GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

KORP & KARP

Språkbanken's infrastructures for corpora and lexicons

EUROPEAN NETWORK OF e-LEXICOGRAPHY

Elena Volodina & Ildikó Pilán

COST ENeL WG3 meeting

Vienna, 12 February 2015

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

KORP & KARP origin

SweFN++ (2009-2015), a project where one of the objectives is to create, curate, and integrate free Swedish lexical resources --> for researchers and language technology applications

META-NORD (2011-2013), a project aiming at establishing an open linguistic infrastructure in the Baltic and Nordic countries.

http://spraakbanken.gu.se/korp
http://spraakbanken.gu.se/karp

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Architecture principles

Front-end (user interface): search, present, edit

Back-end (web services): select from databases, deliver in machine-readable format


```
"Lemma": {
      "FormRepresentation": {
        "feat": [
          {"att": "lemgram",
            "val": "huvuddrag..nn.1" },
          { "att": "partOfSpeech",
            "val": "nn" },
          { "att": "paradigm",
            "val": "nn_6n_bord" }, ]
      } },
```


Databases: access to corpora and lexicons

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Secret of KORP-2-KARP bi-directionality

User interface

Front-end

Web services:
*** reusable**
*** can be combined in any way**
*** allow communication between resources**

API

KARP web services

KORP web services

Resources

KARP-maintained lexicon databases

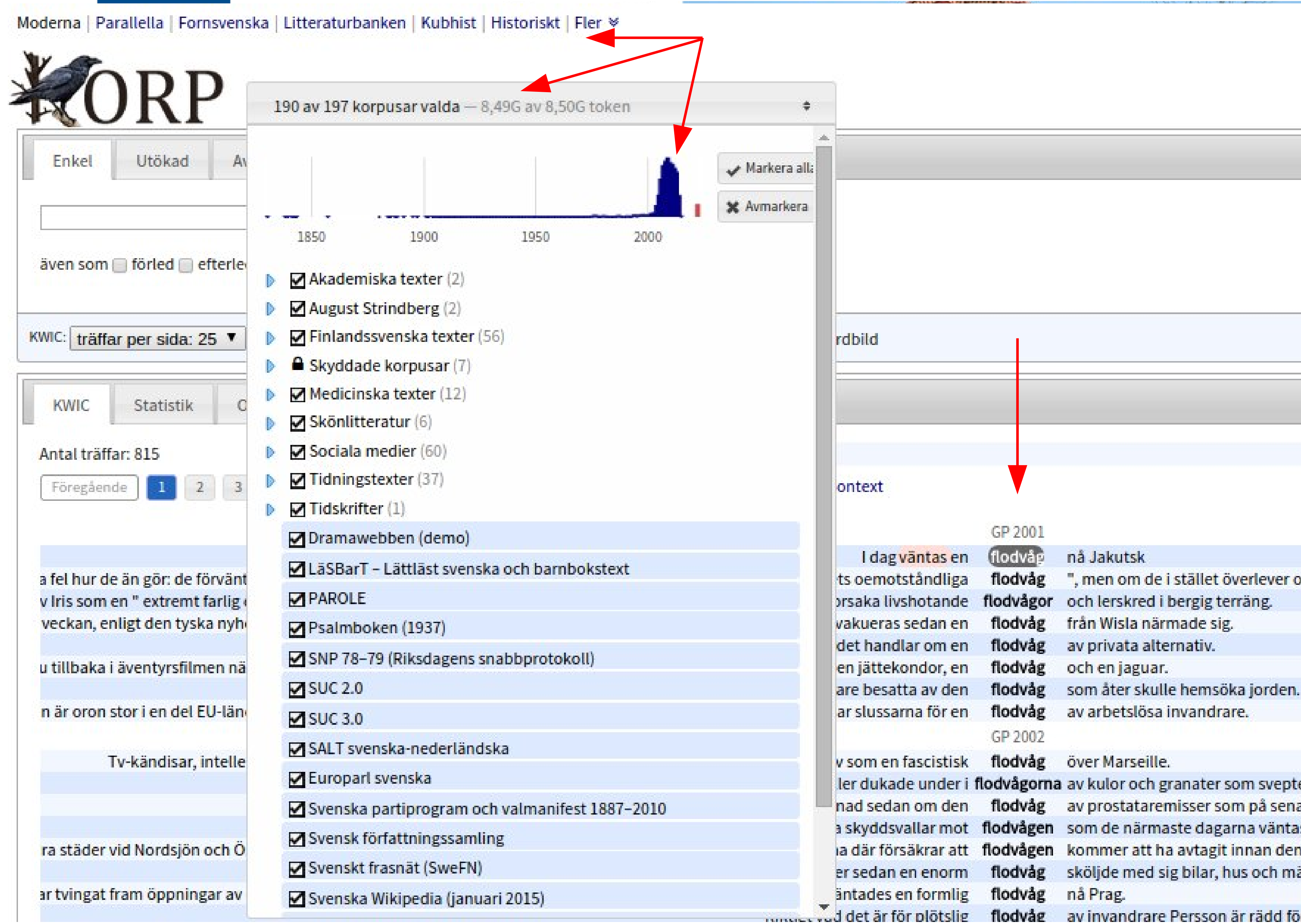
KORP-maintained corpora databases

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

KORP - the corpus infrastructure of SB

- **8,5 billion words** new and old texts – downloadable, scrambled
- **KWIC** vs **context** display
- Linguistic **annotation**: POS, lemgrams, msd, dependencies ...
- **Statistic** functions: pie charts, word lists, trend diagrams ...
- **Word pictures**
- **Save** or **bookmark** searches
- Different **views**: modern, parallel, Old Swedish ...



http://spraakbanken.gu.se/korp

Modern | Parallel | Old Swedish | Litteraturbanken | Kubhist | Old texts | Fler

Cite Korp Svenska | English 5

24 of 197 corpora selected — 321.94M of 8.50G tokens

Search history

Simple Extended Advanced Compare 2

