när avelsanläggningen i Örkelljunga stänger ner, skriver ho
Det är grymt	svårt att få en hyfsad storlek på dem.
rna i England kommer att utsättas för experiment som är grymmare	än de i Sverige.
var tills ridån går upp för kvällens föreställning, komedin Grymt	galet på Skandiateatern i Norrköping.
Under EM 2006. vilken grym	publik det var.



Korp ♥ Karp – the different meanings of *grym*

	GP 2013	grym	.
– Nja, det vet jag inte, men Sean Banan är	grymma		som i verkligheten.
ttelserna om tortyr, övergrepp, skräck blir just så obegripligt	grym		kvalitet i det som de gjorde.
Vi tappade bollen – och de hade sådan	grymt		mycket poäng och Kari är en av allsvenskans bästa vänsterbackar, säger han.
Mika har gjort	grymma		och svårbegripliga kapitel i Sveriges historia.
en mot Bålberget, ger sig Therése Söderlind i kast med detta	grym		klubb, en skön klubb.
Häcken är en	grym		electro.
je månad ser arrangörerna till att ge besökarna tung bas och	grymma		experiment när avelsanläggningen i Örkelljunga stänger ner, skriver hon.
ls beagles till laboratorier i Storbritannien för att användas i	grym		kvalitet i det som de gjorde.
Vi tappade bollen – och de hade sådan	Grymt		gott!
GÖTEBORG:	grym		.
Och då är hon	grymt		att vinna två raka guld, samtidigt måste fokus ligga på varje enskild match.
Det hade varit	grymma		bokningen efter den andra.
Vecka efter vecka presenterar de den ena	grymma		vi är på musik i det här landet! ”
m programledaren Timo Räisänen konstaterade: ”Fatta vad	grym		och berör mig starkt.
Den sista löpningen är	grymma		experiment när avelsanläggningen i Örkelljunga stänger ner, skriver hon.
ls beagles till laboratorier i Storbritannien för att användas i	grymt		svårt att få en hyfsad storlek på dem.
Det är	grymmare		än de i Sverige.
ndarna i England kommer att utsättas för experiment som är	Grymt		galet på Skandiateatern i Norrköping.
gt kvar tills ridån går upp för kvällens föreställning, komedin	grym		publik det var.
Under EM 2006, vilken	Grymt		! ” säger han.
”	grymt		effektiv att rulla bollar.
Man blir	grym		klubb, en skön klubb.
Häcken är en	grymt		bra, säger lagets coach Johanna Ericsson.
i fick en pangstart på matchen, 73 poäng i slutspelsmatch är	Grymt		Galet.
Tillsammans spelar de systrar i komedin			

Korpus

GP 2013

Textattribut

artikelförfattare: Torbjörn Skarhe

artikelavdelning: Kultur Nöje

datum: 2013-02-22

Ordattribut

ordklass: adjektiv

sammansatta lemgram: [tom]

dependensrelation: Subjektspred
(subjektiv predikatsfyllnad)

förled: [tom]

efterled: [tom]

betydelse:

- **grym²**
Visa fler (1)

msd: JJ.POS.UTR.SIN.IND.NOM ⓘ

sammansatta ordformer: [tom]

grundform:

grym

lemgram:

grym (adjektiv)

