SPRÅKBANKENTEXT

A research infrastructure for language data
and a language technology research unit

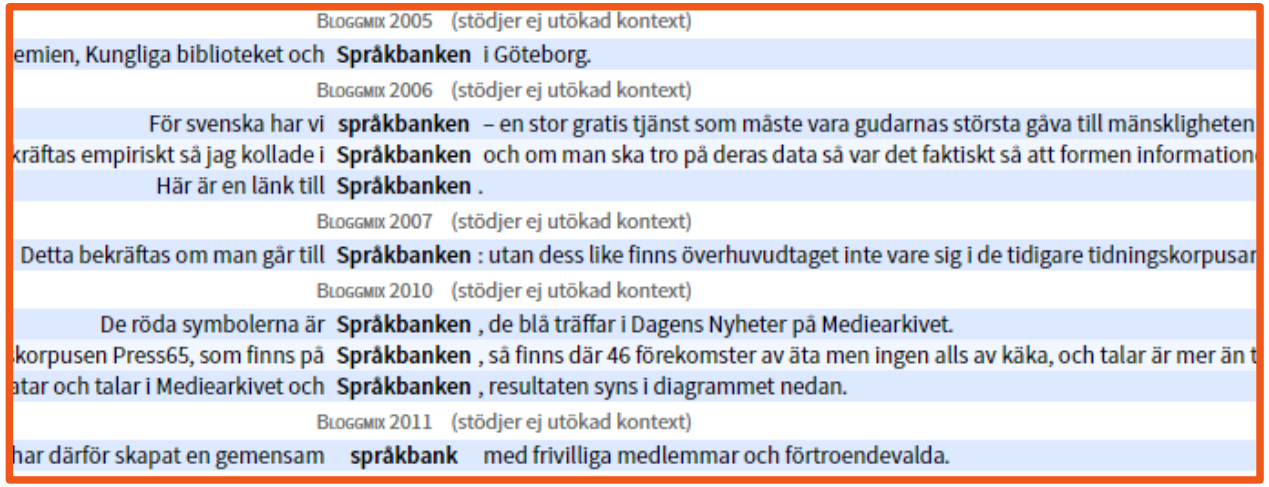
UNIVERSITY OF GOTHENBURG

SPRÅKBANKEN FOR BEGINNERS

2025

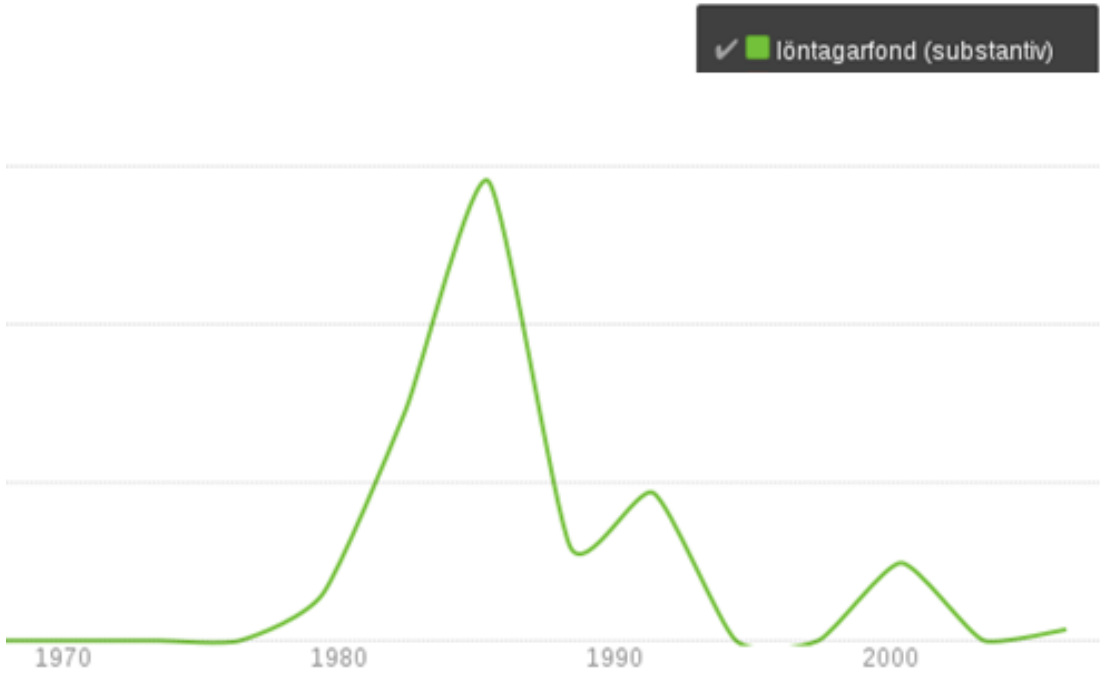
CONTENTS

- What is Språkbanken Text?
- Our data
- Our platforms
- Example applications
- Other things we do
- What we can help you with



WHAT IS SPRÅKBANKEN TEXT?

- A **bank** creates profit
- Språk**banken** creates knowledge



WHO NEEDS SPRÅKBANKEN TEXT?

Any researcher in the field of human communication:

- computer scientists
- philologists
- gender scientists
- historians
- health researchers
- cultural scientists
- linguists
- literary scholars
- media scientists
- education researchers
- psychologists
- social anthropologists
- language technologists
- political scientists
- and more!