ord (noun) Search

also as initial part final part and case-insensitive

Relaterade ord (SWE-FN)
ord, term, fackterm
glosa...

KWIC: hits per page: 25 | Sort within corpora: not sorted | Statistics: compile based on: lemgram | Show word picture

KWIC Statistics Word picture

Results: 88,127

Previous 1 2 3 4 5 6 7 8 9 10 11 .. 3525 3526 Next Show context

ASPAC SVENSKA (does not support extended context)

(Den här gången var hon glad att ingen kunde höra henne, för	**ordet**	lät inte riktigt rätt.
a fall på sig och började läsa, men hennes huvud var så fullt av hummerkadriljen att hon knappt visste vad hon sade, och	**orden**	blev verkligen ganska konstiga:
Jag har glömt	**orden**	.
Jag har hört vartenda	**ord**	ni sade.
Låt mig slippa höra det	**ordet**	mer!
Dessa	**ord**	följdes av en lång tystnad, endast avbruten av ett och
Hon urskilde	**orden**	: - Var är den andra stegen?

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

KWIC | Statistics | Word picture

Results: 88,127

Previous 1 2 3 4 5 6 7 8 9 10 11 .. 3525 3526 Next Show context

ASPAC SVENSKA (does not

(Den här gången var hon glad att ingen kunde höra henne, för **ordet** lät int
fall på sig och började läsa, men hennes huvud var så fullt av hummerkadriljen att hon knappt visste vad hon sade, och **orden** blev v
Jag har glömt **orden**
Jag har hört vartenda **ord** ni sad
Låt mig slippa höra det **ordet** mer!
Dessa **ord** följde
Hon urskilde **orden** : - Var
Men rösten lät hes och besynnerlig, och **orden** var in
Det sista **ordet** kom r
Hon sade det sista **ordet** ett pa
Hon sade de sista **orden** högt,
" och den fanns absolut inte här förut ", sade Alice) och vid flaskhalsen satt fastbunden en papperslapp, och på den var **orden** " DRI
(Alice hade ingen aning om vad latitud och longitud var för något, men hon tyckte det var roligt att säga långa, konstiga **ord** .
Jag begriper inte vad hälften av de där långa konstiga **orden** betyd
Sitt ned båda två, men säg inte ett **ord** förrär
Gripen svarade med nästan exakt samma **ord** som f
stora, tårdränkta ögon men inte sade ett **ord** .
Den börjar med **orden** " När
En del **ord** var nc
ckarn tjöt så att Alice knappt kunde höra **orden** :
iktigt, som om han velat ta reda på vilket **ord** som l
" för kålmasken och det blev alldeles fel **ord** .