Visa dependensträd



Korp ♥ Karp – the different meanings of *grym*

GP 2013	
– Nja, det vet jag inte, men Sean Banan är	grym .
tttelserna om tortyr, övergrepp, skräck blir just så obegripligt	grymma – som i verkligheten.
Vi tappade bollen – och de hade sådan	grym kvalitet i det som de gjorde.
Mika har gjort	grymt mycket poäng och Kari är en av allsvenskans bästa vänsterbackar, säger han.
gen mot Bålberget, ger sig Therése Söderlind i kast med detta	grymma och svärbegripliga kapitel i Sveriges historia.
Häcken är en	grym klubb, en skön klubb.
je månad ser arrangörerna till att ge besökarna tung bas och	grym electro.
als beagles till laboratorier i Storbritannien för att användas i	grymma experiment när avelsanläggningen i Örkelljunga stänger ner, skriver hon.
Vi tappade bollen – och de hade sådan	grym kvalitet i det som de gjorde.
GÖTEBORG:	Grymt gott!
Och då är hon	grym .
Det hade varit	grymt att vinna två raka guld, samtidigt måste fokus ligga på varje enskild match.
Vecka efter vecka presenterar de den ena	grymma bokningen efter den andra.
om programledaren Timo Räisänen konstaterade: " Fatta vad	grymma vi är på musik i det här landet! "
Den sista löpningen är	grym och berör mig starkt.
als beagles till laboratorier i Storbritannien för att användas i	grymma experiment när avelsanläggningen i Örkelljunga stänger ner, skriver hon.
Det är	grymt svårt att få en hyfsad storlek på dem.
ndarna i England kommer att utsättas för experiment som är	grymmare än de i Sverige.
gt kvar tills ridån går upp för kvällens föreställning, komedin	Grymt galet på Skandiateatern i Norrköping.
Under EM 2006, vilken	grym publik det var.
"	Grymt ! " säger han.
Man blir	grymt effektiv att rulla bollar.
Häcken är en	grym klubb, en skön klubb.
vi fick en pangstart på matchen, 73 poäng i slutspelsmatch är	grymt bra, säger lagets coach Johanna Ericsson.
Tillsammans spelar de systrar i komedin	Grymt Galet.

Korpus

GP 2013

Textattribut

artikelförfattare: Henrik Strömbe

artikelavdelning: Kultur Nöje

datum: 2013-02-01

Ordattribut

ordklass: adjektiv

sammansatta lemmagram: [tom]

dependensrelation: Subjektsprec

(subjektiv predikatsfyllnad)

förled: [tom]

efterled: [tom]

betydelse:

- grym
- Visa fler (1)

msd: JJ.POS.UTR+NEU.PLU.IND+DEF.NO

sammansatta ordformer: [tom]

grundform:

grym

lemgram:

grym (adjektiv)



Korp ♥ Karp – the meanings come from Saldo

GP 2013

– Nja, det vet jag inte, men Sean Banan är grym .
Berättelserna om tortyr, övergrepp, skräck blir just så obegripligt grymma – som i verkligheten.
Vi tappade bollen – och de hade sådan grym kvalitet i det som de gjorde.
Mika har gjort grymt mycket poäng och Kari är en av allsvenskans bästa vänsterbackar, säger han.
n, Vägen mot Bålberget, ger sig Therése Söderlind i kast med detta grymma och svärbegripliga kapitel i Sveriges historia.

SALDO ▾ 2 TRÄFFAR (VISAR 2)

BETYDELSE	LEMGRAM	ORDKLASS	PRIMÄR	SEKUNDÄR	BARN (PRIMÄRA)	BARN (SEKUNDÄRA)
grym	grym (adjektiv) 🐦 1087340	adjektiv	hård		bestialisk grymhet kärlekslös	despot
grym ²	grym (adjektiv) 🐦 1087340	adjektiv	bra		grymhet ² lexikograf lexikolog	



Sparv

- Sparv is a tool that can **annotate** your texts (both short and long), i.e., provide them with...
 - lexical analysis
 - syntactic analysis
 - semantic analysis
 - sentiment analysis
 - readability analys
 - etc.

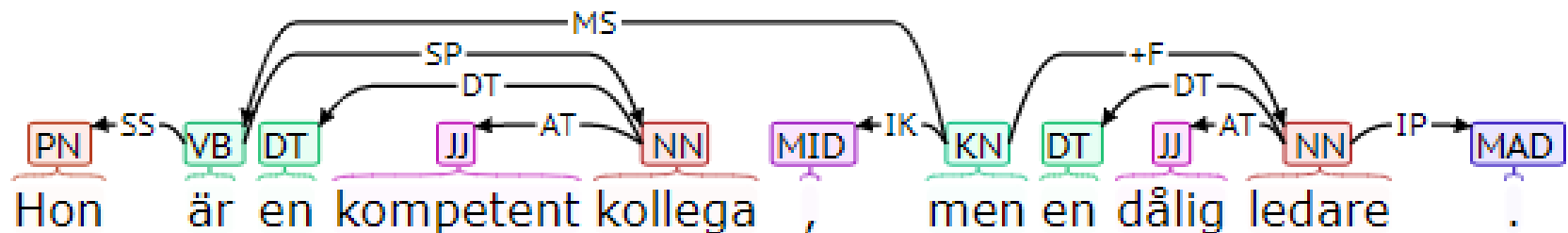
<https://spraakbanken.gu.se/sparv/>



Sparv – syntactic analysis

Hon är en kompetent kollega, men en dålig ledare.

