Language resources and tools

Markus Forsberg

Språkbanken
University of Gothenburg

GF Summer School 2015

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

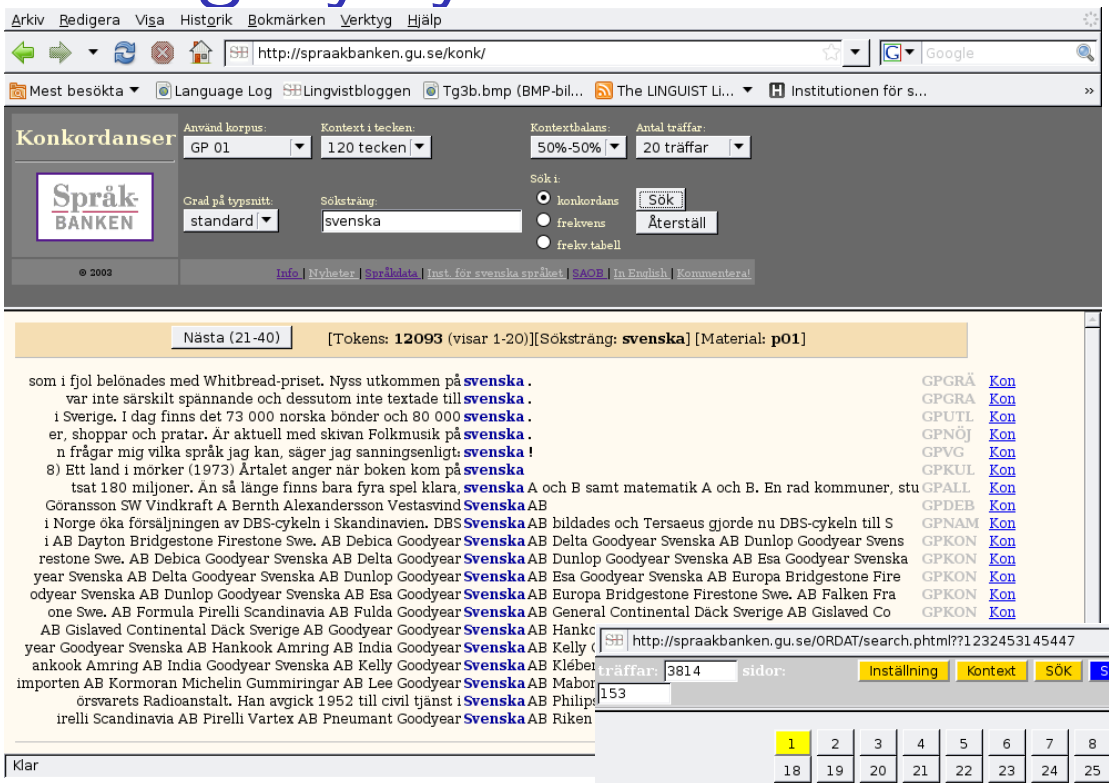
Today's talk

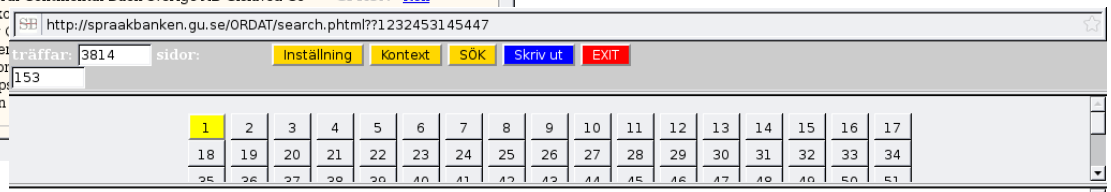
- Language resources and tools at Språkbanken (the Swedish language bank).
- A quick introduction to Corpus Workbench.
- Demonstration of some of Språkbanken's tools.

A couple of years ago: Legacy systems at Språkbanken

- Språkbanken has been around since 1975.
- Service unit for linguists $\Rightarrow$ LT research unit.
- The old way to fly: a language resource = database + interface
- The structure of the LR was largely irrelevant (as long as everything looked nice in the interface).
- Made linguists (somewhat) happy, and LT researchers unhappy.

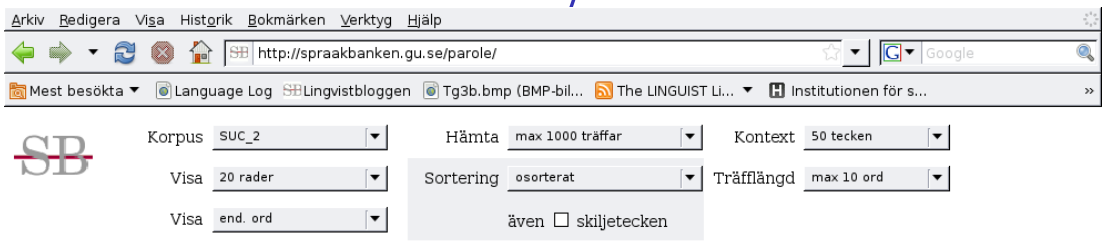
Legacy systems: konk and ORDAT, ...

