



GÖTEBORGS UNIVERSITET  
INST FÖR SVENSKA SPRÅKET

**Språk-**  
**BANKEN**

# Language corpora: Territory or map?

or: What do we study when we study language corpora?

or: the benefits of hoarding

Lars Borin  
Språkbanken • Swe-Clarin  
University of Gothenburg, Sweden

17. rakenduslingvistika kevadkonverents • Tallinn, 19th April, 2018



**SWE-CLARIN**



- ▶ Språkbanken – the Swedish Language Bank
- ▶ “empirical NLP” and language corpora
- ▶ primary data ~ analytical resources in linguistics –  
territory ~ map
- ▶ what we can(not) do with language corpora
- ▶ summary/conclusion



## Språkbanken (at University of Gothenburg) – origins and history

- ~1970: the first Swedish text corpus: Press-65
- 1972: chair in natural language processing
- 1975: Språkbanken established
- 1984: undergraduate program in language technology (LT)
- 2000: GSLT (national graduate school in LT)
- 2004: the Swedish Literature Bank
- 2007: CLT (Centre for Language Technology) started
- 2008: language technology named a strategic research area of the university
- 2009: generous strategic funding for CLT (–2015)
- 2011: version 1 of Korp
- 2013: Korp flies out over the world
- 2014: coordinating node of SWE-CLARIN
- 2015: Språkbanken turns 40
- 2018: close to 15 billion words in Språkbanken
- 2018: National Språkbanken funded by SRC (VR) (–2024)



## what is Språkbanken?

- ▶ a national resource since 1975
- ▶ an R&D-unit in language technology
- ▶ open and free access to sophisticated (linguistic) search in digital, richly annotated language resources (written Swedish representing all historical periods and all genres):
  - ▶ text corpora (monolingual and parallel)
  - ▶ lexical resources (modern and historical, mono- and multilingual)
  - ▶ a language technology infrastructure
  - ▶ downloading of entire resources (under a CC-BY license)
- ▶ (but opportunistic wrt digitization)
- ▶ unique competence in the area of Swedish language resources

## who are our users?

- ▶ language technology researchers
- ▶ linguists and lexicographers working on Swedish
- ▶ educators and students
- ▶ the public





GÖTEBORGS UNIVERSITET  
INSTITUT FÖR SVENSKA SPRÅKET

Språkbanken –  
<https://spraakbanken.gu.se>

Språk-  
BANKEN

Språk-  
BANKEN



SWE-CLARIN

LISTEN | PÅ SVENSKA | A-Ö

Search språkbanken SEARCH

ABOUT US FAQ RESOURCES RESEARCH PUBLICATIONS PHD PROGRAM STAFF

## Språkbanken

Språkbanken (the Swedish Language Bank) is a nationally and internationally acknowledged research unit at the Department of Swedish, University of Gothenburg, established in 1975 in recognition of the groundbreaking corpus linguistic work initiated by Sture Allén. Our work focuses on language technology, in particular methodologies for handling the Swedish language, and the development of linguistic resources and tools for Swedish. These language resources are made available to researchers in language technology and other disciplines, as well as to the general public.

[Read more...](#)

[Språkbanken FAQ](#) with common questions and answers (in Swedish).

## Research

### SweFN++



The Swedish framenet project

### Culturomics



Towards a knowledge-based culturomics

### SweCcn

The Swedish construction project

## Resources



[User manual](#)

Search in the corpus collections

corpora count	461
token (total)	15 204 801 541
sentences (total)	1 060 099 070



[User manual](#)

Search in the lexical resources



Modern | Parallel | Old Swedish | Litteraturbanken | Kubhist | Old texts | More ▾

Log in Svenska | English 0 ⚙



125 of 236 corpora selected — 2.08G of 13.26G tokens

Språk-  
BANKEN



kontrast.nn.1 ▾

Simple Extended Advanced Compare 10

kontrast (noun)

Search ▾

☒ in order and also as ☐ initial part ☐ final part and ☐ case-insensitive

KWIC: hits per page: 25 ▾ sort within corpora: not sorted ▾ Statistics: compile based on: word ▾ ☒ Show statistics ☐ Show word picture

KWIC Statistics Word picture

Results: 21,488

« « 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... » » Go to page  of 860