(Eng: *She is a competent colleague, but a bad manager*)





Sparv – semantic analysis

Hon är en kompetent kollega, men en dålig ledare.

- The word *ledare* can have at least three meanings:

Meaning	Probability in this context (according to Sparv)
manager	0,65
electrical conductor	0,21
editorial	0,14



Sparv – sentiment analysis

Hon är en kompetent kollega, men en dålig ledare.

token	sentimentclass
Hon	
är	neutral
en	
kompetent	positive
kollega	neutral
men	
en	
dålig	negative
ledare	neutral




Strix

- a search engine that...
 - doesn't focus on single **words** (as Korp does), but on whole documents,
 - takes the semantic content of the text into account,
 - can be said to be something in between Korp and Google,
 - is still under development.



Strix – searching for a document

search filter



STRIX

Samling

- Svenska Wikipedia**
- RD - Motion (523 892)
- RD - Betänkande (67 523)
- Välj fler från lista (23 alternativ)

Blingbring

- omfattning (1 265 600)
- prosa (1 216 487)
- inträde (1 181 782)
- Välj fler från lista (994 alternativ)

I följd

Search query

Hittade 248 dokument.

1 2 3 4 5 6 7 8 9 10

Scandal (TV-serie)
Scandal (TV-serie) Scandal är en amerikansk tv-serie skapad av Shonda Rhimes. Serien hade svensk premiär på Kanal 5 den 3 oktober 2012. Scandal utspelar sig i Washington D.C. och kretsar runt Olivia Pope (spelad av Kerry Washington) och arbetet på krisrådgivningsbyrån Pope & Associates. Tittaren får också ...
SVENSKA WIKIPEDIA

Serie A-skandalen 2006
augusti Reggina drogs in i **skandalen** 7 augusti, då det behålla sina poängstraff. Bakgrund **Skandalen** uppdagades av en händelse då
SVENSKA WIKIPEDIA

World Cup i bandy 1999



Strix – searching within a document

Svenska Wikipedia: Serie A-skandalen 2006 ✕

Färgmarkera
ordattribut ▼
inget val ▼

1
2 **Serie A-skandalen 2006**
3 Serie A-skandalen 2006 (italienska: calciopoli eller Moggiopoli) är en härva av brott och andra oegentligheter som upptäcktes av italiensk polis under maj 2006 i den italienska fotbollsligan Serie A. De italienska storklubbarna Juventus, Lazio, Fiorentina och AC Milan anklagades för att ha påverkat matchresultaten genom att välja ut lojala domare och linjemän som dömde deras matcher. Flera högt uppsatta personer inom Italiens fotbollsforbund Federazione Italiana Giuoco Calcio anklagades för att ha låtit domarmanipulationerna fortgå.
4 Under april 2007 rapporterades det att den italienska polisen valt att återuppta undersökningarna om uppgjorda matcher säsongen 2004/2005. Denna gång undersöker man resultaten i 15 nya matcher. De klubbar som straffats 2006 riskerar dock inte några nya straff, däremot riktas anklagelser mot Messinas sportdirektör, samt flera domare och assistentdomare som inte tidigare har förekommit i undersökningen.
5 14 juli
6 Enligt den dom som meddelades den 14 juli blev Juventus fråntagna sina två senaste mästerskapstitlar och även nedflyttade till Serie B där de startade säsongen 2006/2007 med 30 minuspoäng. Fiorentina och Lazio startade säsongen med 12 respektive 7 minuspoäng. AC Milan slapp nedflyttning men straffades med 44 minuspoäng för säsongen 2005/2006 och startade dessutom säsongen 2006/2007 i Serie A med 15 minuspoäng. Flera personer inom Federazione Italiana Giuoco Calcio samt klubbledare och domare dömdes till böter och blev