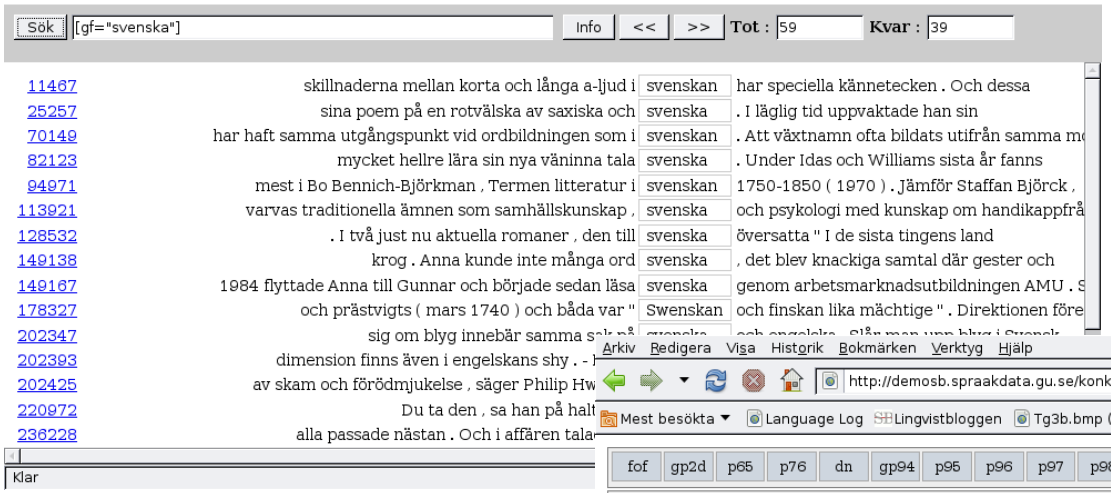


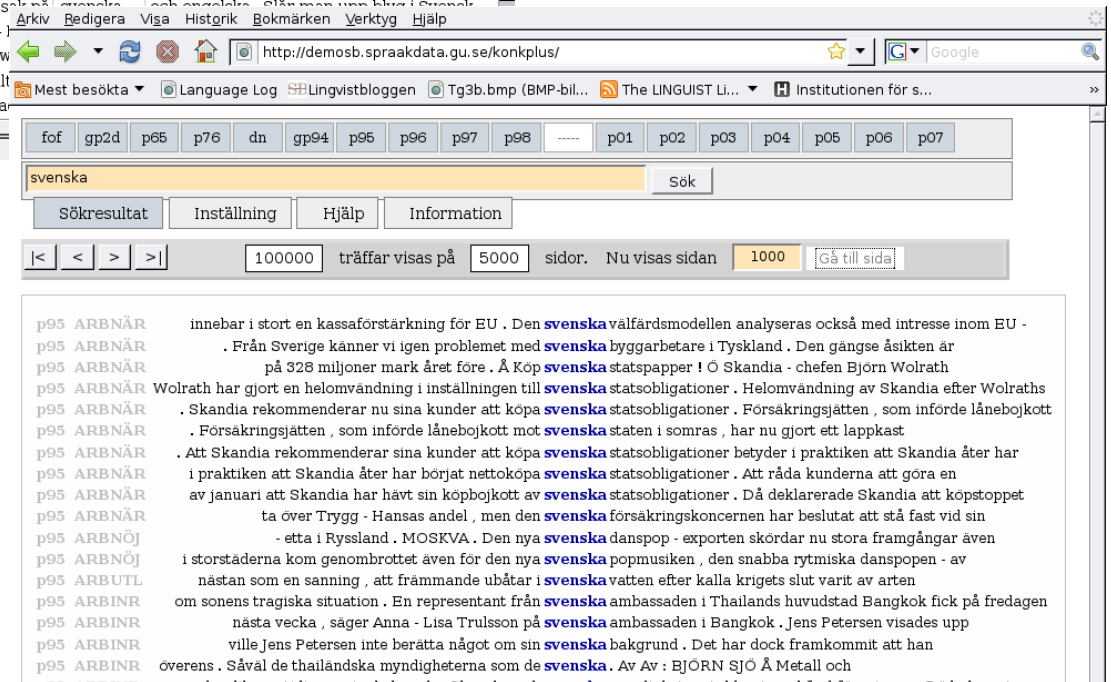


1923 rav bättre anpassade bestämmelser . 1923 års svenska motorkrönika rymmer visserligen utöver den nä
1923 ari på Edsviken främst , och de därvid satta svenska hastighetsrekorden för såväl bil som motorcyk
1923 r sig svårt att ge ett allmänt omdöme om den svenska penningmarknaden år 1923 . Från viss synpunkt
1923 g . I samband härmed bör kanske nämnas , att svenska folkets sammanlagda tillgodohavande å sparkas
1923 ller mindre delar av sina fonder , må nämnas Svenska Handelsbanken (105 , 0 mill . kr. avskr . tä
1923 lda bank (19 , 0 resp. 17 , 7 mill .) samt Svenska lantmännens bank (18 , 7 resp. 16 , 1 mill .
1923 För att rädda banken inträdde då som känt är svenska staten såsom huvuddelägare i banken . Karakte
1923 ar riksbankens räntepolitik betingats av den svenska kronans ställning till dollarn och därmed til
1923 utländska valutorna i allmänhet har dock den svenska kronan ej blott hävdat sin ställning , utan n
1923 lit under året icke blott i förhållande till svenska kronor utan även till pund . Den fruktan för
1923 siffror få gälla som norm . Jordbruket . Det svenska jordbruket har under de senaste åren kämpat m
1923 mport av vete under 1923 (se nedan) är det svenska jordbrukets läge f. n . bekymmersamt . För at
1923 sättningen av åtminstone större delen av det svenska vetet ha de större kvarnarna i september före
1923 betalningsbalans , vilken under den moderna svenska industriens grund läggningsperiod i slutet av
1923 913 . Den är så mycket egendomligare som den svenska textilindustrien i allmänhet synes ha god sys
1923 et givetvis ännu återstår mycket , innan den svenska industrien kan anses arbeta för full drift .
1923 egeringen , universitetskonsistoriet och den svenska minoriteten uppburna s. k. linjedelningsförsl
1923 , som om det antagits varit ägnat att trygga svenska språkets framtida ställning vid universitetet

. . . Parole/SUC and Konkplus, . . .