SPRÅKBANKENTEXT

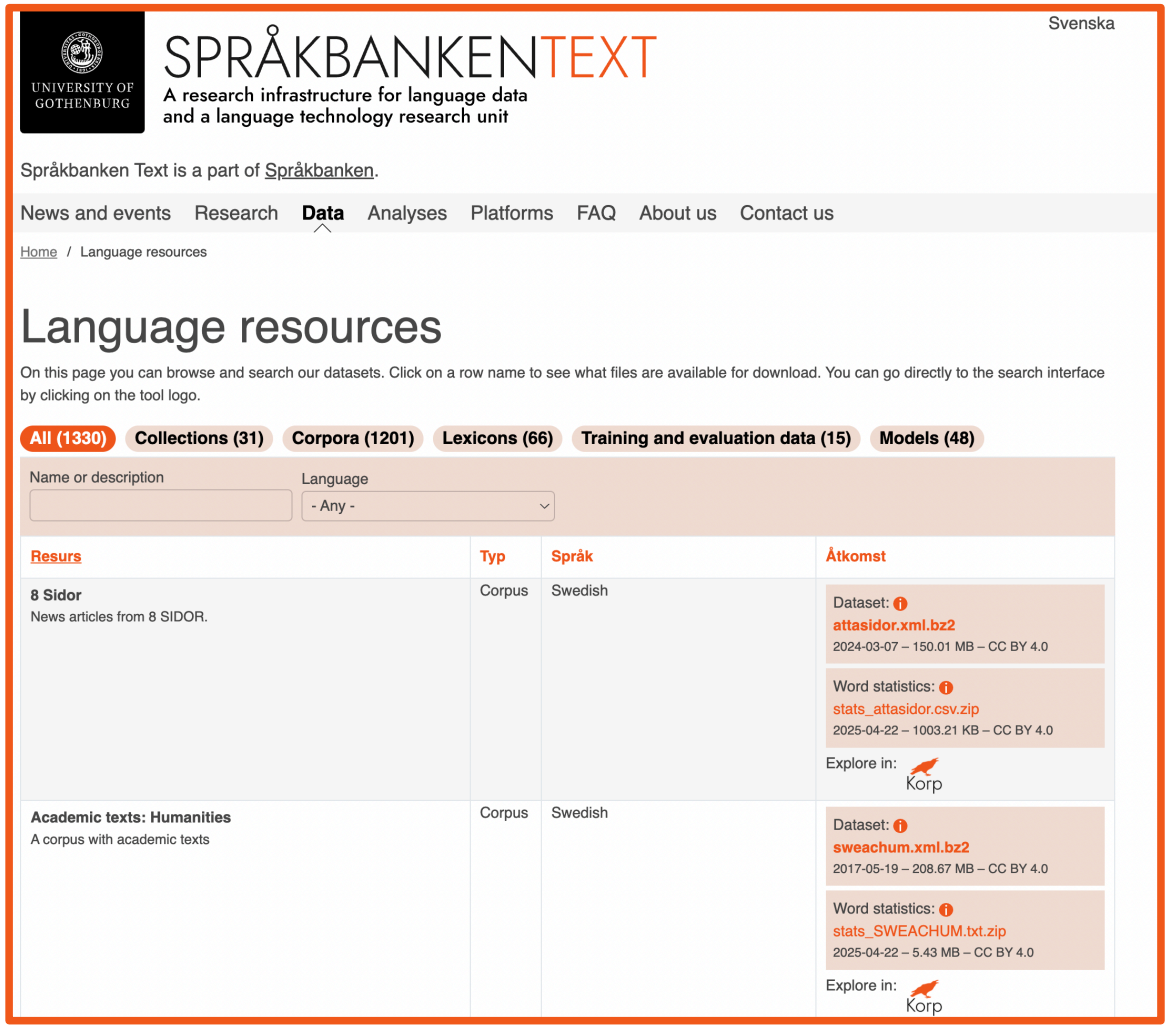
THERE ARE MORE THAN ONE SPRÅKBANK

- **Språkbanken** is a research infrastructure financed jointly by the Swedish Research Council and 10 universities and government organisations. It consists of four divisions.
- This presentation is about **Språkbanken Text**, which works with written language data.
- **Språkbanken Tal** is situated in Stockholm and works with spoken language.
- **Språkbanken Sam** works with archival material, together with Isof.
- **Språkbanken Clarin** is our connection to the international research infrastructure Clarin.

OUR DATA

- More than 1000 corpora
- More than 50 lexicons

You can seach for and download them on our webpage:
spraakbanken.gu.se/en/resources



WHAT IS A CORPUS?

- A large searchable collection of texts.
- To improve searchability, corpora are often annotated, that is, linguistic information has been added.
- The annotation is often done by a computer, and may contain errors.

Annotation gives information:

- **Morphological**: *Boken* is a noun, singular definite form, the same word as *bok*, *böcker* and *böckerna*
- **Syntactic**: *Jag läser en bok*: *bok* är the object
- **Semantic**: Vilken sense of *bok* – "book" or "beech"?
- **Metadata**: In what context is the word used? Who wrote the text? When?

SOME OF SPRÅKBANKEN'S CORPORA

- fiction
- newspapers
- magazines
- social media
- blogs
- webforums
- Twitter
- government documents
- Wikipedia
- Finland Swedish
- historical texts
- Old Swedish
- modern Swedish
- parallel texts
- the same texts in 25+ languages
- the Gigaword corpus
- a large, balanced corpus of different genres of modern Swedish from 1950 forward

SOME OF SPRÅKBANKEN'S LEXICONS

- SALDO: a semantisc and morphological lexicon
- Hellquist's Swedish etymological dictionary
- Bliss symbol lexicon
- Schlyter's and Söderwall's dictionaries of medieval Swedish
- attitude lexicon
- lexicon of borrowed words
- Swesaurus (a Swedish word net, a Swedish conceptual lexicon)

Our lexicons are made for automatic text analysis.

SPRÅKBANKEN'S MOST IMPORTANT PLATFORMS

Korp 
Språkbanken's word research platform

Karp 
Språkbanken's data publishing platform

Sparv 
Språkbanken's analysis platform

Strix 
Språkbanken's text research platform

Mink 
Språkbanken's data platform

- **Korp:** a search engine which gives you access to 30 000 million words in Språkbanken's corpora.
- **Karp:** a platform for working with our lexicons.
- **Sparv:** an annotation platform for adding morphological, syntactic and semantic information to text.
- **Strix:** a search engine which is document centered and can focus more on the semantic content of the document.
- **Mink:** a platform where you can apply our language technology methods to your own texts.

WHAT IS KORP?

Korp

Språkbanken's word research platform

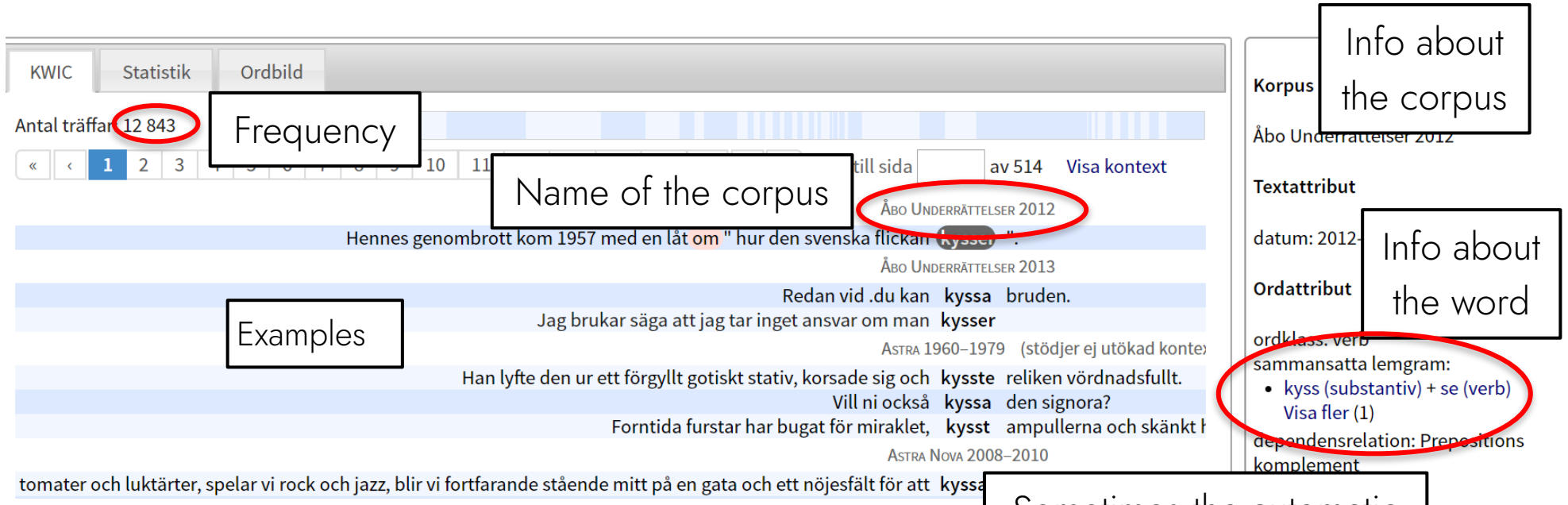
- a search engine
- made for investigating linguistic phenomena

WORD SEARCH

Korp

Språkbanken's word research platform

- Gives information about the frequency of a word and shows all its occurrences in context. The word *kyssa* "kiss":