Show context

ÅBO UNDERRÄTTELSER 2012

I " Stundom i dröm jag dröjer " blev det en pikant	<b>kontrast</b>	mellan den modernistiska texten och sångaren som komrade sig sjä
Konst speglar språkets	<b>kontraster</b>	skapas av spjutets vasshet och de små tyglapparna.
Det som jag på något sätt saknar är större	<b>kontraster</b>	, med de spänningsfält sådana förmår skapa, och kanske en tydligare i
ga, flinande turkgubbar och kuvade, medskyldiga kvinnor i	<b>kontrast</b>	till helyllensvenskarna Bosse och Barbro.
la så högt och aggressivt jag kunde där, bara för att skapa	<b>kontraster</b>	.
nom att kombinera trädet med ljusa pasteller skapar hon	<b>kontraster</b>	.
Estetiken på scen framstod ibland som en	<b>kontrast</b>	till de mer psykologiskt laddade kompositionerna, som till exempel A
Den stakar ut en framtid i bjärt	<b>kontrast</b>	mot vår gemensamma historia.
t tystnaden, fågelsången och lugnet utgör en fascinerande	<b>kontrast</b>	till det liv som fördes här.

## Corpus

Åbo Underrättelser 2012

## Text attributes

date: 2012-11-20

## Word attributes

sense:

- **kontrast**

compound lemmings:

- **kont** (noun) + **rast** (noun)
- Show more (2)

initial part:

kona (noun)  
kont (noun)



GÖTEBORGS UNIVERSITET  
INSTITUT FÖR SVENSKA SPRÅKET

Språkbanken –  
Karp

Språk-  
BANKEN



Svenska | English

Choose a resource below or use the standard Karp selection.



SALDO

Lexical resource for Swedish language technology.

Old Swedish

Schlyter and Söderwall

SweFN

Swedish Framenet

A Russian constructicon

Hellquist's Swedish etymology

By Rolf Hellquist, digitized from the first (1922) edition.

Dalín

Dalín's dictionary ~ 19th century Swedish

Constructicon

A database of Swedish constructions.

Svenskt kvinno biografiskt lexikon

Bliss

A symbol lexicon

Swedberg's Svensk ordabok

18th century dictionary



Sparv

*Språkbanken's annotation tool*

Språk-  
BANKEN



Language of analysis:

Swedish ▼

Load example:

Drama

Åtta sidor

Talbanken

Lasbart

Ikea

Exempelkorpus

Editor

Upload

☒ Plain text ☐ XML

```
1 Vem som helst kan väl bli bråkig och besvärlig, men inte hur som helst.  
2 Dagens ungdom är förfärlig!
```

☒ Lexical analysis ☒ Compound analysis ☒ Dependency analysis ☒ Sentiment analysis ☐ Named entity



Exercise Generator

Hit-Ex

Annotation editor

Text evaluation




Language Acquisition Reusing **Korp**

**Lärka** - "LÄR språket via **KorpusAnalys**" - with its English equivalent "Lark" (**L**anguage **A**cquisition **R**eusing **K**orp)  
- is a freely available online platform developed in **Språkbanken** for learning Swedish and Swedish linguistics.

LärkaLabb, the version of Lärka under active development, currently includes:

- exercises for linguists (e.g. parts of speech)
- exercises for learners (e.g. word guessing)
- Texteval, a text difficulty evaluation for Swedish as a second (or foreign) language
- Hit-Ex, a sentence selection tool
- Cefrit, an annotation editor





STRIX

Collection

RD - Matton (1,403)  
RD - Beträkande (1,191)  
RD - Protokoll (995)

Select from list (20 options)

Blingbring

usling (1,152)  
Rantropi (1,063)  
handelsövers (502)

Select from list (594 options)

LIX

0-3940

Nk

0-744

QVik

0-1700

VERA PLUM PLUM

swedish  
document type (RD)  
status  
organ  
party (SD)  
speaker  
recipient  
year  
party  
topic

In order

Found 6,354 documents.