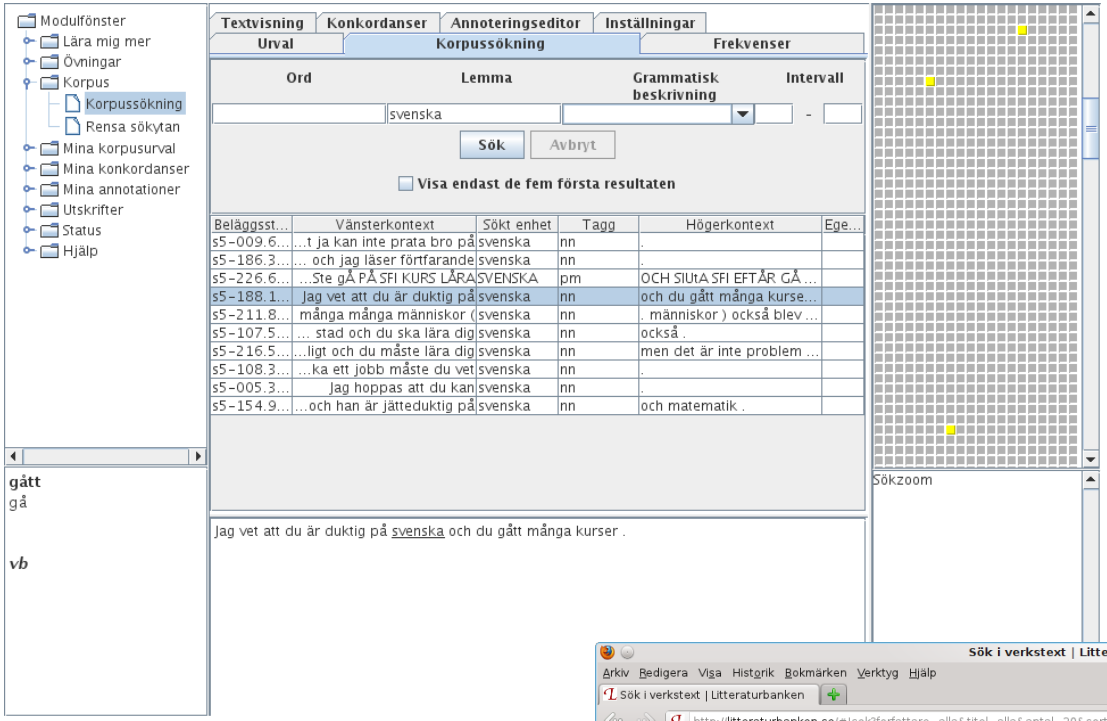


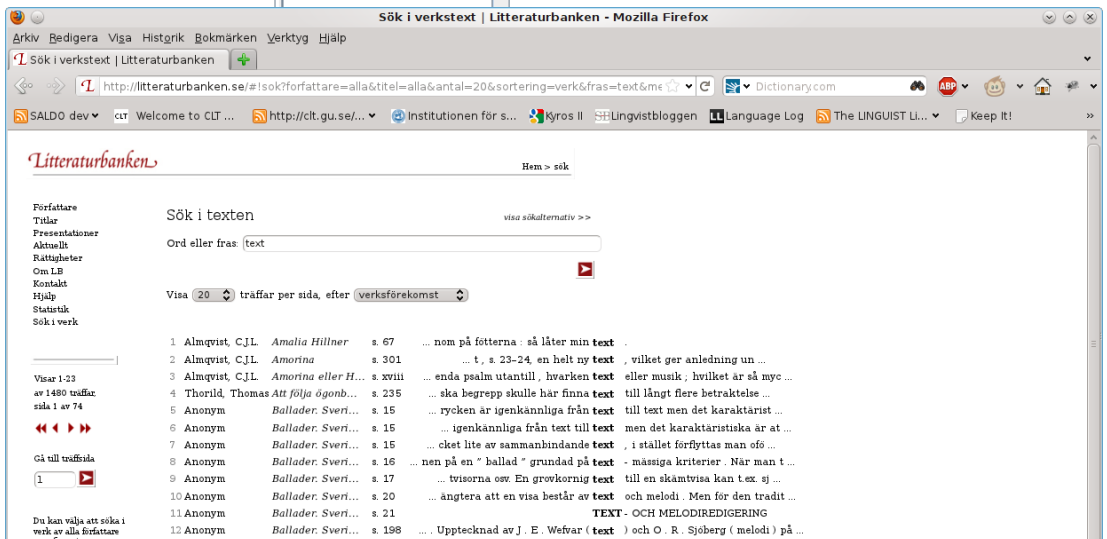




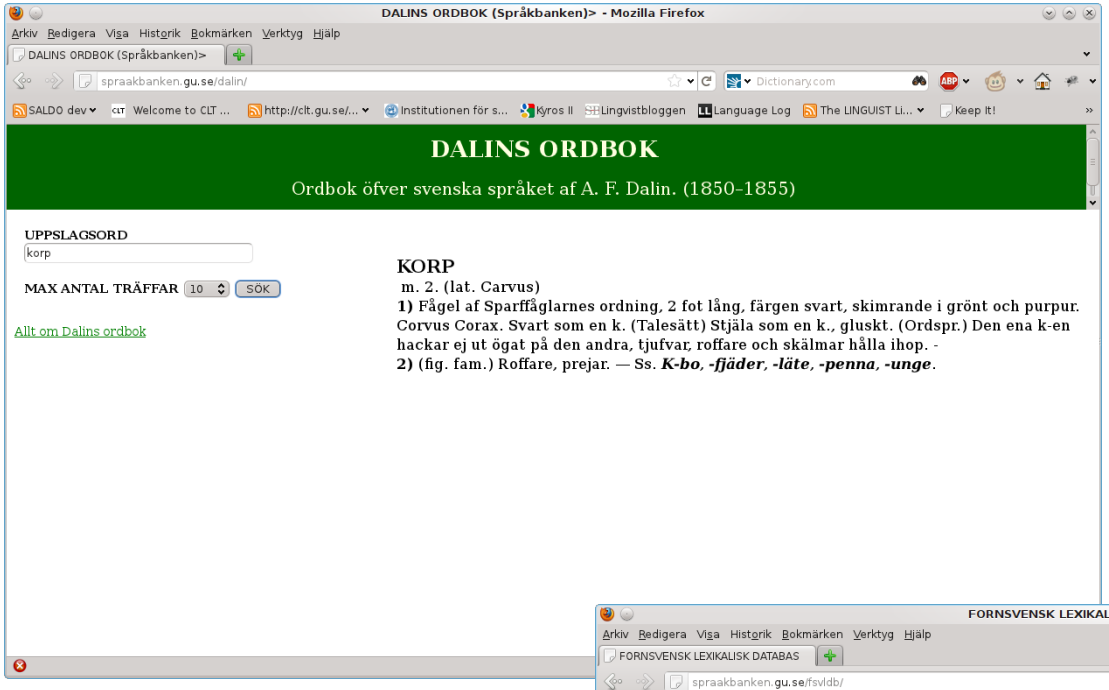
. . . ITG and Litteraturbanken, et cetera, moreover