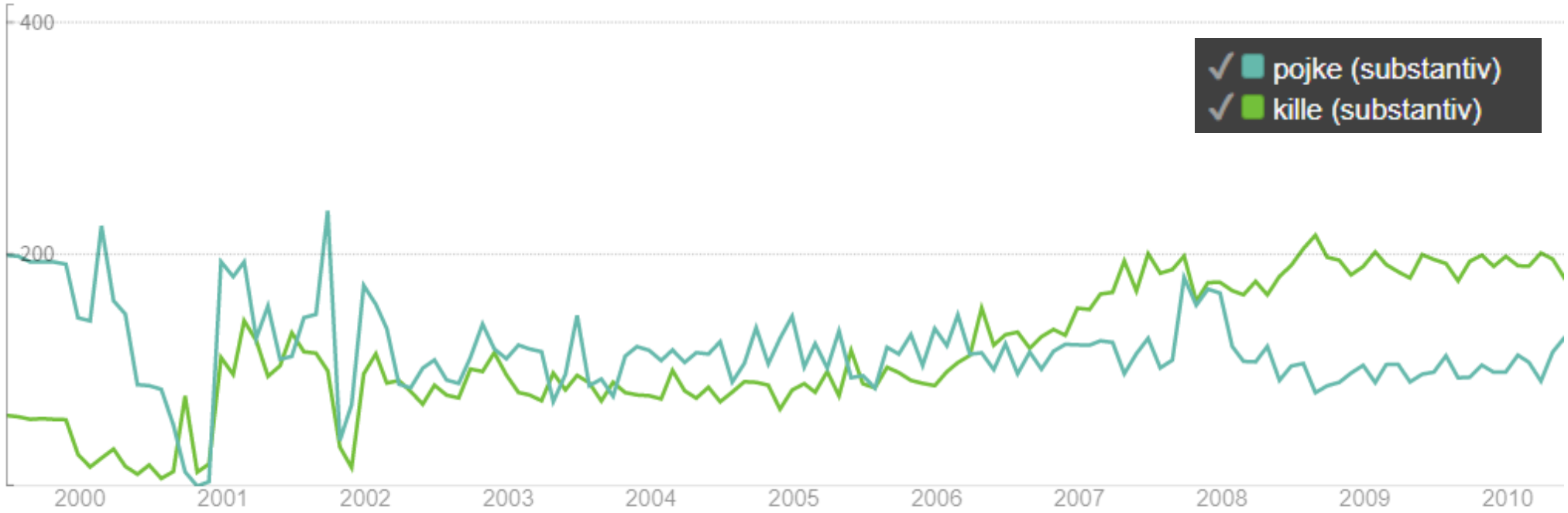


SPRÅKBANKENTEXT

TREND DIAGRAMS

Korp
Språkbanken's word research platform

- Language change:



COMPARISONS BASED ON PROBABILITY

Korp
Språkbanken's word research platform

- Which political parties focus more on education, and which focus more on healthcare?

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

SPRÅKBANKENTEXT

WORD PICTURES

Korp
Språkbanken's word research platform

- Who kisses whom/what, and in what way?

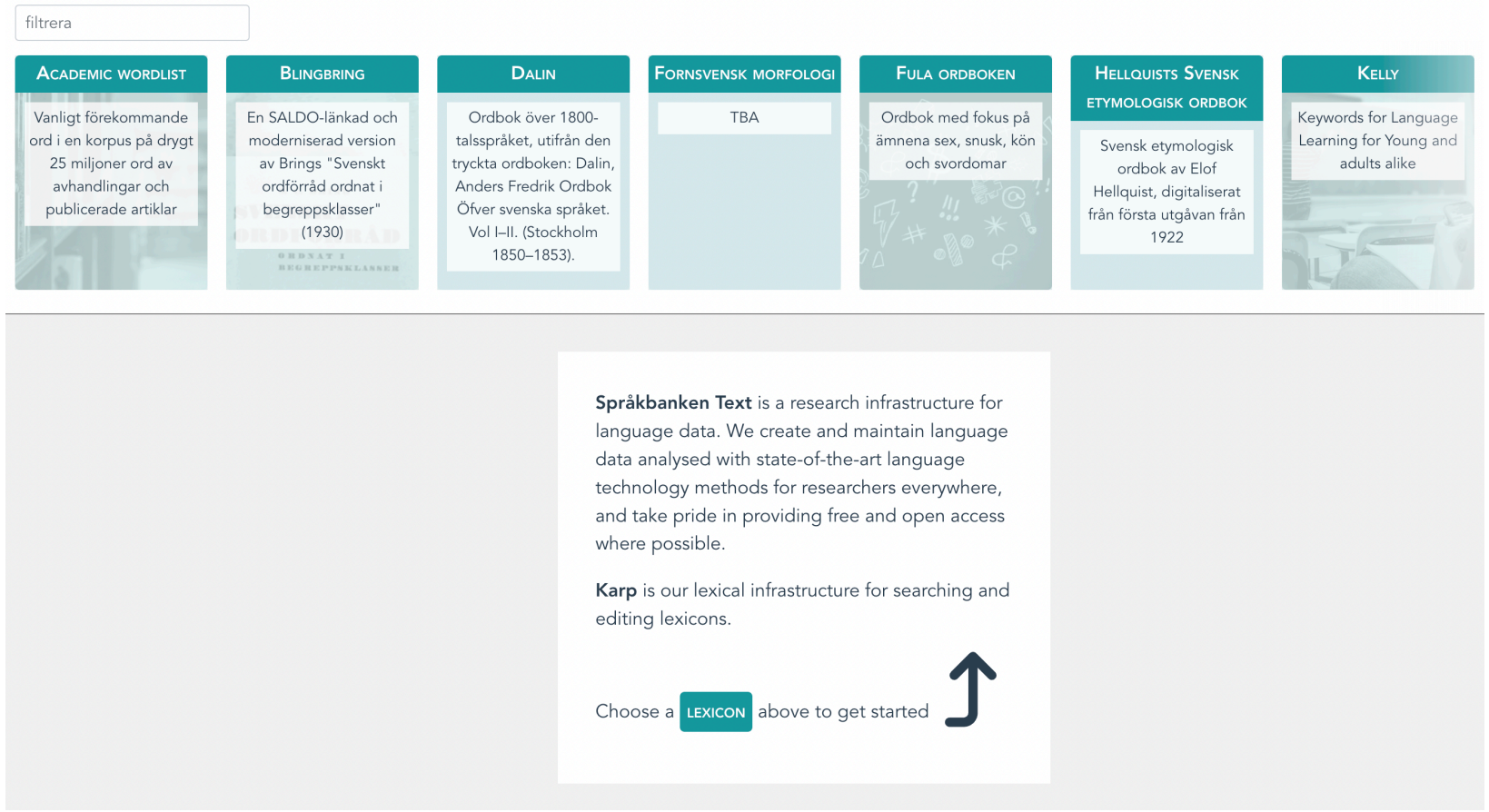
Subjekt		kyssa		Objekt		Adverbial	
1. pojke	41	1. tjej	181	1. på kind	86
2. främling	25	2. främling	97	2. på mun	72
3. kille	42	3. fot	136	3. i regn	52
4. man	46	4. groda	93	4. i nacke	49
5. kvinna	39	5. flicka	103	5. senast	63
6. hand	18	6. hand	107	6. på panna	33
7. kille[2]	23	7. kille	97	7. på hand	33
8. tjej	26	8. klubbmärke	45	8. första gången	38
9. oskuld	10	9. kille[2]	73	9. gång[2]	72
10. gaypar	8	10. kompis	71	10. gång	55
11. gång[2]	22	11. kind	51	11. gång[3]	55
12. hemlandshustru	6	12. tomte	51	12. sen	68
13. sol	20	13. tomt	51	13. i arsle	18
14. sol[2]	20	14. läpp	57	14. på hals	17
15. påve	11	15. farväl	48	15. under vecka	25

SPRÅKBANKENTEXT

WHAT IS KARP?