[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [...](#)

**Invandring**  
**Invandring** Andelen invandrare av världens befolkning i vandrare globalt har ökat. **Invandring** eller immigration är en form tvä utrikes födda föräldrar. **Invandring** kan bland annat ske för exempelvis allvarlig sjukdom. **Invandring** kan också förädlas av skäl om begreppet även innefattar illegal **invandring**, det vill säga personer  
SWEDISH WIKIPEDIA

**Fr invandring**  
**Fr invandring** Argument mot fr **invandring** är oftast antingen kulturella eller faktorer som orsakas av fr **invandring** är mer än rättfärdigade av livskvaliteten i allmänhet. Fr **invandring** och krig Kaos i samband de facto leda till fr **invandring**. Den naturliga försöker att  
SWEDISH WIKIPEDIA

**Invandring till Kanada**  
**Invandring till Kanada** **Invandring till Kanada** är processen där  
SWEDISH WIKIPEDIA

**Invandring**  
Statistik m.fl. (v) **Invandring** Förslag till riksdagsbeslut Riksdagen tillkännager här idag hårda regler för **invandringen**, samtidigt som vi om vår mening bör reglerna för **invandring** till Sverige förenklas och krävs  
RD - NOTION

**Invandring till Finland**  
**Invandring till Finland** **Invandringen** till Finland är den process delar av Finlands historia har **invandringen** varit en viktig källa till befolkningstillväxt och kulturella förändringar. **Invandringens** ekonomiska, sociala och politiska  
SWEDISH WIKIPEDIA

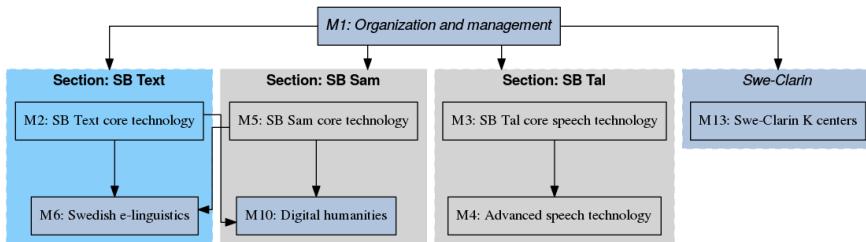
**Invandring**  
har idag hårda regler för **invandringen**, samtidigt som vi om min mening bör reglerna för **invandring** till Sverige förenklas och krävs  
RD - NOTION

**Fr information om invandringen**  
Fr information om **invandringen** Fr information om **invandringen** var en tidskrift utgiven av  
SWEDISH WIKIPEDIA

**Illegal invandring**  
**Illegal invandring** **Illegal invandring** refererar till företeelsen att människor bygger i praktiken på illegal **invandring**, eftersom ansökan om asyl längre. Med begreppet **illegal invandring** avses vanligen uppehåll i landet  
SWEDISH WIKIPEDIA



- ▶ *Språkbanken & Swe-Clarin (National Språkbanken – NSB)*  
funded by the Swedish Research Council 2018–2024
- ▶ three main sections:
  1. SB Text (= Språkbanken/U Gothenburg)
  2. SB Tal (SB Speech) (= KTH)
  3. SB Sam (SB Society) (= Institute of Language and Folklore)
- ▶ (plus 7 Swe-Clarin partners)







Types of multilingual resources in SB Text (all of course primarily for **LT**, but also designed/suitable for contrastive **(C)** or/and typological/areal/genealogical **(T)** linguistic studies) –

C: parallel (and comparable) corpora of the  
“traditional” kind

T: (massively multilingual corpora)

C/T: multilingual lexical resources

T: typological databases



T: interlinear glossed text (ITG)

Note that lexical and textual resources are normally **interlinked**, through tools for automated linguistic annotation.