Corpus

ASPAC svenska

Text attributes

title: Alice i Underlandet
author: Lewis Carrol
language: swe
description: Översättning av Harry Lundin

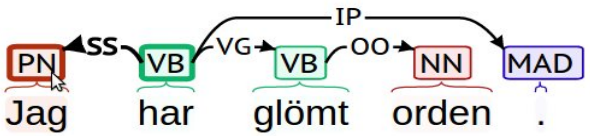
Word attributes

part-of-speech: noun
baseform:
ord
lemgram:
ord (noun)
sense:
ord
final part: *[empty]*
initial part: *[empty]*
dependency relation: Direct object
msd: NN.NEU.PLU.DEF.NOM

Show Dependency Tree

Dependency Tree

part-of-speech: pronoun



GÖTEBORGS UNIVERSITET

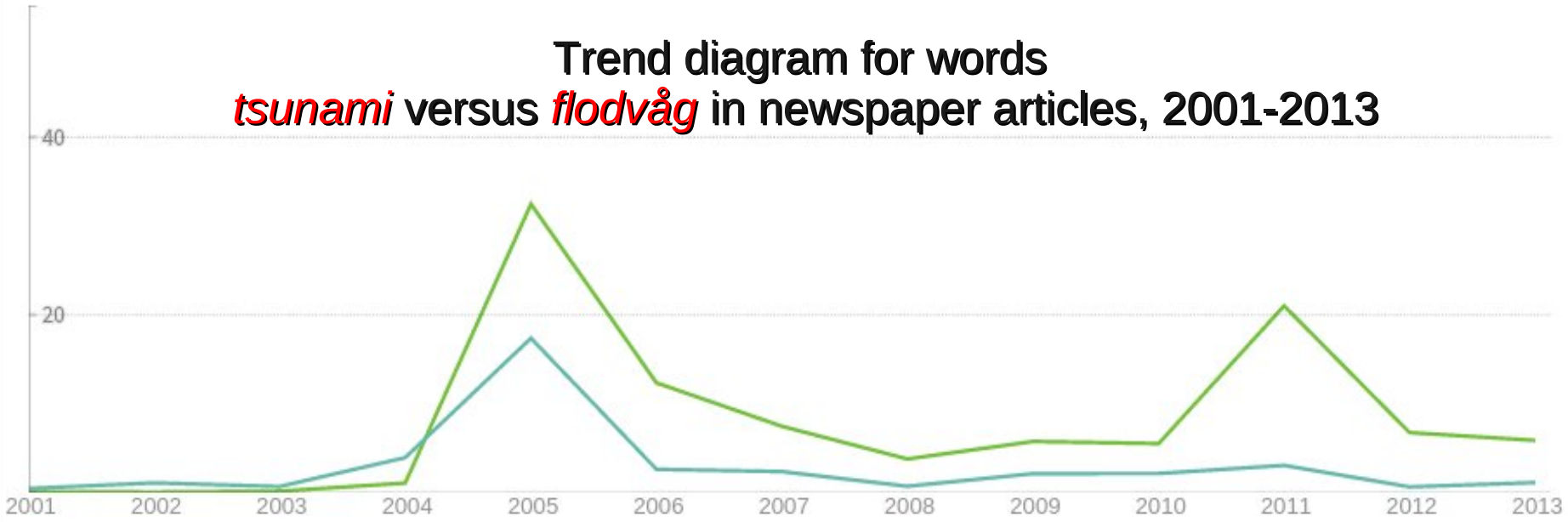
CLT Språk-BANKEN

KWIC | Statistics | Word picture | Graph

Line | Bar | Table

flodvåg (substantiv)
tsunami (substantiv)
Σ

Trend diagram for words *tsunami* versus *flodvåg* in newspaper articles, 2001-2013



GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Word picture for the noun *ord* (*word*):

Word's grammatical and collocational behavior
Based on MI, statistics from the selected corpora

ord (noun) Search

also as initial part final part and case-insensitive

Relaterade ord (SWE-FN)
ord, term, fackterm
glosa...