Karp
Språkbanken's data publishing platform

- A tool for working with Språkbanken's lexicons.

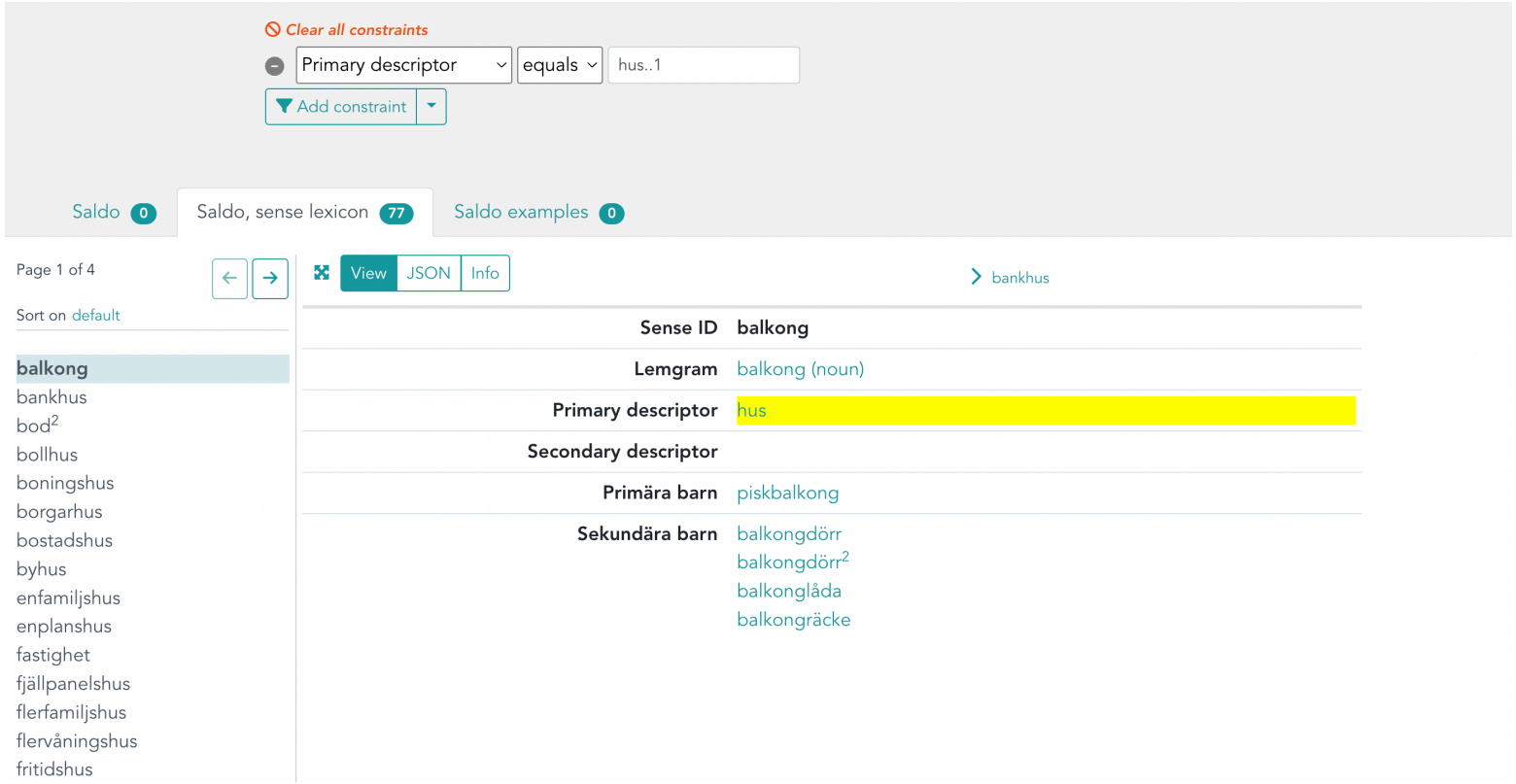


SPRÅKBANKENTEXT

SALDO

Karp
Språkbanken's data publishing platform

- What is the semantic neigbourhood for *hus* "house"?

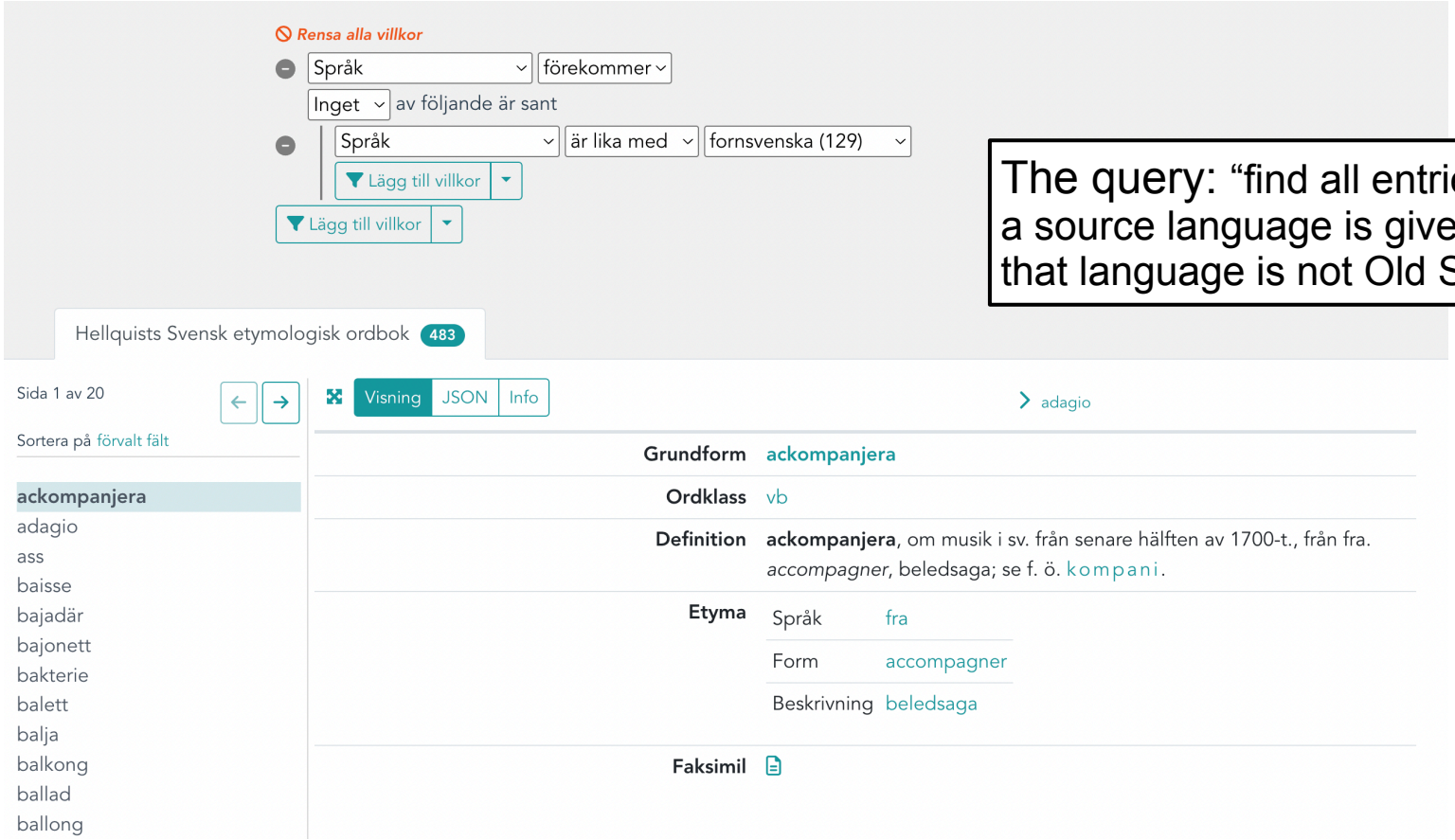


SPRÅKBANKENTEXT

ETYMOLOGICAL DICTIONARY

Karp
Språkbanken's data publishing platform

- From which languages does Swedish have the most loan words?