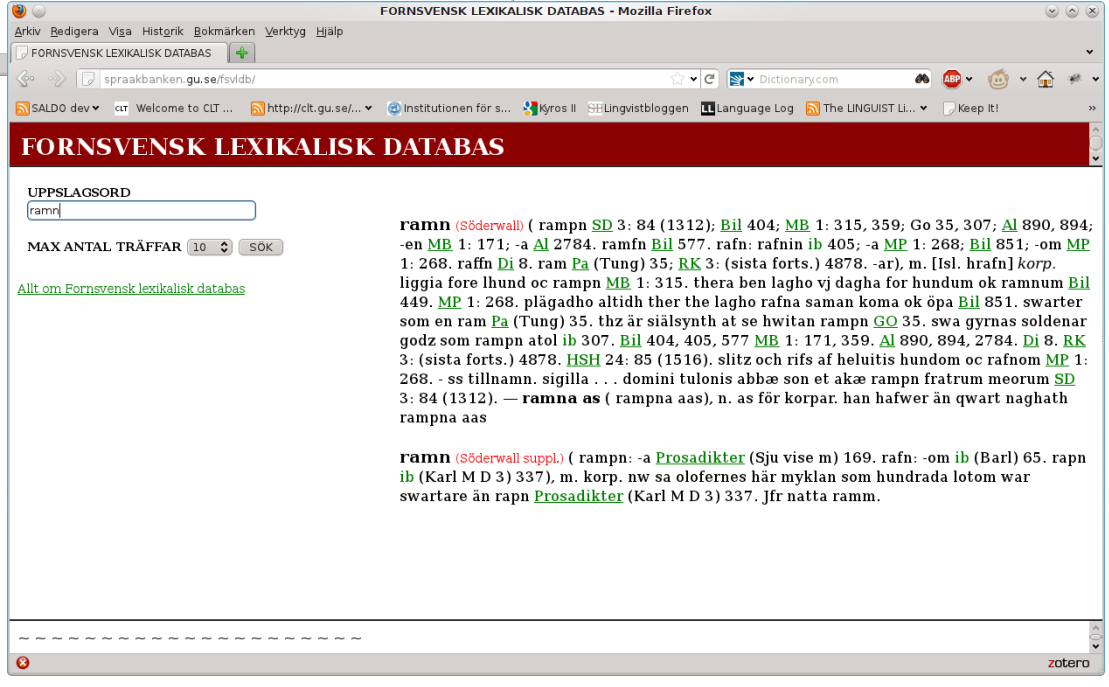
. . .



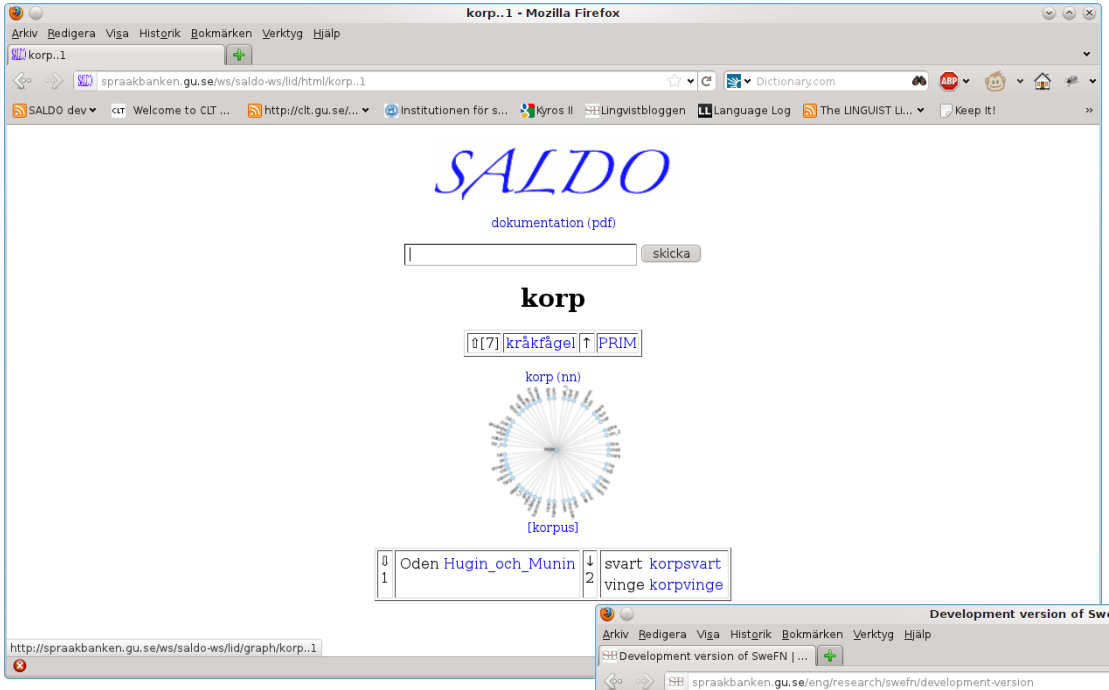


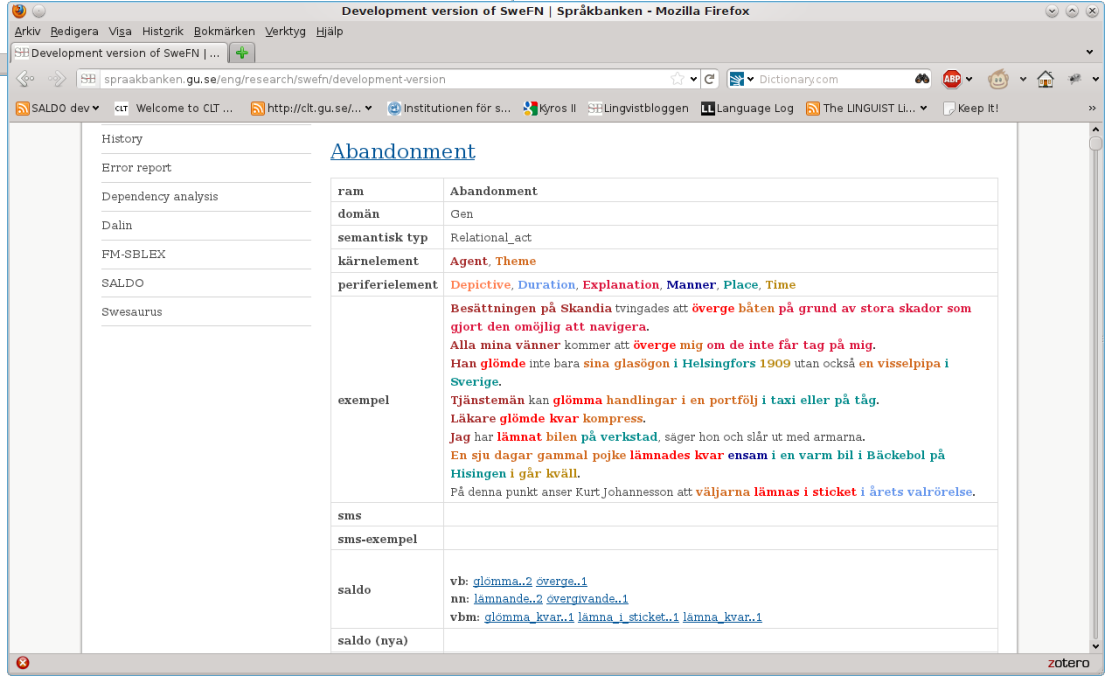
. . . Dalin and Old Swedish, . . .