KWIC: hits per page: 25 | sort within corpora: not sorted | Statistics: compile based on: lemgram | Show word picture

KWIC | Statistics | Word picture

Show word class

Preposition		Pre-Modifier		ord	Post-Modifier		ord	Verb		ord	Verb	
1. med	9191	1. fri	2353		1. betyda	358		1. stå	506		1. använda	994
2. utan	494	2. ful	314		2. återkomma	57		2. betyda	242		2. skräda	256
3. bakom	235	3. vacker	376		3. beskriva	58		3. vara	2530		3. säga	1219
4. mot	484	4. hård	395		4. använda	80		4. använda	243		4. sätta	585
5. bortom	50	5. tom	208		5. som bäst beskriva	28		5. väga	126		5. höra	456
6. i brist på bättre	23	6. ont	96		6. i gp	62		6. väga[3]	104		6. uttala	180
7. mellan	125	7. vänlig	120		7. i ki-duks	20		7. väga tungt	62		7. yttra	141
8. ute efter	14	8. engelsk	179		8. av kärlek	30		8. komma	540		8. växla	175
9. trots	57	9. fin	229		9. för dag	41		9. säga	258		9. byta	273
10. om och om igen	6	10. klok	100		10. på guldvåg	20		10. höra	106		10. hitta	332
11. me	6	11. latinsk	59		11. i mun	34		11. räcka[2]	92		11. väga[3]	146
12. af	6	12. fel	111		12. kunna	111		12. räcka	92		12. väga	149
13. för övrigt	9	13. förflugen	37		13. dyka upp	32		13. yttra	34		13. få	804
14. till sist	7	14. ond	117					14. uttala	42		14. lära sig	140
15. till intet förpliktigande	3	15. grekisk	84					15. forsa	30		15. förstå	150

Modern | Parallel | Old Swedish | Litteraturbanken | Kubhist | Old texts | Fler

Cite Korp Svenska | English 5

KORP

37 of 38 corpora selected — 354.58M of 356.10M tokens

Search history

Extended

English

26 languages paired with Swedish

baseform is

sun Aa

or

Add lang Search

Parallel corpora view

KWIC: hits per page: 10 sort within corpora: not sorted Statistics: compile based on: word

KWIC Statistics

Results: 135

Previous 1 2 3 4 5 6 7 8 9 10 11 .. 13 14 Next Show context

ASPAC SVENSKA-ENGELSKA (does not support extended context)

er sat still just as she left her, leaning her head on her hand, watching the setting **sun** , and thinking of little Alice and all her wo

kvar där hon hade lämnat henne . Hon satt med huvudet lutat mot handen och tittade på solnedgången och tänkte på lilla Alic

They very soon came upon a Gryphon, lying fast asleep in the **sun** .

De kom snart till en grip , som låg och sov i solskenet .

- timidly at first - to push a charming little sprig inoffensively upward toward the **sun** .

Corpus

ASPAC svenska-engelska

Text attributes

title: Alice in Wonderland (1865)
author: Carroll, Lewis
language: eng
description: *[empty]*

Word attributes

KWIC: hits per page: 10 ▾ | sort within corpora: not sorted ▾ | Statistics: compile based on: word ▾

KWIC | Statistics

Results: 135

Previous 1 2 3 4 5 6 7 8 9 10 11 .. 13 14 Next

Show context

ASPAC SVENSKA-ENGELSKA (d

ter sat still just as she left her, leaning her head on her hand, watching the setting **sun** , anc
t kvar där hon hade lämnat henne . Hon satt med huvudet lutat mot handen och tittade på

They very soon came upon a Gryphon, lying fast asleep in the **sun**
De kom snart till en grip , som låg och sov i solskenet .

- timidly at first - to push a charming little sprig inoffensively upward toward the **sun** .
ett näpet och vänligt skott mot solen .

" I should like to see a sunset... Do me that kindness... Order the **sun** to s
- Jag skulle så gärna vilja se en solnedgång ... Var snäll och befall solen att

But if you tame me, it will be as if the **sun** cam
Om du tämjde mig skulle det lysa upp min tillva

Just so. Everybody knows that when it is noon in the United States the **sun** is se
Ja , sannerligen - när klockan är tolv i Amerika , går ju solen som alla vet ned i Fra

" And I was born at the same moment as the **sun** ... "
Och jag föddes just som solen ...