# SB Text – parallel corpora

Modern | Parallel | Old Swedish | Litteraturbanken | Kubhist | Old texts | More ▾

Log in Svenska | English  



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



✓ Select all

✗ Select none

▾ ☒ Europarl3 (10)

- ☒ Europarl svenska-danska
- ☒ Europarl svenska-engelska
- ☒ Europarl svenska-finska
- ☒ Europarl svenska-franska
- ☒ Europarl svenska-grekiska
- ☒ Europarl svenska-italienska
- ☒ Europarl svenska-nederländska
- ☒ Europarl svenska-portugisiska
- ☒ Europarl svenska-spanska
- ☒ Europarl svenska-tyska

▾ ☒ SALT (1)

- ☒ SALT svenska-nederländska

▾ ☒ ASPAC (26)

-  The English-Swedish Parallel Corpus (ESPC)

14,023,276 sentences in the selected corpora

Språk-  
BANKEN



SWE-CLARIN

kontrast..nn.1 ▾

☒ Show statistics



## SB Text – multilingual lexical resources

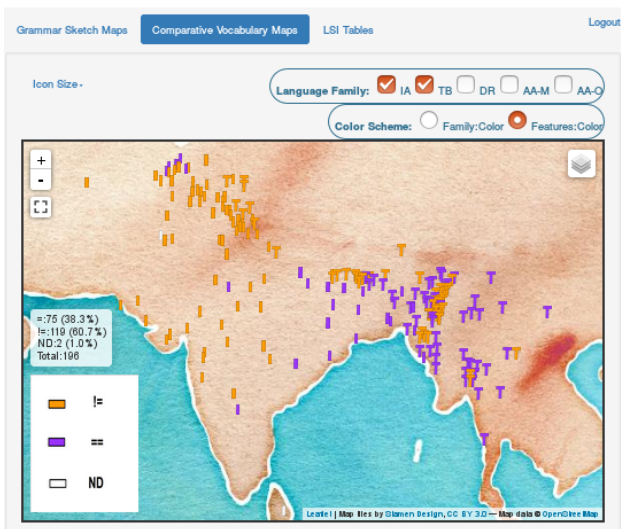
LWT ▾ 2 TRÄFFAR (VISAR 2)

SALDO- BETYDELSE	LWT-ID	UTTRYCK	DEFINITION	EXEMPEL
⚙ land <sup>2</sup>	S01.210	land <small>SWB ↕</small>	the hard surface of the earth, when compared to the area covered by sea	The captain sighted land in the distance.
⚙ jord	S01.212	jord <small>SWB ↕</small>	the substance that plants naturally grow in	The soil is pretty good in this area.
<div>engelska danska tyska grekiska franska italienska portugisiska spanska ryska nederländska jiddish tibetanska hindi kotgarhi nepali marathi telugu</div>				



## SB Text – typological databases

### South Asia as a Linguistic Area





chuk-hin-am, nga-rang-gi shak-po mu-la rang-thak che-pa. Daji kho thu-gu  
*wherefore, my friends with feast to-make. But that son*  
 chhungã yong-wa; kho-rang-su nor tshang-ma na-jung-la tang-wa-zin-song,  
*young came; him-by property all harlots-to to-give-finished,*  
 khe-rang-su kho-la za-ja thung-ja tang-we.' Kho-rang-su zer-wa, 'to  
*you-by him-to eating drinking gavest.' Him-by said, 'O*  
 nga-rang-gi thu-gu, khyot nga-rang-dang; da-rung chi hin-na nga-rang-gi  
*my son, thou me-with; and what is my*  
 nang-la thob-ong, kho khe-rang-la tshang-ma hin. Nga-rang-la do-chuk  
*house-in will-be-found, that thee-to all is. Us-to go*  
 kham-zang; khe-rang-gi no shi-sha-wa, tak-sang sanyo doi;  
*merry; your younger-brother dead-was, now alive went;*  
 tor song-wa hin, tak-sang thop-song.'  
*lost gone was, now found-was.'*