The query: "find all entries where a source language is given, and that language is not Old Swedish"

SPRÅKBANKENTEXT

MORE ABOUT KORP

Korp
Språkbanken's word research platform

- Let's search for the word *grym* "cruel, wicked":

Different senses!

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 svårbegripliga kapitel i Sveriges historia.
Häcken är en	**grym**	klubb, en skön klubb.
en varje månad ser arrangörerna till att ge besökarna tung bas och	**grym**	electro.
ndratals 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.
Som programledaren Timo Räisänen konstaterade: ” Fatta vad	**grymma**	vi är på musik i det här landet! ”
Den sista löpningen är	**grym**	och berör mig starkt.
ndratals 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.
glehundarna i England kommer att utsättas för experiment som är	**grymmare**	än de i Sverige.
nu långt 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.

SPRÅKBANKENTEXT

AUTOMATIC TEXT ANALYS

Korp
Språkbanken's word research platform

- Senses in Korp:

	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.
an, Vägen mot Bålberget, ger sig Therése Söderlind i kast med detta	grymma	och svårbegripliga kapitel i Sveriges historia.
Häcken är en	grym	klubb, en skön klubb.
gen varje månad ser arrangörerna till att ge besökarna tung bas och	grym	electro.
ındratals 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.
Som programledaren Timo Räisänen konstaterade: ” Fatta vad	grymma	vi är på musik i det här landet! ”
Den sista löpningen är	grym	och berör mig starkt.
ındratals 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.
ıglehundarna i England kommer att utsättas för experiment som är	grymmare	än de i Sverige.
ınu långt 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.

Gå till sida [ ] av 19

Korpus

GP 2013

Textattribut

artikelförfattare: Torbjörn Skarhed
artikelavdelning: Kultur Nöje
datum: 2013-02-22

Ordattribut

ordklass: adjektiv
sammansatta lemgram: *[tom]*
dependensrelation: Subjektspredikativ (subjektiv predikatsfyllnad)
förled: *[tom]*
efterled: *[tom]*
betydelse:
- grym[2]
  Visa fler (1)

msd: JJ.POS.UTR.SIN.IND.NOM
sammansatta ordformer: *[tom]*
grundform:
grym
lemgram:
grym (adjektiv)

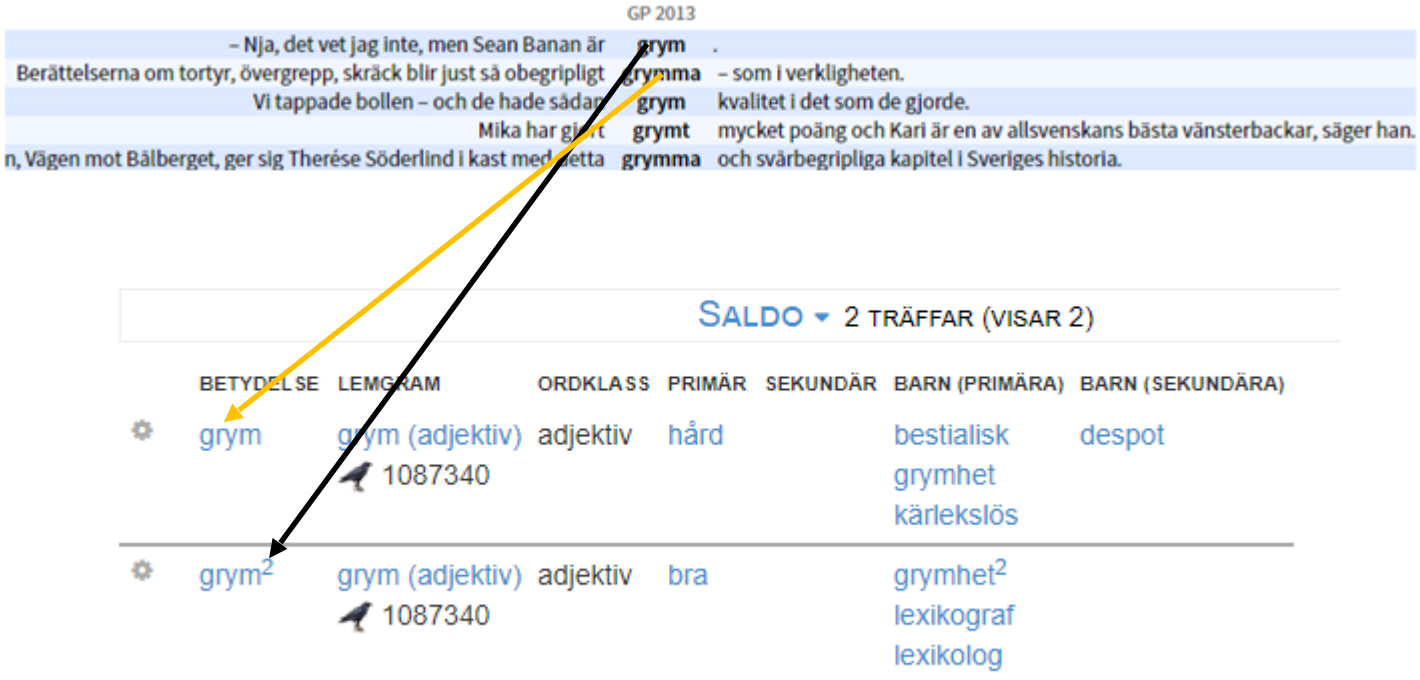
Visa dependensträd

SPRÅKBANKENTEXT

AUTOMATIC TEXT ANALYS

Korp
Språkbanken's word research platform

- The senses in Korp come from the SALDO lexicon:



WHAT IS SPARV?

Sparv
Språkbanken's analysis platform

Sparv can annotate your texts, and add:

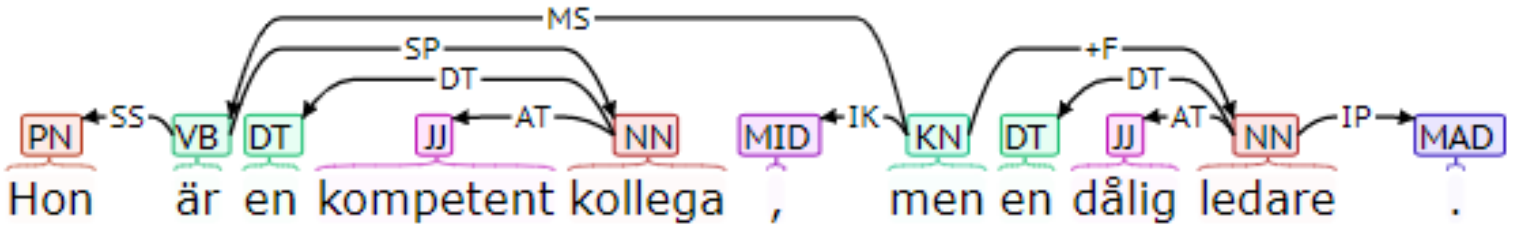
- lexical analysis
- syntactic analysis
- semantic analysis
- sentiment analysis
- readability analysis
- and more!