: hurt any more, he still had the delicious taste of cinnamon bun in his mouth, the **sun** was
längre , han hade fortfarande den härliga kanelbullesmaken kvar i mun , solen sken in gen

Corpus

ASPAC svenska-engelska

Text attributes

title: Alice in Wonderland (1865)
author: Carroll, Lewis
language: eng
description: *[empty]*

Word attributes

part-of-speech: NN
baseform: sun

Parallel corpora search results

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Svenska English

ORP Annotation Laboratory

Language of analysis: English

Load example: Example

```
1  Lexicography
2  Practical lex
3  Theoretical l
4  syntagmatic a
5  developing th
6  the needs for
7  the data inco
8  'metalexicogr
9  focuses on th
10 provide a des
11 dictionary or
12 compilation,
13 (relatively r
14 e.g. legal le
15 and following
16 dictionaries.
```

```
eparate but equally important groups:
or craft of compiling, writing and editing dictionaries.
holarly discipline of analyzing and describing the semantic,
ionships within the lexicon (vocabulary) of a language,
components and structures linking the data in dictionaries,
 in specific types of situations, and how users may best access
nd electronic dictionaries. This is sometimes referred to as
ed to lexicography is called a lexicographer. General lexicography
, use and evaluation of general dictionaries, i.e. dictionaries that
age in general use. Such a dictionary is usually called a general
uage for General Purpose). Specialized lexicography focuses on the design,
 specialized dictionaries, i.e. dictionaries that are devoted to a
guistic and factual elements of one or more specialist subject fields,
ctionary is usually called a specialized dictionary or LSP dictionary
lized dictionaries are either multi-field, single-field or sub-field
```

- Bulgarian
- Dutch
- English
- Estonian
- Finnish
- French
- German
- Italian
- Latin
- Polish
- Portuguese
- Russian
- Slovak
- Spanish
- Swedish
- Swedish-dev

Positional attribute g

word | pos | msd

Show advanced setti

Run!

<sentence id="en847b0-en80d1f"> [Show XML]

word	pos	msd	lemma
Lexicography	noun	NN	lexicography

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

KORP Annotation Laboratory

Hide advanced settings

Run!

Corpus name: untitled

Root tag **Structural attributes** +

Tag: text

Show Makefile | Show JSON Settings | XML

Show tags Run!

<sentence id="en847b0-en80d1f"> [Show XML]

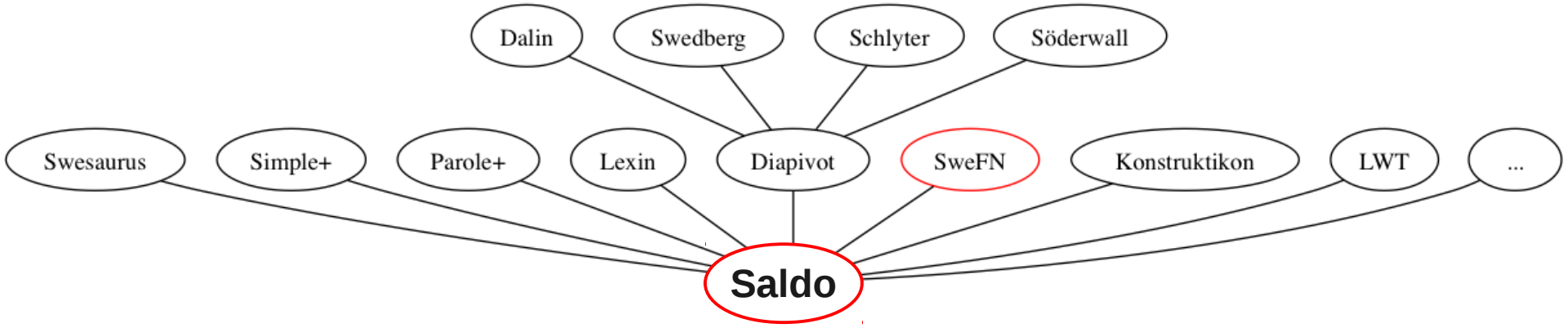
word	pos	msd	lemma
Lexicography	noun	NN	lexicography
is	verb	VBZ	be
divided	verb	VBN	divide
into	preposition / subordinating conjunction	IN	into
two	numeral	Z	2
separate	adjective	JJ	separate
but	coordinating conjunction	CC	but
equally	adverb	RB	equally
important	adjective	JJ	important
groups	noun	NNS	group
:	punctuation	Fd	:
Practical	adjective	JJ	practical

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

KARP - Lexical infrastructure of SB

Modern (13) **Historical** (9) **Diachronic pivot**



All resources (24) are **inter-linked** via Saldo, a Swedish association lexicon, that functions as a pivot for all lexicon resources.