Sökträffar
(2)

Textattribut

SWEEP


Activity_resume
Fining
Sentencing

BLINGBRING

domare
lagkarl
lagskipning

LÄSBARHETSINDEX
47.87

OVIX
57.29



Sparv figures out the main themes of the text



Strix – searching within a document

Show all words that denote organisations

Färgmarkera **ordattribut** **swefn** **Organization**  

1

2 **Serie** A-skandalen 2006

3 Serie A-skandalen 2006 (italienska: calciopoli eller Moggiopoli) är en härva av brott och andra oegentligheter som upptäcktes av italiensk polis under maj 2006 i den italienska fotbollsligan Serie A. De italienska **storklubbarna** Juventus, Lazio, Fiorentina och AC Milan anklagades för att ha påverkat matchresultaten genom att välja ut lojala domare och linjemän som dömde deras matcher. Flera högt uppsatta personer inom Italiens **fotbollsförbund** Federazione Italiana Giuoco Calcio anklagades för att ha låtit domarmanipulationerna fortgå.

4 Under april 2007 rapporterades det att den italienska polisen valt att återuppta undersökningarna om uppgjorda matcher säsongen 2004/2005. Denna gång undersöker man resultaten i 15 nya matcher. De **klubbar** som straffats 2006 riskerar dock inte några nya straff, däremot riktas anklagelser mot Messinas sportdirektör, samt flera domare och assistentdomare som inte tidigare har förekommit i undersökningen.

5 14 juli

6 Enligt den dom som meddelades den 14 juli blev Juventus fråntagna sina två senaste mästerskapstitlar och även nedflyttade till Serie B där de startade säsongen 2006/2007 med 30 minuspoäng. Fiorentina och Lazio startade säsongen med 12 respektive 7 minuspoäng. AC Milan slapp nedflyttning men straffades med 44 minuspoäng för säsongen 2005/2006 och



Strix – searching within a document

Show all words that denote the start of a process

Färgmarkera **ordattribut** swefn Process_start

mästerskapstitlar och även nedflyttade till Serie B där de **startade** säsongen 2006/2007 med 30 minuspoäng. Fiorentina och Lazio **startade** säsongen med 12 respektive 7 minuspoäng. AC Milan slapp nedflyttning men straffades med 44 minuspoäng för säsongen 2005/2006 och **startade** dessutom säsongen 2006/2007 i Serie A med 15 minuspoäng. Flera personer inom Federazione Italiana Giuoco Calcio samt klubbledare och domare dömdes till böter och blev avstängda från fotbollen.

7 Domslutet innebar att Inter och Roma var direktkvalificerade till Champions League säsongen 2006/2007. Palermo och Chievo fick kvala till turneringen. Toscana-klubbarna Empoli och Livorno är kvalificerade till Uefacupen säsongen 2006/2007. Tidigare nedflyttade trion Treviso, Messina och Lecce undgick nedflyttning till Serie B.

8 25-26 juli

9 25 juli föll domen efter överklaganden från de dömda parterna.

10 Milan får sitt poängstraff reducerat från 15 till 8 minuspoäng kommande säsong. Dessutom fick laget ett mindre avdrag för föregående säsong, 30 istället för 44 minuspoäng, och därmed fick man delta i Champions League. (Där de vann turneringen 2006/2007)

11 Både Lazio och Fiorentina fick **börja** säsongen i Serie A med poängavdrag, men slapp alltså nedflyttning. Fiorentina fick **börja** på minus 19 poäng och Lazio på minus 11 poäng. Juventus fick **börja** Serie B med 17 poängs avdrag istället för 30. Juventus fick som tidigare beslutats lämna ifrån sig mästartiteln från säsongen 2004/2005. Den 26 juli meddelade en