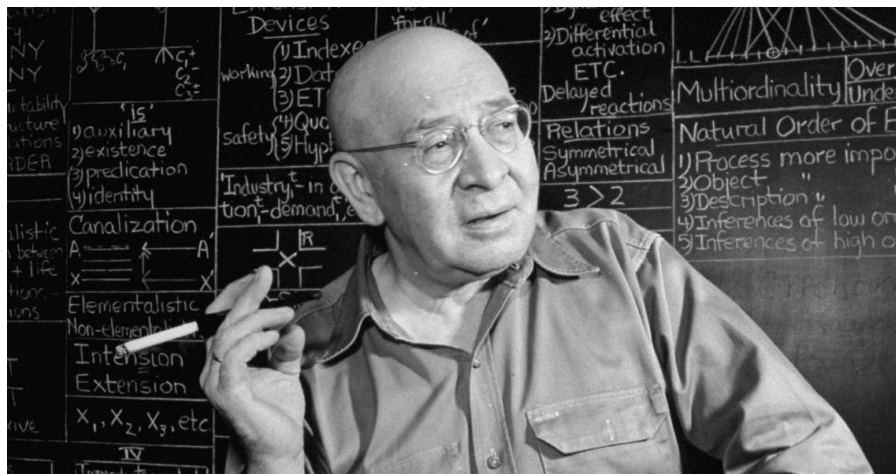
- ▶ in 2002, Språkbanken held <150 million words
- ▶ in the period 2003–2010, the corpus data grew by about 20 million words/year (basically one year's worth every year of GP, a Gothenburg-based newspaper)
- ▶ ⇒ in 2010, there were <250 million words in Språkbanken
- ▶ with Korp (released in 2011), we started collecting social-media text, first blogs, later online forums and tweets
- ▶ today, Språkbanken offers access to ~15 billion words (~13.5 BW modern, ~1.5 BW historical texts)
- ▶ ... but what can we do better with 100 times more text?



GÖTEBORGS UNIVERSITET  
INSTITUT FÖR SVENSKA SPRÅKET

"A map is not the territory it  
represents" (Alfred Korzybski 1933)

**Språk-**  
**BANKEN**



(From <<http://http://wheretimeturnsintospace.blogspot.com/2013/03/general-semantics-dune-and-whole-system.html>>)



## another mathematician on maps and territories:

"That's another thing we've learned from your Nation," said Mein Herr, "map-making. But we've carried it much further than you. What do you consider the largest map that would be really useful?"

"About six inches to the mile."

"Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"

"Have you used it much?" I enquired.

"It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well. (...)"

*Lewis Carroll: Sylvie and Bruno Concluded (1893)*





- ▶ Workshop on Very Large Corpora (WVLC): 1993, 1994, 1995, 1996, 1997, 1998
- ▶ Conference on Empirical Methods in Natural Language Processing (EMNLP): 1996, 1997, 1998
- ▶ Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: 1999, 2000
- ▶ EMNLP (occasionally joint with CoNLL or HLT): 2001–
- ▶ “Perhaps the most immediate reason for this empirical renaissance is the availability of massive quantities of data: text is available like never before” (Church 1999)

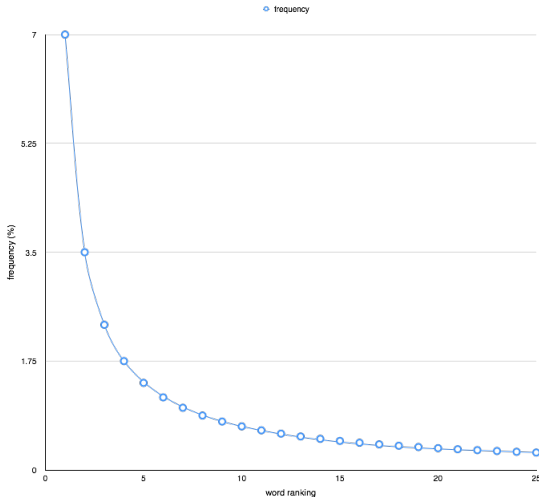


“Every time I fire a linguist, the performance of our speech recognition system goes up.” (c:a 1985)

(Frederick Jelinek (1932–2010), speech technology pioneer)



# Zipf's law and its consequences



Most linguistic phenomena are **very rare** (LNRE: Baayen 2001)



## "Deep learning"

```
226ms [[["Estonia",0.7507997155189514],["Latvia",0.6499463319778442],["Finland",0.6434240341186523],  
["Estonian",0.6195367574691772],["Lithuania",0.6182000637054443]]
```

If you don't get "queen" back, something went wrong and baby SkyNet cries.

Try more examples too: "he" is to "his" as "she" is to ?, "Berlin" is to "Germany" as "Paris" is to ? (click to fill in).

is to

as

is to

```
205.1ms [[["footwear",0.4086417853832245],["Head",0.3736826181411743],["heads",0.3688754439353943],  
["shoes",0.36462873220443726],["sandal",0.3555208444595337]]
```

If you don't get "queen" back, something went wrong and baby SkyNet cries.