Saldo contains around 130 000 entries, is organized by sense and each entry has a persistent identifier (PID).

Diapivot is a mapping between Saldo entries and entries (+ inflectional variants) in historical lexicons [SKOS principle; 20 names for *hens*; or "*bakvade*"-example]

http://spraakbanken.gu.se/karp

KARP

Log in | Download resources | Cite Karp | Svenska | English | About Karp

24 lexical resources selected (all) — 720,246 lexical entries

Search history

Simple | Extended | Editor

tsunami (substantiv) Search through Diapivot

Hits per page: 25

*"kan idag verkligen inte fatta att det var så stort att få sitt första barn att man missar hela **tsunami** katastrofen"*

Full Hits (3) | Statistics

Saldo | Saldos morfologi | Lexin

Saldo (1)

Sense	Lemgram	POS	Primary	Secondary	Children (primary)	Children (secondary)
tsunami	*tsunami (noun)* (21399)	noun	*flodvåg*	*förödande*	*tsunamikatastrof* –	*tsunamivarning* *varning*

Saldo morphology (1)

Lemgram	POS	Inherent	Paradigm	Morphology
tsunami (noun) (21399)	noun	u	nn_vu_safari	tsunami (nn_vu_safari) …

Lexin (1)

KARP

Log in | Download resources | Cite Karp | Svenska | English | About Karp

24 lexical resources selected (all) — 720,246 lexical entries

Search history

Simple | Extended | Editor

Search for entries that match

Phonetic form | equals

or

Sort by

Search

Hits p

- Wordform
- Baseform
- Msd
- Phonetic form
- Saldo Sense
- Lemgram
- Part of speech
- Primary descriptor
- Secondary descriptor
- Grammar note
- Text in definitions
- Text in examples
- Paradigm
- Cefr
- Form information
- Raw frequency
- Wpm
- Source
- Rank

Simple | Extended | Editor

Search for entries that match

Phonetic form | equals | çö:k

or

Sort by Wordform | ascending

Search

Hits per page: 25

Full Hits (3) | Statistics

Lexin (3)

Baseform	Phonetic form	POS	Rank	Lexin-ID	Morphology	Gram
kök	çö:k	noun	1637	1078455	kök ...	
translations						

rum där man lagar mat [Lexintema 10 11]

en lägenhet på två rum och kök
det kom härliga dofter från köket
köks|bord

KARP – editing

Password protected
Goal: to edit all resources

Name dubbelimperativ

Illustration Fortsätt ställ frågor i frågestunden HÄR!

Berkeley ID

Type
- konstruktion
- direktiv huvudsats

Category S

FrameNet

Definition
Konstruktionen omfattar [verb]$_V$ som inte kräver infinitivmärke i infinitivfras, i samordning.
Konstruktionen har vardaglig prägel. [Det inledande verbet]$_{Modifierarverb}$ har modifierande funktion.

Structure [$V_{1,imp}$ $V_{2,imp}$]

Inheritance
- pseudosamordning

Keywords

Common words
- fortsätta[1]
- sluta[1]

Construction elements, internal
- name=Modifierarverb, cat=V, msd=imp, other=Modifierarverb
- name=V, cat=V, msd=imp

Construction elements, external

Examples
- [[Fortsätt]$_{Modifierarverb}$ [sälj]$_V$]$_{dubbelimperativ}$ biljetter SJ – och sluta släng av barnen !
- [[Sluta]$_{Modifierarverb}$ [skriv]$_V$]$_{dubbelimperativ}$ så flummigt och redogör hur det verkligen ligger till , tack .