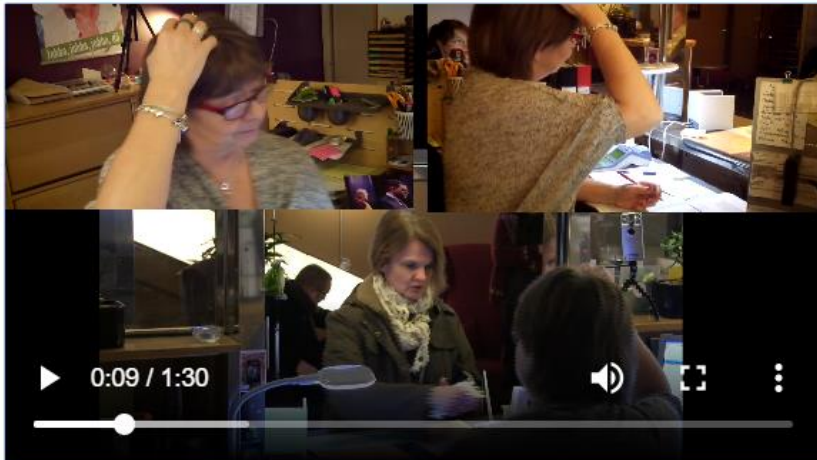
Strix – multimodal corpora

- For some data (e.g., recordings of conversations), one can get access to audio and video sources

Colorize word embeddings

1 (4.0)
2 #så:dä:r#
3 hej
4 hej (0.5)
5 skulle hä
6 stället (
7 [ska vi s
8 [aa .aa]
9 (3.5)
10 en två so
11 (4.9)
12 eh du har
13 [fem] i [stället] ja
14 [mm]
15 (1.6)
16 aa de e vi samma bord så ja bara plockar bort [en ultav dom

i



stället (0.6) där var nu[mret där]



Lärka

- **Lärka** (LÄR språket via KorpusAnalys) is a learning platform and research tool. Lärka is able to...
 - generate exercises for language learners
 - find relevant examples
 - assist in assessing...
 - student essays
 - text difficulty
 - etc.

<https://spraakbanken.gu.se/larka/>



Lärka – generate exercises...

Vocabulary Multiple Choice

C1 ▼

Set level!

Click to generate!

- | | |
|---|---|
| 5 | Hjulen , förstår ni , fru Dina , de får aldrig _____ fast ! |
| 4 | Krämig risotto , gärna smaksatt med _____ , passar fint till . |
| 3 | Endast Gud hade hon att _____ hos . |
| 2 | Slå blandningen i pajen och _____ allt 30 minuter i 225 graders ugn . |
| 1 | Det är olagligt att behålla _____ som är under 45 cm lång . |

besluta

pest

klaga

grädda

öring



...and correct the answers automatically!



Lärka – assess texts

Skriv eller klistra in texten i rutan nedan.

Vad gör man i sådana situationer i Sverige? Vad är traditionen när det gäller att en ny familj bildas och barnen är fram och tillbaka varannan vecka till bonusmamma och bonuspappa? Bestämmer man själv hur relationerna ska funka? Vad är det som är acceptabelt? Jag har som sagt en sådan bred bakgrund med kulturer att jag aldrig sett så mycket konflikt och bitterhet som i Sverige och det förvånar mig eftersom jag hade helt en annan bild av svensk kultur och relationerna. |

Vilken typ av text vill du bedöma? ?

Uppsats

Läsförståelse

Visa alla ord för följande CEFR nivåerna ?

☒ A1

☒ A2

☒ B1

☒ B2

☒ C1

Ytterligare funktioner ?

☐ Markera möjliga fel

☐ Använd stavningskontroll



Lärka – assess texts

Vad gör man | sådana situationer | Sverige ? Vad är traditionen när det gäller att en ny familj bildas och barnen är fram och tillbaka varannan vecka till bonusmamma och bonuspappa ? Bestämmer man själv hur relationerna ska funka ? Vad är det som är acceptabelt ? Jag har som sagt en sådan bred bakgrund med kulturer att jag aldrig sett så mycket konflikt och bitterhet som i Sverige och det förvånar mig eftersom jag hade helt en annan bild av svensk kultur och relationerna .