SYNTAKTIC ANALYSIS

Sparv
Språkbanken's analysis platform

Hon är en kompetent kollega, men en dålig ledare.

"She is a competent colleague, but a poor leader."



SEMANTISK ANALYS

Sparv
Språkbanken's analysis platform

Hon är en kompetent kollega, men en dålig ledare.

The word *ledare* has at least three senses:

Sense	Probability in this sentence, according to Sparv
leader	0,65
electric conductor	0,21
editorial	0,14

SENTIMENT ANALYSIS

Sparv
Språkbanken's analysis platform

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

WHAT IS STRIX?

Strix
Språkbanken's text research platform

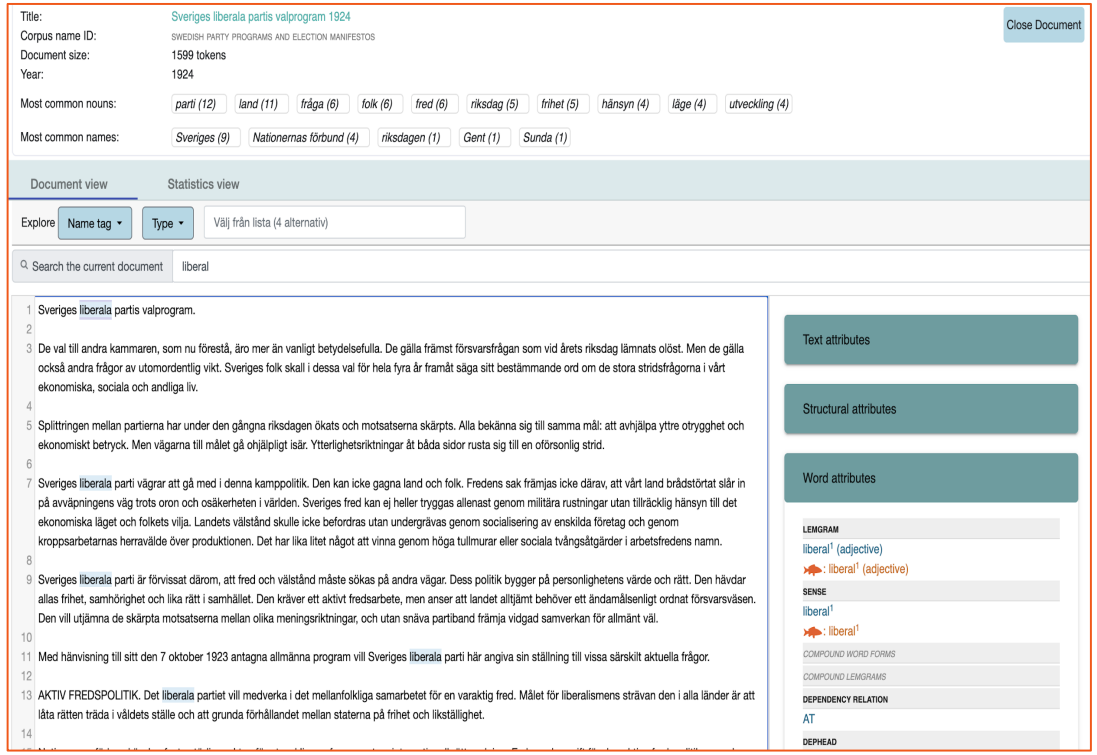
A search engine which:

- is focused on entire documents (unlike Korp, which focuses on words)
- analyses the semantic contents of the document

WHAT CAN STRIX DO?

Strix
Språkbanken's text research platform

- text analysis
- find keywords
- find names
- sentiment analysis
- topic classification
- show entire annotated documents
- work with collections of documents
- show metadata

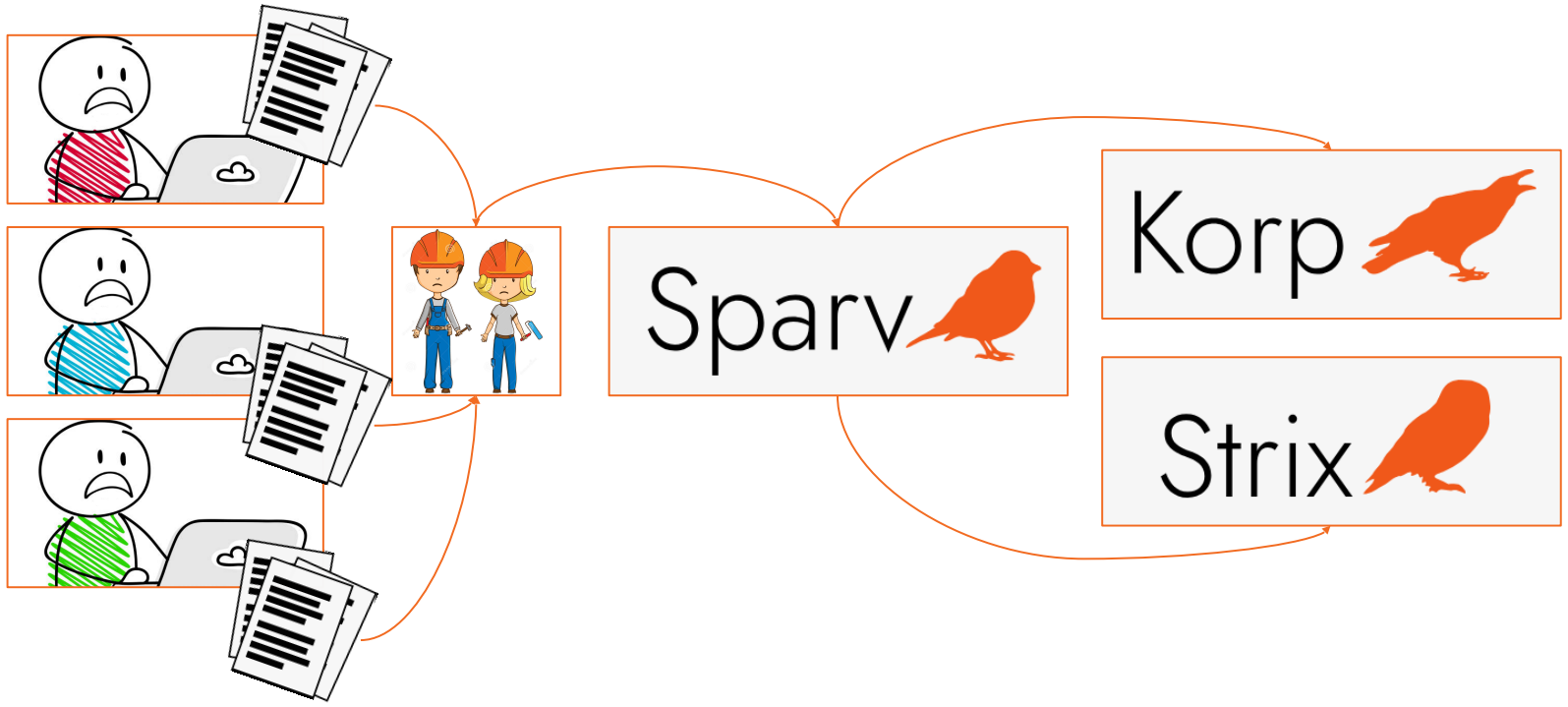


SPRÅKBANKENTEXT

WHAT IS MINK?