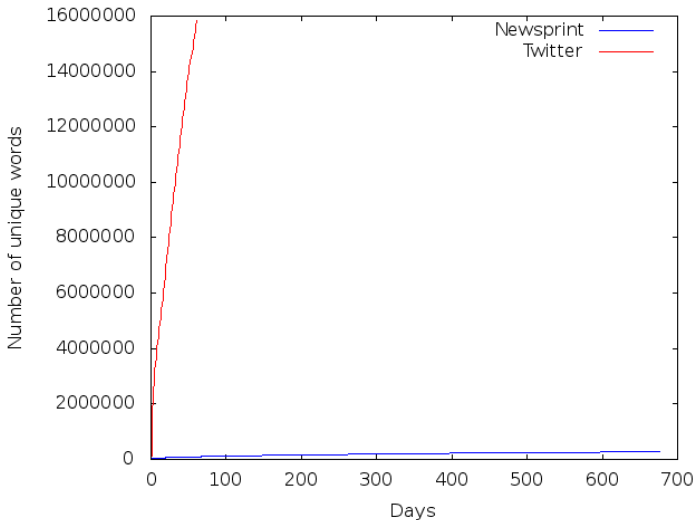
Try more examples too: "he" is to "his" as "she" is to ?, "Berlin" is to "Germany" as "Paris" is to ? (click to fill in).

is to

as

is to

(From <<http://rare-technologies.com/word2vec-tutorial>>,  
model trained on 100 billion words)



(Image source: Magnus Sahlgren, Gavagai, Inc.)



```
File Edit Options Buffers Tools Help
1 12,4°C
1 124:e
1 124hpåmig
1 1259kr
2 1.25AH
2 12.5°C
1 125ggr 01F 01F 01F 01F 01F
486 486 486 486 487 68A
1 1,25h
1 125km
1 12,5mkr
1 12,5st
1 12.67U
1 12,6°C
1 12.6°C
1 12-6h
1 12+6ish
1 126m
1 12,7°C
1 12.7°C
1 1.27gig
4 12.7k
1 127timmar
1 1280X720
6 12,8°C
1 12.8°C
1 128e
1 128GB
1 128k
1 12.8k
4 128mdr
2 1:29:10h
1 1299kr 01F
1 +12.9C
1 12,9°C
4 12.9°C
3 129kr
6 12a
1 1-2a
U:--- tpd-131006-nanal.txt 1% L1429 (Text)
```

```
File Edit Options Buffers Tools Help
68 Truuuuu
64 nuuuuu
37 huuuuuuuuuur
30 sjuuuuukt
27 duuuuu
25 nuuuuuu
21 kuuuuul
17 sjuuuuuukt
17 guuuuud
16 juuuuu
15 nuuuuuuu
15 guuuuuud
13 huuuuur
12 Nuuuuu
12 juuuuuu
11 nuuuuuuuu
11 kuuuuuul
10 shuuuuum
9 Huuuuur
9 duuuuuu
9 Buuuuuuteeeeeeee
8 huuuuuur
7 uuuuu
7 herreguuuuud
7 herreguuuuud
6 Åherreguuuuud
6 suuuuuger
6 sluuuuut
5 suuuuuper
5 Kuuuuul
5 Huuuuugh
5 Guuuuuud
5 duuuuuuu
5 azuuuuum
4 youuuuu
4 uuuuuuuuu
4 uuuuuuh
4 sluuuuuta
U:--- uuuuu-131006.txt Top L25 (Text)
```



- ▶ Himmelmann (1998 and elsewhere) introduces the notion of **language documentation** (or documentary linguistics),
- ▶ distinguishing it from descriptive linguistics:
- ▶ “language description aims at the record of *a language*”  
(analytical resources: ‘map’?: **≈written**)
- ▶ “language documentation, on the other hand, aims at the record of *the linguistic practices and traditions of a speech community*”  
(primary data: ‘territory’?: **≈spoken**)