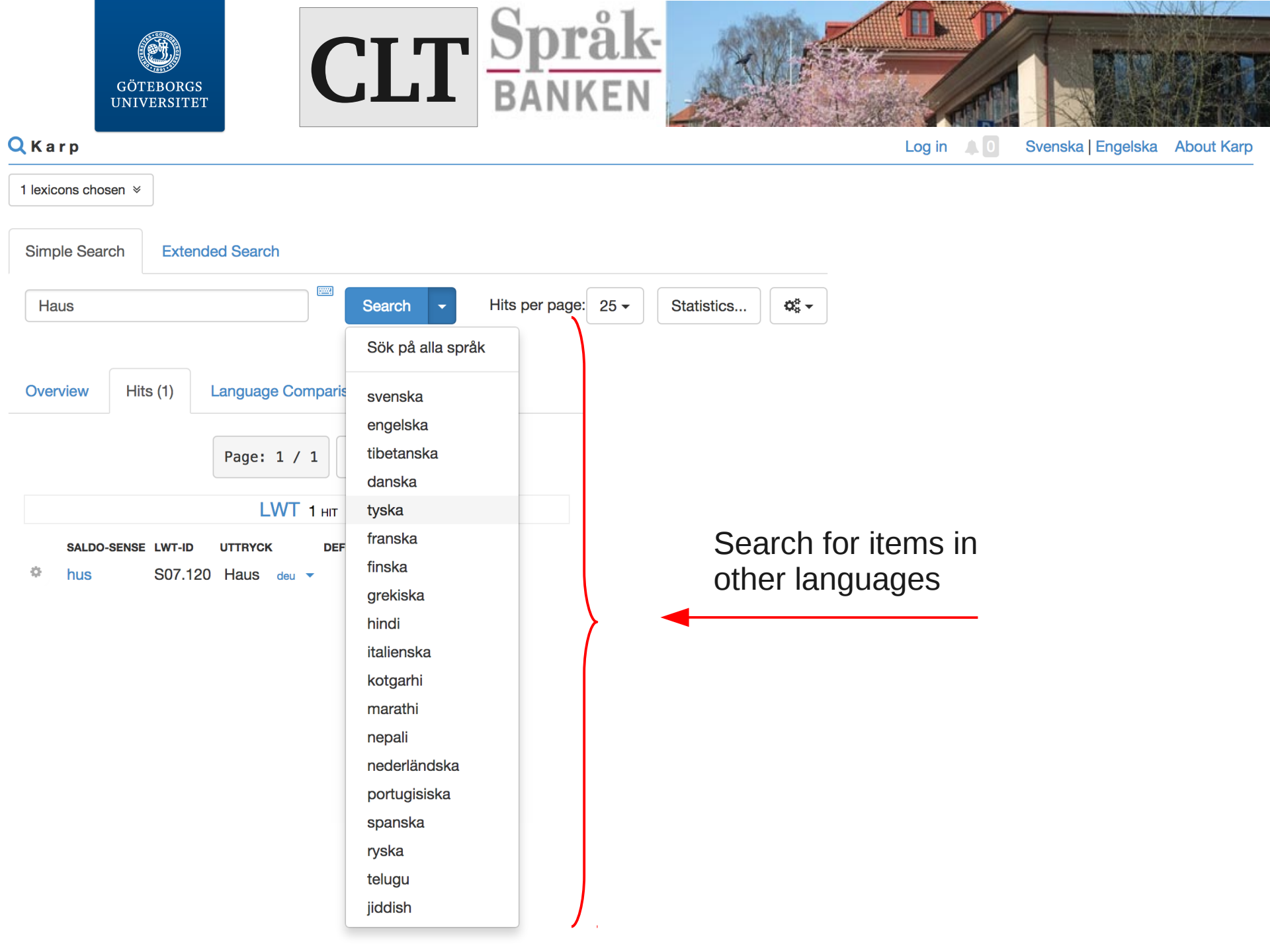
SALDO MORPHOLOGY 5 HITS

LEMGRAM	PART OF SPEECH	INHERENT	PARADIGM	BÖJNING
fisk (noun) 384421	noun	u	nn_2u_sten	fisk ...
fisk[2] (noun) 294213	noun	u	nn_0u_månsing	fisk ...
som fisk i vattnet (multiword adverb) 8	multiword adverb		abm_i_till_exempel	som fisk i vattnet (oböjl.)
som en fisk (multiword adverb) 3348	multiword adverb		abm_i_till_exempel	som en fisk (oböjl.)
som en fisk på torra land (multiword adverb) 135	multiword adverb		abm_i_till_exempel	som en fisk på torra land (oböjl.)