. . . SALDO and SweFN, et cetera, et cetera

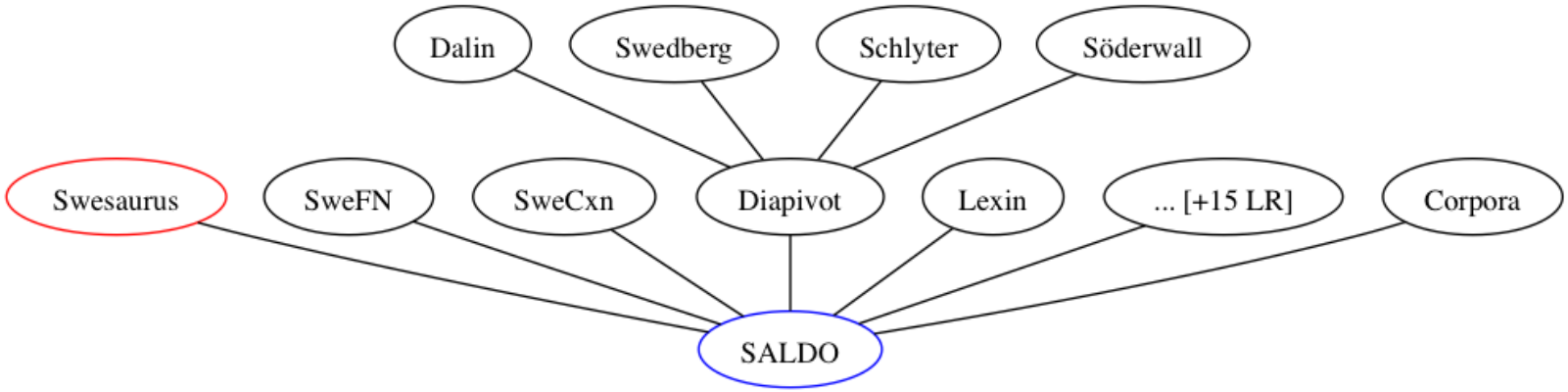




What to do?

Changing the situation

- Put the resources in the center, not the interfaces (**downloadable** resources in a common format, so far IPR permits)
- Centralize and think in term of research infrastructure (= technological solutions that try to enable as much new research as possible)
  - Korp – corpora infrastructure
  - Karp – lexical infrastructure
- Link all the resources to a pivot resource (GF speak: a lexical abstract syntax), SALDO; that is, create a large LT resource network (a macro-resource).



SALDO

- **SALDO** is a full-scale ($\sim$ 130k word senses, 2M word forms) lexical-semantic resource for Swedish with semantic relations between all word senses (including MWE).
- Available under an open license: CC-BY.
- SALDO is a directed graph with so called **primary** and **secondary** relations.
- The fundamental unit is the **word sense** (the first version of SALDO did only contain word senses).
- All word senses is given one or more formal descriptions, referred to as **lemgrams** (lemgram=paradigm+lemma $\rightarrow$inflection table)

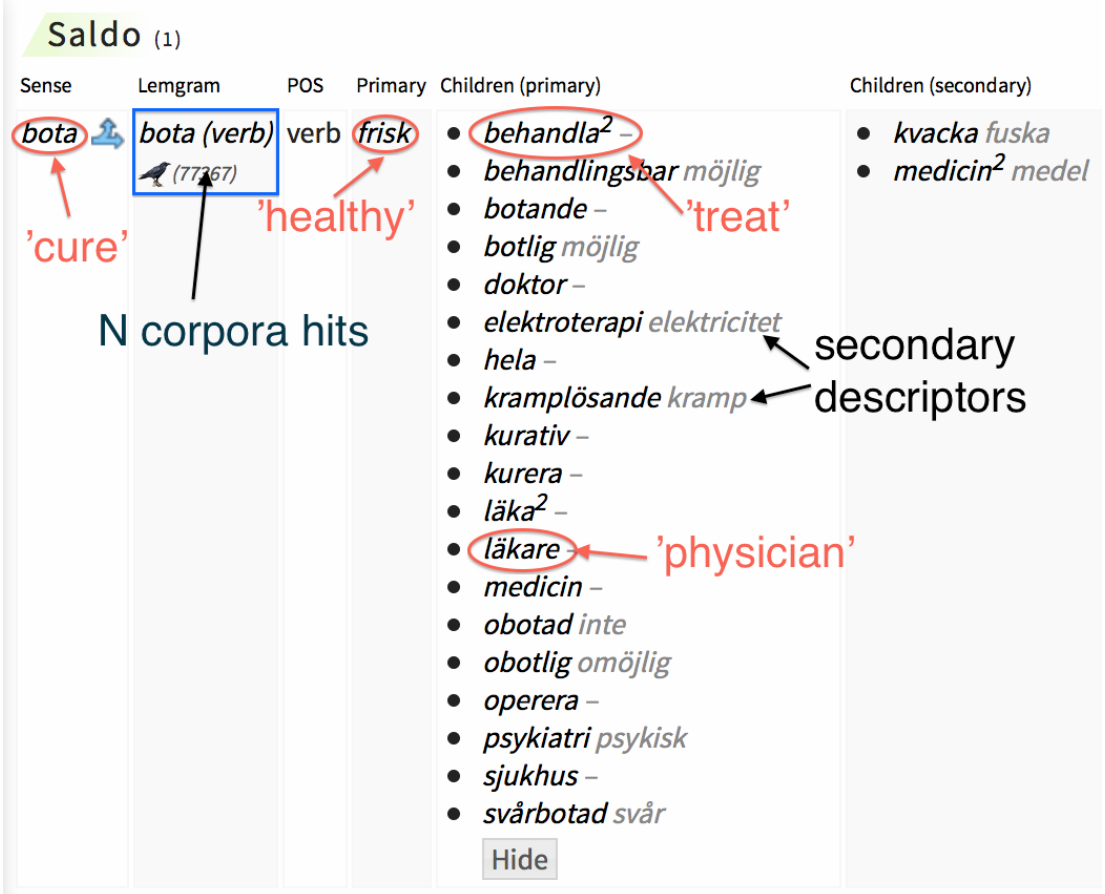
SALDO "PIDs"