Statistisk information

Automatisk bedömning av CEFR: C1

Detaljerad bedömning

Antal meningar	5
Antal ord	87
Antal icke-lemmatiserade ord	0
Genomsnittlig meningslängd	17.4
Genomsnittlig ordlängd	4.38
Genomsnittlig dependenslängd	2.29
LIX värde	38 (lättläst)
Nominalkvot	0.68
Pronomen-substantiv kvot	0.64

Vilken typ av text vill du bedöma? ⓘ

Uppsats

Läsförståelse

Visa alla ord för följande CEFR nivåerna ⓘ

- ☒ A1
- ☒ A2
- ☒ B1
- ☒ B2
- ☒ C1



Lärka – results

Vad gör man | sådana situationer | Sverige ? Vad är traditionen när det gäller att en ny familj bilds och barnen är fram och tillbaka varannan vecka till bonusmamma och bonuspappa ? Bestämmer man själv hur relationerna ska funka ? Vad är det som är acceptabelt ? Jag har som sagt en sådan bred bakgrund med kulturer att jag aldrig sett så mycket konflikt och bitterhet som | Sverige och det förvånar mig eftersom jag hade helt en annan bild av svensk kultur och relationerna .

CEFR Receptive Distribution

A1: 54.88 %

A2: 8.54 %

B1: 9.76 %

B2: 7.32 %

C1: 8.54 %

?: 10.98 %

CEFR Productive Distribution

A1: 53.66 %

A2: 10.98 %

B1: 12.20 %

B2: 1.22 %

C1: 2.44 %

?: 19.51 %



Installing the tools for yourself

- Most of our tools are open source
- Korp, Karp and Sparv can be used through an API
- Sparv can be installed as a command-line pipeline
- Korp can be adapted to other corpora and other languages, and is currently being used in Norway, Finland, Denmark, Iceland, Estonia and Italy



The need for more data – open data

- The usefulness of our tools depend on good annotations
- Manual annotation is expensive
 - so only few resources are manually annotated
 - most of our data is automatically annotated
 - to be able to do that we need good NLP models
- The quality of the data impact the quality of the models
- We need all kinds of text data for training different models
 - newspaper texts
 - social media, discussion forums
 - official publications, reports, informational web pages, etc.
 - blog posts, personal web pages, etc.