Number of lemgram occurrences in Korp

Statistics over form usage in Korp

sg indef nom	**fisk: 215**
sg indef gen	**fisks: 2**
sg def nom	**fisken: 84**
sg def gen	**fiskens: 3**
pl indef nom	**fiskar: 79**
pl indef gen	**fiskars: 0**
pl def nom	**fiskarna: 45**
pl def gen	**fiskarnas: 2**
ci	**fisk-: 1 fisk: 215**
cm	**fisk-: 1 fisk: 215**
sms	**fisk-: 1**


GÖTEBORGS UNIVERSITET
CLT
Språk-BANKEN
Karp
Log in
0
Svenska | Engelska
About Karp
1 lexicons chosen
Simple Search
Extended Search
Haus
Search
Hits per page:
25
Statistics...
Sök på alla språk
svenska
engelska
tibetanska
danska
tyska
franska
finska
grekiska
hindi
italienska
kotgarhi
marathi
nepali
nederländska
portugisiska
spanska
ryska
telugu
jiddish
Overview
Hits (1)
Language Comparis
Page: 1 / 1
LWT 1 HIT
SALDO-SENSE
LWT-ID
UTTRYCK
DEF
hus
S07.120
Haus
deu
Search for items in other languages

Karp

Log in 1 Svenska | Engelska About Karp

24 lexicons chosen

Simple Search Extended Search

hus Search

Hits per page: 25 Statistics...

Overview Hits (725) Language Comparison

svenska engelska tibetanska danska tyska franska finska grekiska hindi italienska kotgarhi marathi nepali nederländska portugisiska spanska ryska telugu jiddish

SWE	ENG	DEU	FRA	FIN	HIN	ITA	NLD	POR	SPA	RUS
hus	the house, house	Haus	maison	talo	घर, gʰar	casa	huis, huis	casa	casa	дом

Karp

Log in 0 Sver

24 lexicons chosen

Simple Search | Extended Search

Find entries where baseform equals fyra or fem + or

Add term

Search

Reset Hits per page: 25 Statistics...

Overview | Hits (65) | Language Comparison

svenska | engelska | tibetanska | danska | tyska | franska | finska | grekiska | hindi | italienska | kotgarhi | marathi | nepali | nederländska | portugisiska | spanska | ryska | telugu | jiddish

SWE	fyra	fem
ENG	four	five
BOD	bźi	lṅa
DEU	vier	fünf
FIN	neljä	viisi
NEP	cāra	pāṅca
RUS	четыре	пять

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Korp for Karp

- Examples for FrameNet
- Examples for search items
- Evidence for Constructicon
- Polysemy identification
- New vocabulary candidates
- Evidence for target learner vocabulary

UNIVERSITY OF GOTHENBURG Språk-BANKEN

Korp & Karp in Lärka

LÄR språket via KorpusAnalys

- Corpus examples
- Swe FrameNet
- Saldo morphology
- Lexin
- Kelly
- Wikipedia
- Wiktionary

research - Advanced English

research

▸ **n**

1. systematic investigation to establish facts
Hypernyms: investigation
Hyponyms: operations research, field work, marketing research, microscopy, probe, scientific research
2. inquiry, enquiry - a search for knowledge
"their pottery deserves more research than it has received"

UNIVERSITY OF GOTHENBURG Språk-BANKEN

SaldoM exercise: training inflection patterns

Lärka
LÄR språket via KorpusAnalys

Inflection paradigm (SaldoM as source)

Reference materials

e.g. Saldo morphology

Korp sentences

Language learners | Train inflections, multiple-choice | General purpose vocabulary | 1 of 6 learner levels selected

Fully automatic

self-study mode | test mode | timed test | Generate

Result Tracker

Exercise name	Correct/Total
Learners/Inflections, självstudier	1/2

Train inflections, multiple-choice

Choose an appropriate word form for the gap

TIPS

Nr	Sentence		Correct answer	Links
2	Ännu en skatteförvaltning har slarvat bort _____ .	Select a word		
1	Husfasaderna pryds av enorma **målningar** med tecknade figurer som motiv .	målningar	✓ målningar	

Select a word
deklaration
deklarationen
deklarationens
deklarationer
deklarationerna
deklarationers

Saldo morphology: *målning*

MÅLNING
lemgram: målning..nn.1;
part-of-speech: nn;
saldo-paradigm: nn_2u_mening;
inherent: u;

sg		
indef	målning	
nom		
sg		
indef	målnings	
gen		
sg		
def	målningen	
nom		
sg		
def	målningens	
gen		
pl		
indef	målningar	
nom		

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

KORP & KARP export

We are exporting our tools, if anyone is interested in using them for their corpora or lexica. They are free of charge, and we help with installation.

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Thank you!