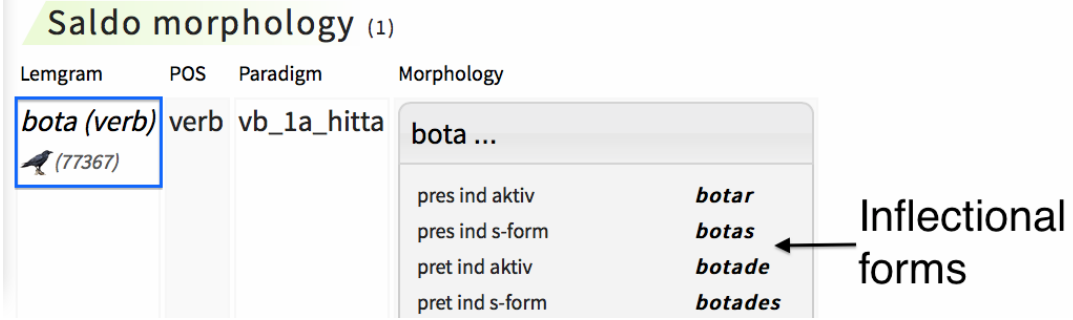
- SALDO has id's for:
  - senses (`grad..1`)
  - lemgrams (`grad..nn.1`)
  - parts of speech (`nn`)
  - paradigms (`nn_3u_film`)
- the id's are designed to be
  - **unique** (no other id's should be necessary, e.g., database keys)
  - **atomic** (no built-in assumptions about sense–subsense relationships, etc.)
  - **usable in Semantic Web formalisms** (RDF, OWL): id's are well-formed XML names
  - **human-readable** (makes resources easier to work with)

Details about SALDO

- All (except a few) have a obligatory **primary descriptor**, and an optional set of **secondary descriptors**.
- 41 senses lack primary descriptor, joined together with an artificial zero-sense *PRIM..1* (E.g., färg 'color', 'rak' 'straight', tänka 'think', ...)
- A primary descriptor should be **semantically close** and more **central**: more frequent, stylistically more neutral, morphologically simpler, and more.
- The secondary descriptors help discriminate the sense (no special criteria).

SALDO example: bota 'cure'





Linking backwards in time (I)

- Linking SALDO and Dalin (19th century Swedish) is relatively straightforward.
- The vocabulary differences are mainly in the compounds, e.g.:
  - *bäfverhund* 'dog used for beaver hunt'
  - *bäfverhund* → modernize → *bäverhund* → compound analysis → bäver..nn.1+**hund..nn.1**

Linking backwards in time (II)

- Linking the Old Swedish to SALDO is more challenging. An illustrative example:
  - bakvaþi *fatal accident resulting from a sword being struck backwards without the striker looking in that direction beforehand*
- Link to what? Accident? Sword? Both? Others?

Korp pipeline: the annotation lab

Svenska English

KORP **Annotation Laboratory**

Language of analysis: Swedish

Load example: Drama | Åtta sidor | Talbanken | Läsbart | Exempelkorpus

Editor | Upload

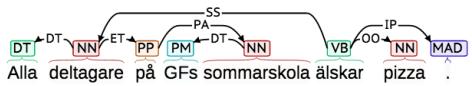
```
1 Alla deltagare på GFs sommarskola älskar pizza.
```

☑ **Lexical analysis** ☑ **Compound analysis** ☑ **Dependency analysis**

Show advanced settings

Run!

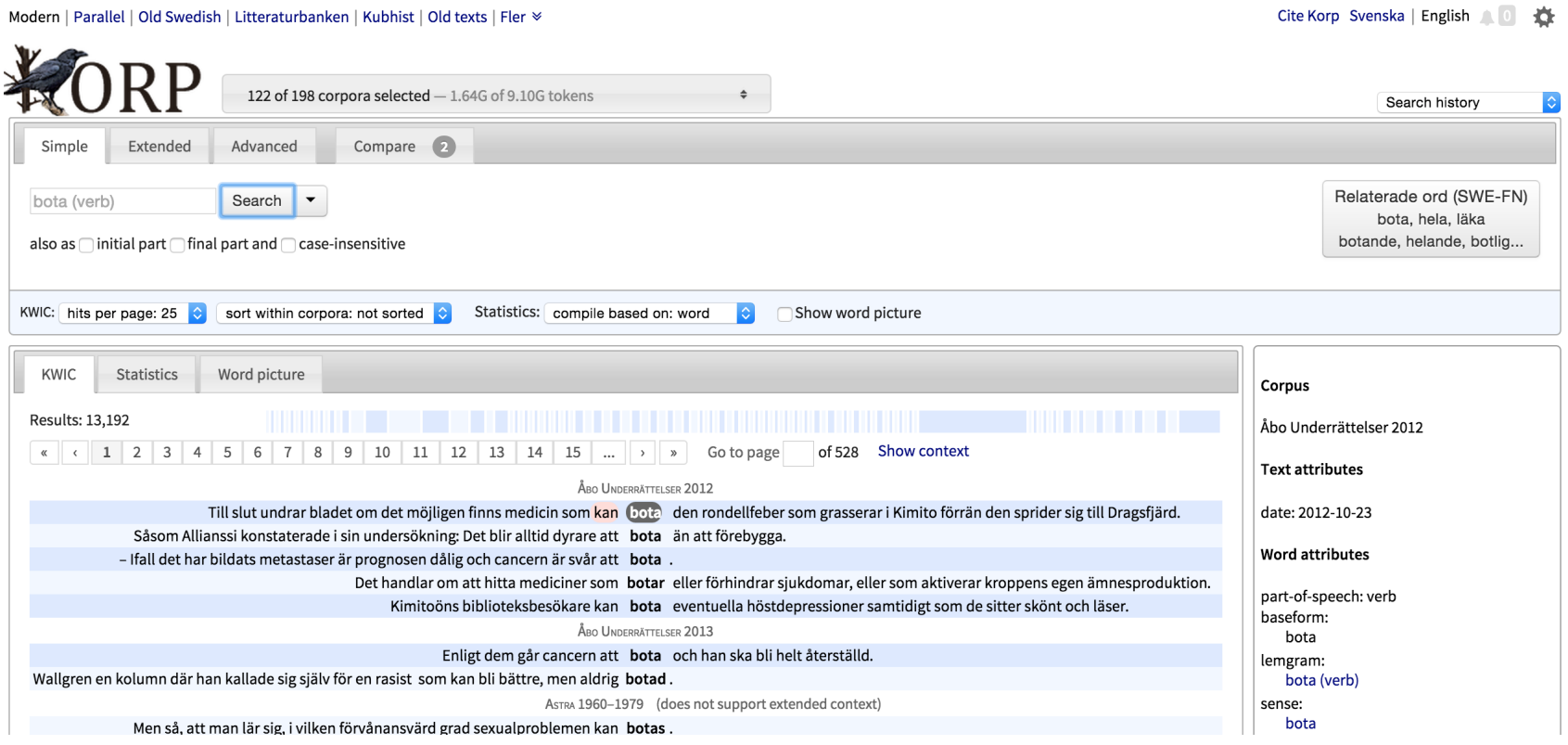
```
<paragraph> [Show XML]
<sentence id="8f7-83b"> [Show XML]
```