GÖTEBORGS
UNIVERSITET

SBTEXT

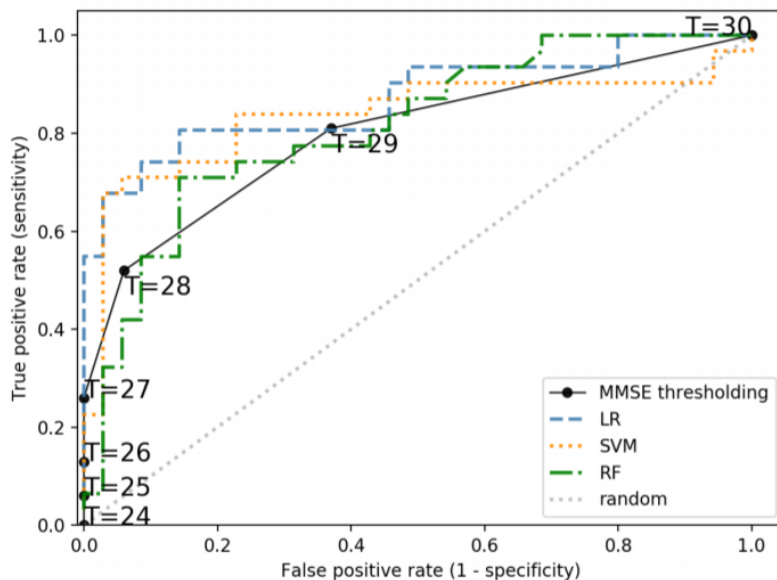
SOME EXAMPLES OF OUR RESEARCH



Photo: Elena Volodina

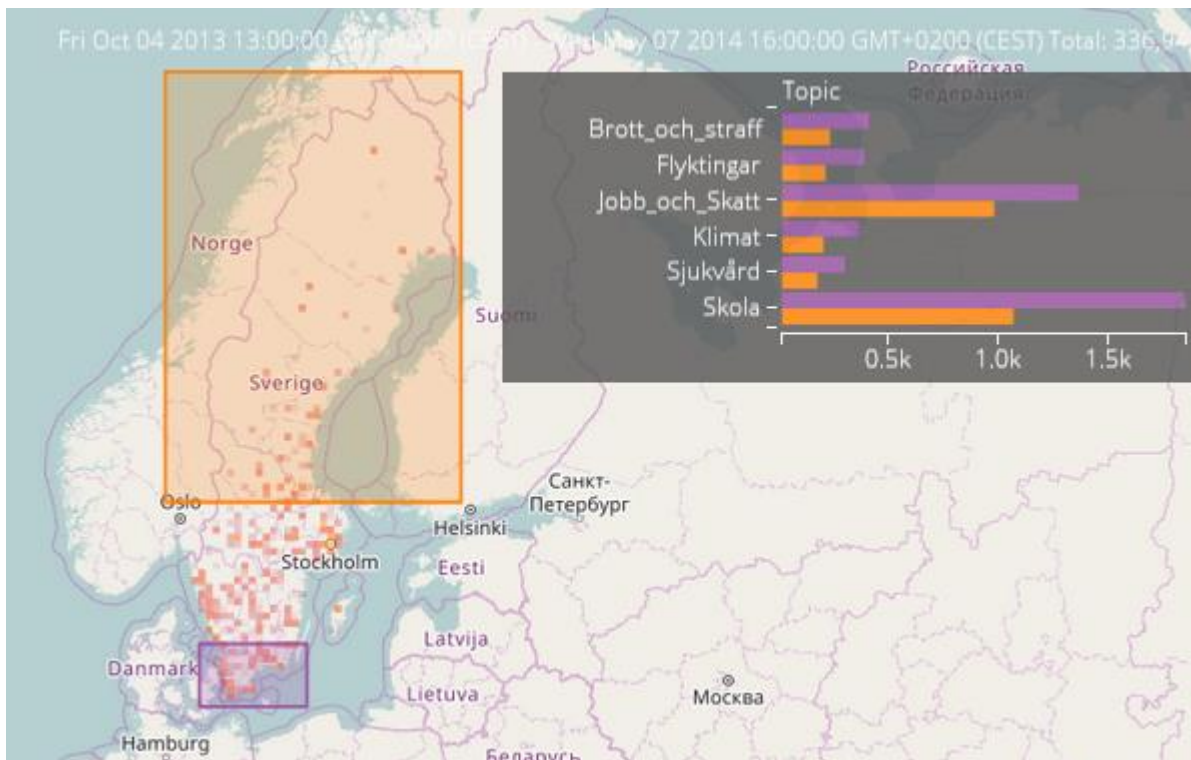
Language technology supports medical diagnostics

- A common test that is used for diagnosing mild cognitive impairment (MCI) has the accuracy (“AUC score”) **68%**
- If we enhance the test with automatic analysis of the patient’s language, the accuracy is improved to **87%**



Fraser, Lundholm Fors, Eckerström, Themistocleous, and Kokkinakis (2018). *Improving the sensitivity and specificity of MCI screening with linguistic information*

What are Swedes discussing on Twitter right now?



Orange vs purple:
hur mycket is the
theme discussed in
northern vs southern
Sweden.

Simplifying the reading of historical texts

I.

Crister ær först i laghum uarum. tha ær kristna uaar.
 Uarther barn boret till kirkyu. **och** bethes kristnu. tha skal
 father **oc** mother. guthfathur **och** gutmothor til fa **oc** salt.
 tha skall barn til kirkyu bæra **oc** a præst kalla. han skal a
 kirkyubole boa. barn skal brimsigna for utan kirkyu dör.
 sithan skal font uia. **oc** præster skal barn döpa **oc**
 guthfater **oc** guthmother a halda **oc** til namn sigiæ.
 præster skall biuda huru længe fater **oc** mothor skal
 uarthueta. Hænder soot a uægh **oc** ma ey til kirkyu koma.
 tha skal döpa i uatne. i namn fathurs **och** sons. **ok** thæs
 hælgha anda. þa skal thæt i kirkyu garth grafua **oc** arff
 taka. Komber liuande barn fram tha skal thæt brimsigna **oc**
 kristna. **oc** optare ey döpa. Hænder barn syukdomber **oc** ær
 ey kona een inne. þa skal hun döpa thæt **oc** thy namn
 gifua tho skal thæt thy huaro i kirkyu garth grafua **oc** arff
 taka. Uærther barn brimsignat **oc** ey döpt. tha skal þæt ey i

ok, 14 träffar (ok/uk/iak/oker)