Mink
Språkbanken's data platform

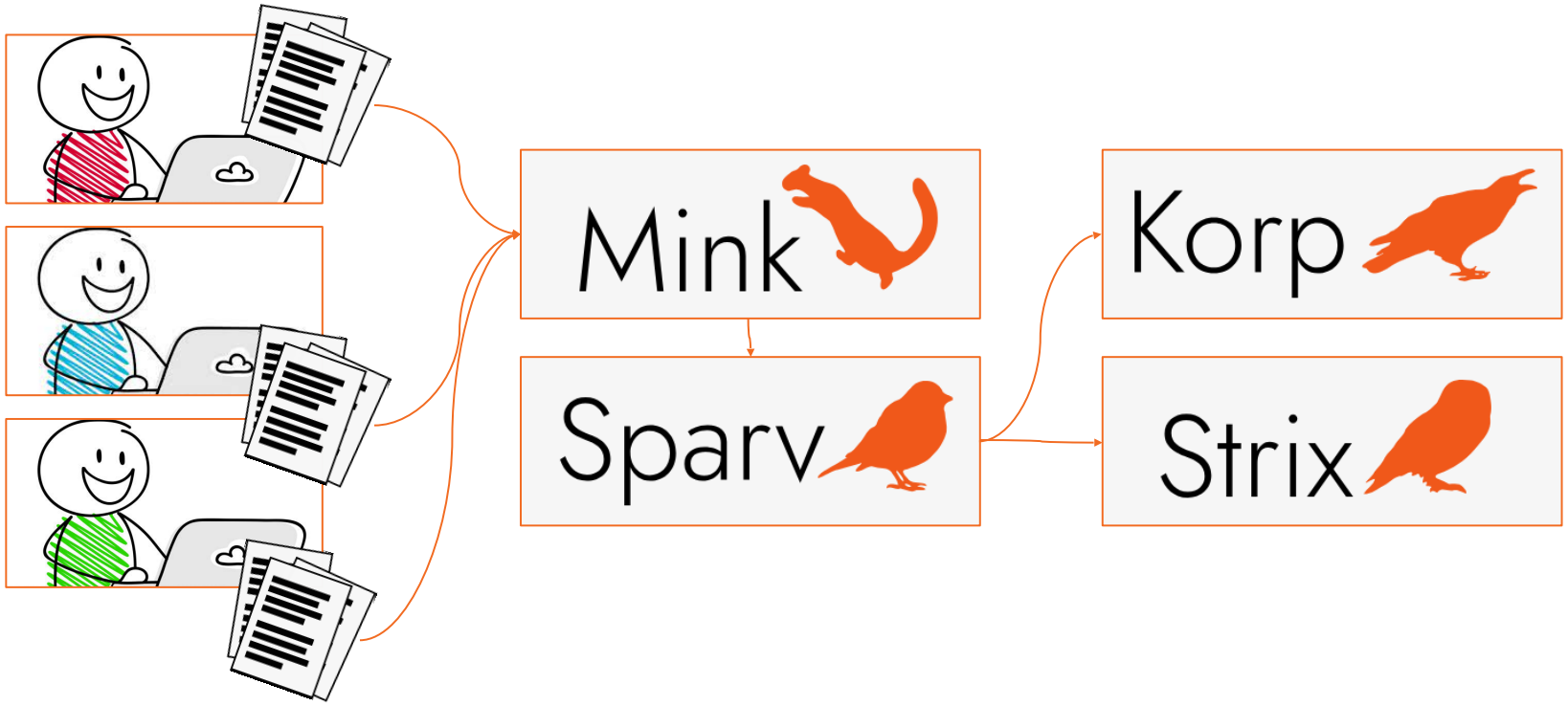
Adding new corpora involves Språkbanken Text.



WHAT IS MINK?

Mink
Språkbanken's data platform

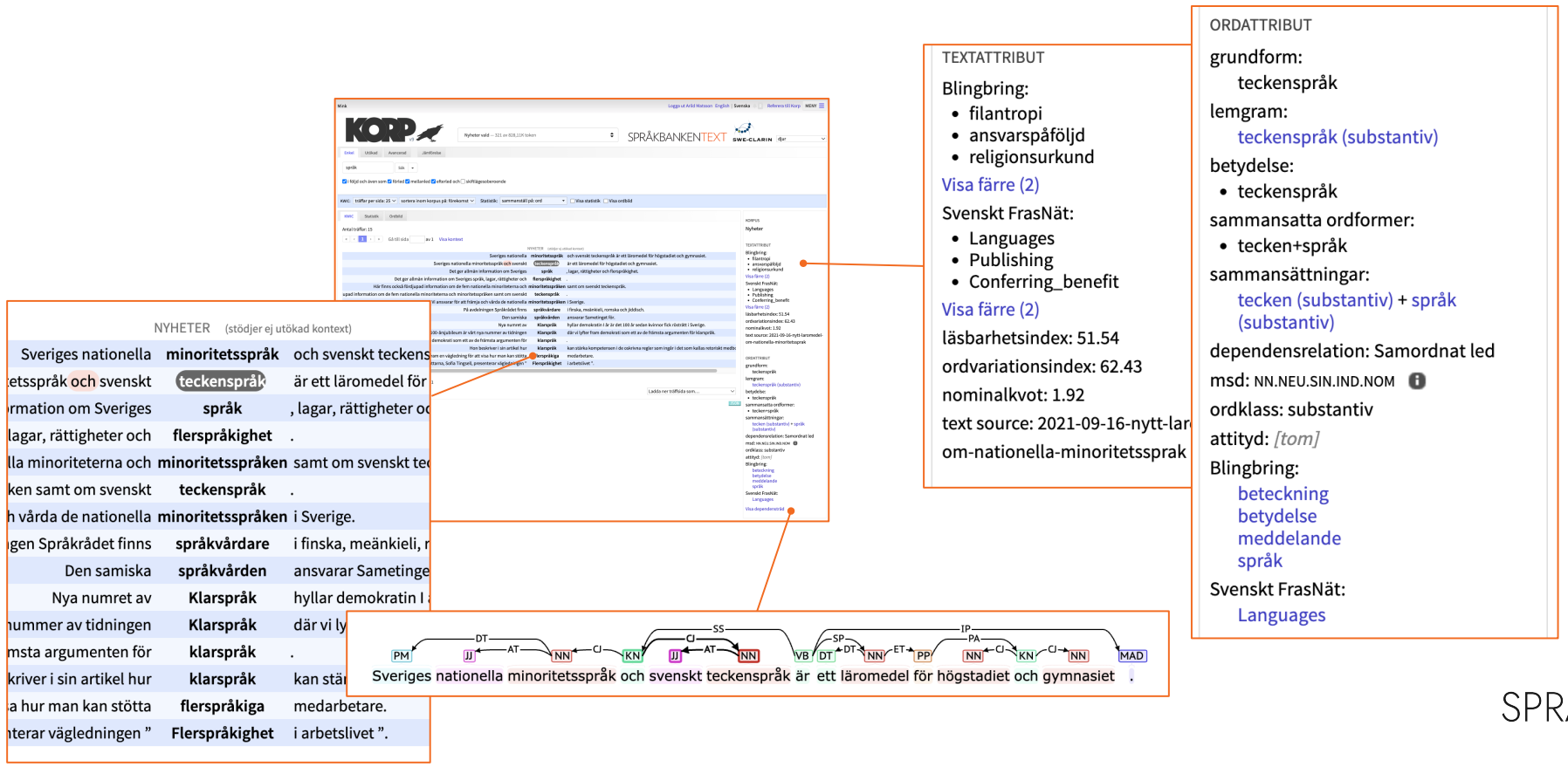
With Mink, you can add and analyse your own texts!