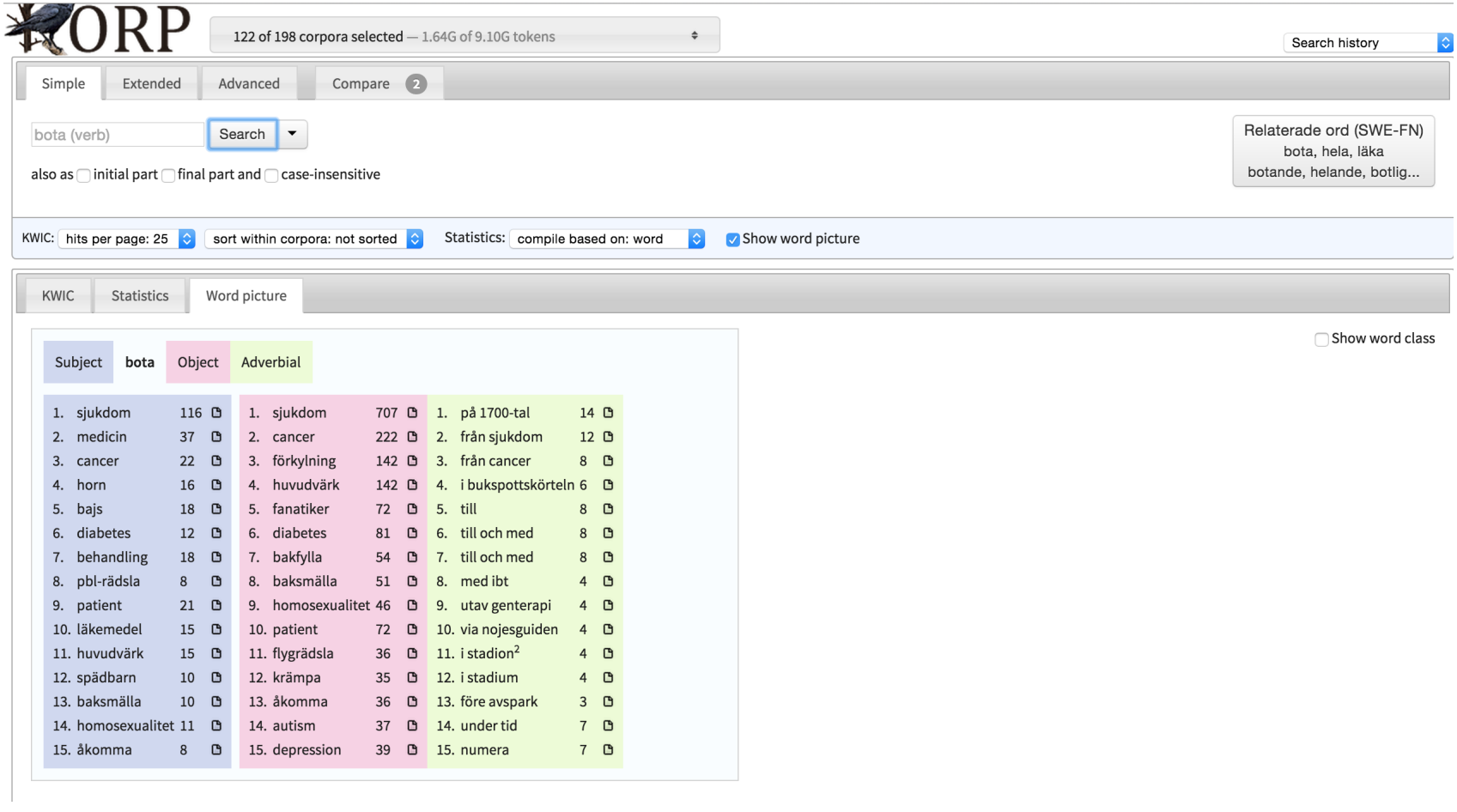
word	msd	lemma	lex	saldo	prefix	suffix	ref	dephead	deprel
Alla	DT. UTR+NEU. PLU. IND+DEF	all	all..pn.1	all..1			1	2	DT
deltagare	NN. UTR. PLU. IND. NOM	deltagare	deltagare..nn.1	deltagare..1			2	6	SS
på	PP	på	på..pp.1	på..1			3	2	ET
GFs	PM. GEN						4	5	DT
sommarskola	NN. UTR. SIN. IND. NOM				sommar..nn.1	skola..nn.1	5	3	PA
älskar	VB. PRS. AKT	älska	älska..vb.2, älska..vb.1	älska..2, älska..1			6		ROOT
pizza	NN. UTR. SIN. IND. NOM	pizza	pizza..nn.1	pizza..1			7	6	OO
.	MAD						8	6	IP

```
</sentence>
</paragraph>
```