- ok
 - Schlyter
Ok (hoc, Sk.*), 1) conj. oc
 - ...
 - Schlyter nn **Ok**, n. se Uk.
 - Söderwall kn
ok (oc . ock . och
 - ...
 - Söderwall nn **ok**, n. se uk.
 - Söderwall supp ab, kn
ok (oc . och . ogh
 - ...
 - Söderwall supp nn **ok**, n. se uk.

Fig. 2. The FSvReader highlights all text words linked to the same lexicon entry, in this case *ok* ('and') in the start passages of the *Younger Västgöotalagen*

Etcetera

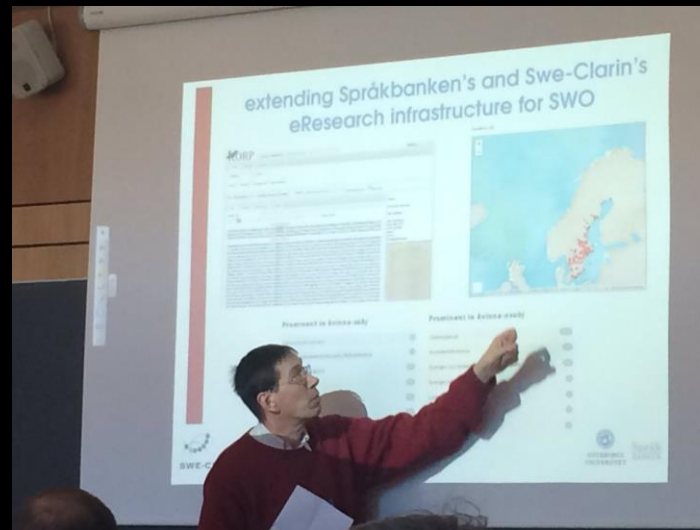
- How can one see the growth of the consumer society in older Swedish novels (1830–1860)?
- Grammar books contain valuable information on thousands of languages. Can this information be extracted automatically?
- How can *data science* and language technology be used to analyse language change and societal change over time?

<https://spraakbanken.gu.se/swe/forskning>

<https://spraakbanken.gu.se/swe/publikationer>



TODO TO DO TOGETHER



Things we do

- We educate in language technology
 - and supervise master theses
- We supervise PhD students
 - <https://spraakbanken.gu.se/swe/phd-program>
- We inform about our resources, our research, and language
 - workshops
 - the Språkbanken blog: <https://spraakbanken.gu.se/blogg>
- We answer questions from researchers, and from the public
 - <https://spraakbanken.gu.se/swe/info>

Things we can do, todo to do

- We can organise a workshop to help you or your students to get started with Språkbanken and language technology
- We can help you to use Språkbanken for your own research
- We can cooperate with you to find new research problems



GÖTEBORGS
UNIVERSITET



Yvonne Adessm



David Alter



Kristoffer Andersson



Malin Antonsson



Aleksandrs (Sasha) Berticevskis



Lars Borin



Gerlof Bouma



Dana Dannels

CONTACT

[HTTPS://SPRAAKBANKEN.GU.SE/](https://spraaakbanken.gu.se/)

[SB-INFO@SVENSKA.GU.SE](mailto:sb-info@svenska.gu.se)



Lelf-Jöran Olsson



Ildikó Pilán



Jacobo Rouces



Johan Roxendal



Nina Tahmasebi



Charalambos (Harris) Themistocleous



Elena Volodina



Niklas Zechner



Anna Lindahl



Peter Ljunglöf



Arild Matsson



Martin Hammarstedt



Richard Johansson



Jenny Kierkemann



Dimitrios Kokkinakis



Stian Rødven Eide



Kristina Lundholm Fors



Markus Forsberg



Luis Nieto Piña



Dan Rosen



Carl-Johan Schenström



Anne Schumacher



Jonathan Uppström



Shafqat Mumtaz Virk



Maria Ohrman