WHAT IS MINK?

Mink
Språkbanken's data platform

Then Sparv runs in the background and annotates your text.



SPRÅKBANKENTEXT

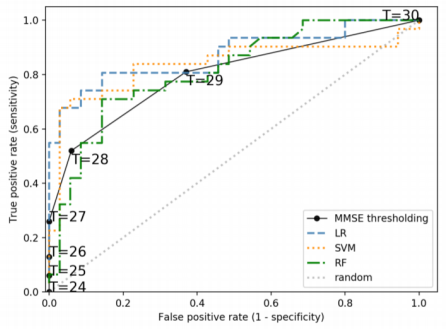
TECHNICAL USE OF OUR PLATFORMS

- Most of our tools are open source.
- Korp and Sparv can be used through an API.
- Korp can be adapted to other corpora and other languages, and is used in Italy, Estonia, Finland, Iceland, Denmark och Norway.

EXAMPLES OF APPLICATIONS

Language technology supports medical diagnostics:

- A common test used to diagnose mild cognitive impairment (MCI) has an accuracy ("AUC score") of 68%.
- If we supplement the test with automatic analysis of the patient's language, the accuracy is improved to 87%.

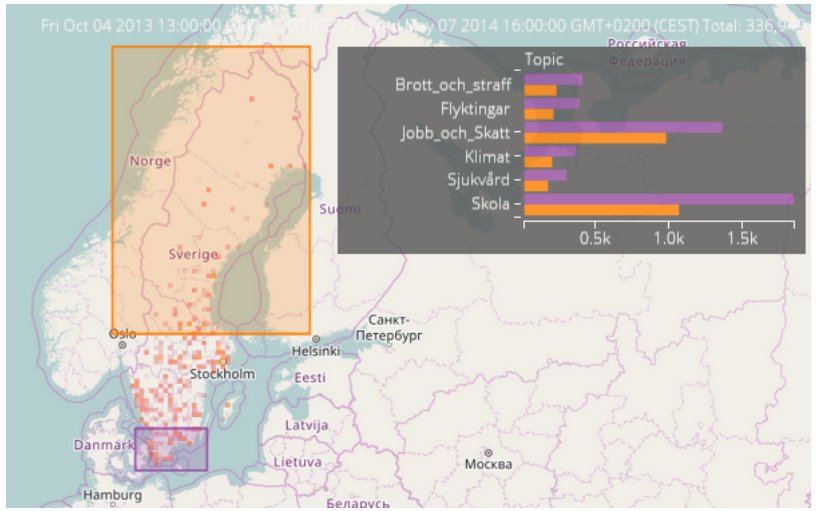


Fraser, K., Lundholm Fors, K., Eckerström, M., Themistocleous, C., & Kokkinakis, D. (2018). Improving the sensitivity and specificity of MCI screening with linguistic information.

EXAMPLES OF APPLICATIONS

What are Swedes discussing on X (Twitter) right now?

- Orange versus purple: How much is the topic discussed in northern versus southern Sverige?
- We see no big differences in the distribution.



Borin, L., & Kosiński, T. 2016. Towards interactive visualization of public discourse in time and space

EXAMPLES OF APPLICATIONS

Simplifying the reading of historic texts.

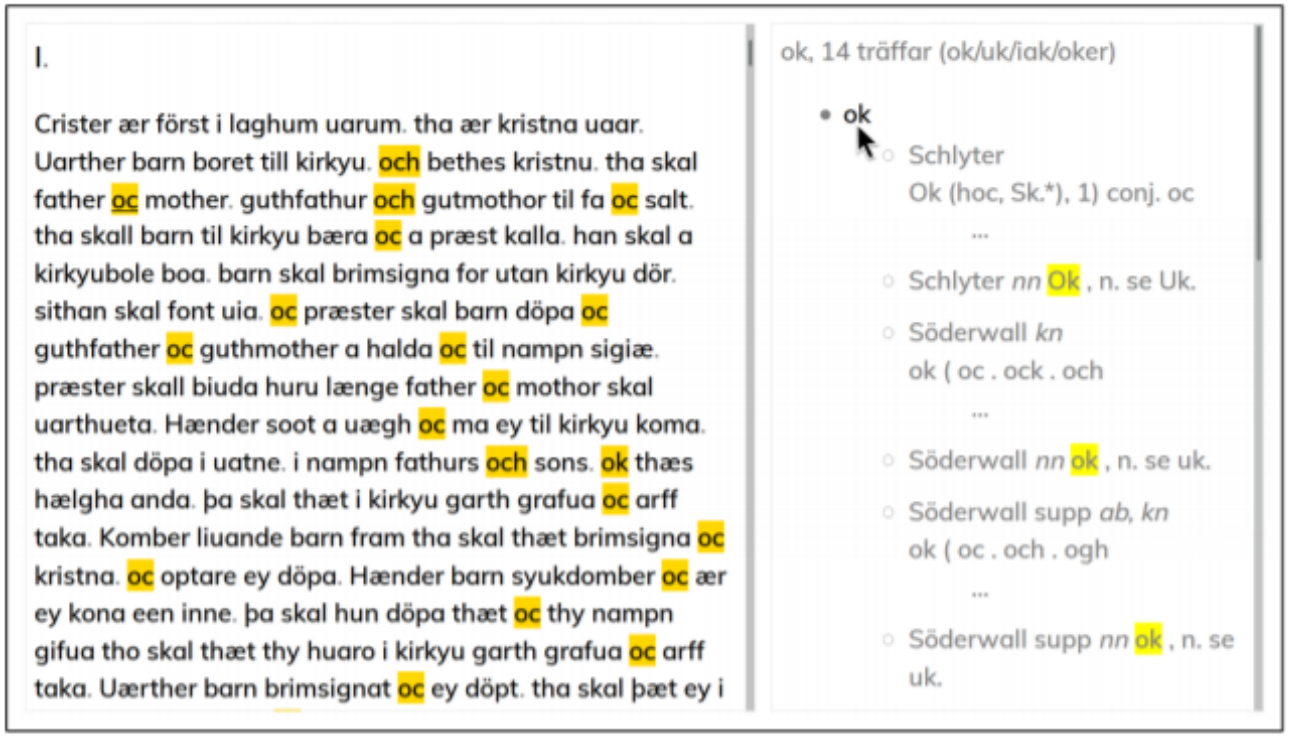


Fig. 2. The FSvReader highlights all text words linked to the same lexicon entry, in this case *ok* ('and') in the start passages of the *Younger Västgötalagen*

Adesam, Y., Ahlberg, M., & Bouma, G. 2018. FSvReader – Exploring Old Swedish cultural heritage texts.

EXAMPLES OF APPLICATIONS

More examples of research where language technology plays a part:

- How can we see the growth of consumerism in older Swedish novels (1830–1860)?
- Grammar books contain valuable information about thousands of languages. Can we extract that information automatically?
- How can we use "big data" and language technology to analyse changes in language and society over time?

spraakbanken.gu.se/en/research

OTHER THINGS WE DO

- Teach language technology
- Supervise doctoral students
- Inform about our resources, our research, and language in general
- Answer questions from researchers and the public
- Arrange workshops
- Arrange user days
- Blog

WHAT CAN WE HELP YOU WITH

Other things we can do:

- Arrange a workshop to help you or your students get started with Språkbanken Text and language tecnology.
- Help you use Språkbanken Text for your research question.
- Cooperate to find new research questions.

SPRÅKBANKENTEXT

GÖTEBORGS UNIVERSITET

A research infrastructure for language data and a language technology research unit

spraakbanken.gu.se