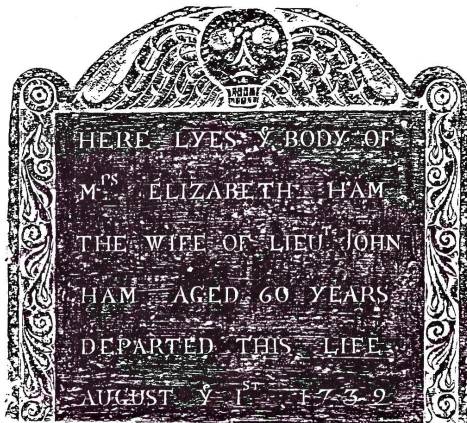


GÖTEBORGS UNIVERSITET  
INST FÖR SVENSKA SPRÅKET

## language corpora – a bit of both worlds

**Språk-**  
**BANKEN**

Somehow, language corpora combine aspects of map and territory, especially if they contain linguistic and other kinds of annotation. They are perhaps more like gravestone rubbings than like maps:



(Image from Dover N.H. Public Library website)



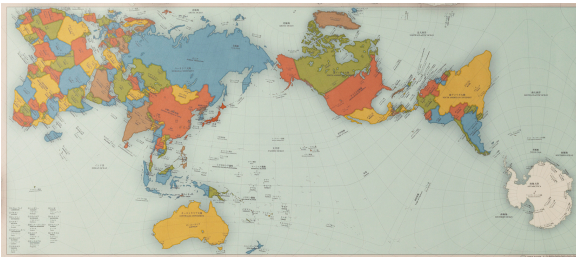
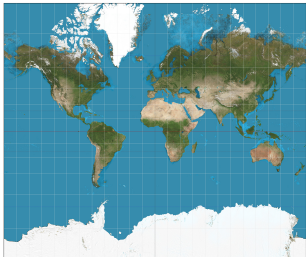


GÖTEBORGS UNIVERSITET  
INST FÖR SVENSKA SPRÅKET

so, just like maps, language  
corpora may distort the world

**Språk-**  
**BANKEN**

i.e., their **representativity** may be less than desired



(Images from: Wikimedia Commons (Mercator) • Alexcious (AutoGraph))



GÖTEBORGS UNIVERSITET  
INST FÖR SVENSKA SPRÅKET

they don't show us everything

**Språk-**  
**BANKEN**



© aultparksunrise.com 2011

(Photo from <<http://aultparksunrise.com/>>)



“Sometimes I do wish that the informants would be more careful in pronunciation and follow some system which would conform to theory. . . . Apparently no excuse, excepting that informants are too lazy to use it correctly.”

(Father Berard Haile in a letter to Edward Sapir  
30 March, 1931, quoted by Darnell 1990: 257)



and when it is avoided, there may  
be surprises

Technology Intelligence

## Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours



Microsoft's new teenage chat-bot CREDIT: TWITTER

(From <<https://www.telegraph.co.uk>>)



## what we can(not) expect to find out about language from corpora

- ▶ possibly (even likely):
  - ▶ many highly frequent formal aspects of written language (especially lexicon and phraseology)
  - ▶ second-language acquisition (primarily written language)
  - ▶ first-language acquisition (with directed documentation)
  - ▶ translation phenomena
  - ▶ typological phenomena
- ▶ not so possibly (even unlikely):
  - ▶ spoken language (especially dialog),
  - ▶ including multilingual interaction
  - ▶ rare linguistic phenomena (LNRE strikes again!)
  - ▶ semantic and pragmatic phenomena
- ▶ (... and always in a **data-hungry** way (remember deep learning!))



- ▶ language corpora combine the territory ("primary data") and map ("analytical resources") aspects:
- ▶ they may distort the object of study
- ▶ as **primary data**, they obey Zipf's law (meaning that hoarding is in fact a good thing in this context)
- ▶ as **analytical resources**, they may reflect (crypto-) normativity and preconceptions on what is to be found
- ▶ as **both**, they suffer from representativity issues wrt to the investigated domain/phenomena
- ▶ but still – perhaps exactly because of their double nature – they tend to beat linguistic intuition (and speculation) flat out



GÖTEBORGS UNIVERSITET  
INST FÖR SVENSKA SPRÅKET

Thank you for your attention!

**Språk-**  
**BANKEN**

And my deepest gratitude to our systems developers and researchers in Språkbanken who develop and maintain our language tools and resources!



(Foto: Boyd Michalovsky)