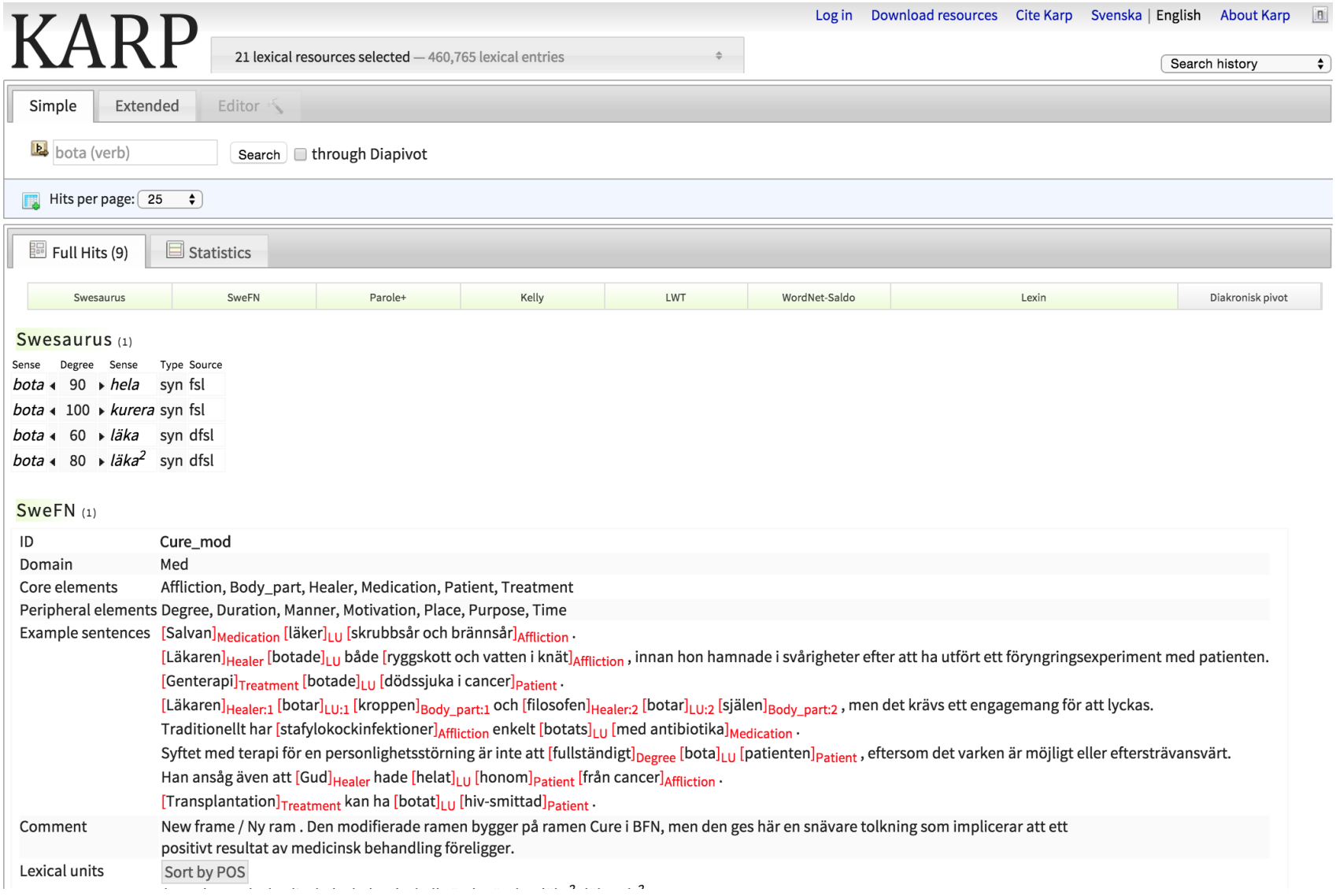
Korp: the corpora infrastructure



Korp: word picture



Karp


KARP
Log in
Download resources
Cite Karp
Svenska | English
About Karp
21 lexical resources selected — 460,765 lexical entries
Search history
Simple
Extended
Editor
bota (verb)
Search
through Diapivot
Hits per page: 25
Full Hits (9)
Statistics
Swesaurus
SweFN
Parole+
Kelly
LWT
WordNet-Saldo
Lexin
Diakronisk pivot
Swesaurus (1)
Sense Degree Sense Type Source
bota 90 hela syn fsl
bota 100 kurera syn fsl
bota 60 läka syn dfsl
bota 80 läka2 syn dfsl
SweFN (1)
ID Cure_mod
Domain Med
Core elements Affliction, Body_part, Healer, Medication, Patient, Treatment
Peripheral elements Degree, Duration, Manner, Motivation, Place, Purpose, Time
Example sentences [Salvan]Medication [läker]LU [skrubbsår och brännsår]Affliction .
[Läkaren]Healer [botade]LU både [ryggskott och vatten i knät]Affliction , innan hon hamnade i svårigheter efter att ha utfört ett föryngringsexperiment med patienten.
[Genterapi]Treatment [botade]LU [dödssjuka i cancer]Patient .
[Läkaren]Healer:1 [botar]LU:1 [kroppen]Body_part:1 och [filosofen]Healer:2 [botar]LU:2 [själen]Body_part:2 , men det krävs ett engagemang för att lyckas.
Traditionellt har [stafylokockinfektioner]Affliction enkelt [botats]LU [med antibiotika]Medication .
Syftet med terapi för en personlighetsstörning är inte att [fullständigt]Degree [bota]LU [patienten]Patient , eftersom det varken är möjligt eller eftersträvansvärt.
Han ansåg även att [Gud]Healer hade [helat]LU [honom]Patient [från cancer]Affliction .
[Transplantation]Treatment kan ha [botat]LU [hiv-smittad]Patient .
Comment New frame / Ny ram . Den modifierade ramen bygger på ramen Cure i BFN, men den ges här en snävare tolkning som implicerar att ett positivt resultat av medicinsk behandling föreligger.
Lexical units
Sort by POS

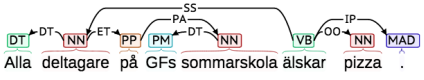
An quick introduction to Corpus Workbench

- A database system for querying annotated texts.
- Uses regular expressions over attributed words.
- Part of the backend of Korp.
- Input format:

<paragraph> [Visa XML]

<sentence id="8f7-83b"> [Visa XML]

PM: Egennamn



ord	msd	lemma	lex	saldo	prefix	suffix	ref	dephead	deprel
Alla	DT. UTR+NEU. PLU. IND+DEF	all	all..pn.1	all..1			1	2	DT
deltagare	NN. UTR. PLU. IND. NOM	deltagare	deltagare..nn.1	deltagare..1			2	6	SS
på	PP	på	på..pp.1	på..1			3	2	ET
GFs	PM. GEN						4	5	DT
sommarskola	NN. UTR. SIN. IND. NOM				sommar..nn.1	skola..nn.1	5	3	PA
älskar	VB. PRS. AKT	älska	älska..vb.2, älska..vb.1	älska..2, älska..1			6		ROOT
pizza	NN. UTR. SIN. IND. NOM	pizza	pizza..nn.1	pizza..1			7	6	OO
.	MAD						8	6	IP

</sentence>

</paragraph>

Corpus Query Language (CQL)

- Basic form, a box = word/token

  `[attr=value]`

- Example:

  `[word="pizza"]`

- Regular expression:

  `[word="pizz(a|or)"]`

- Boolean expression:

  `[word="pizz(a|or)" & (pos="VB" | pos="JJ")]`

Corpus Query Language (CQL)

- Comparisons: =, !=, <=, >=, !<=, !>=, (==, !==)

  ```
  [c >= 5]
  ```

- Sequences of tokens/words

  ```
  [word="älskar"] []{0,3} [word="pizza"]
  ```

- A longer example

  ```
  "catch|caught" [tag="DT"] [tag="JJ"]* [tag="N.*"] |
    [tag="N.*"] "was|were" "caught"
  ```

Demonstration: overview

1. Korp annotation lab
   <http://spraakbanken.gu.se/korp/annoteringslabb>
2. Korp
   <http://spraakbanken.gu.se/korp>
3. Karp
   <http://spraakbanken.gu.se/karp>