



GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

Språkteknologin till din tjänst

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

Leif-Jöran Olsson

Språkbanken, Göteborgs universitet, CLARIN ERIC, Swe-Clarin

2016-09-12



Översikt

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

- Översikt
- Visualisations
- Thematic examples
- Sociograms and relations
- Training and evaluation
- Final Comments

► Språkteknologin till din tjänst



Acknowledgements

GÖTEBORGS
UNIVERSITET

 **Språk-**
BANKEN

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

- ▶ For Swe-Clarin I gratefully acknowledge financial support from the Swedish Research Council 2014–2019.
- ▶ For parts relating to Dramawebben (The Swedish Drama Web) the Swedish Research Council (VR Dnr: 2011-6202) specifically.



Who are Swe-Clarin?

- ▶ Språkbanken, Göteborgs universitet
- ▶ SND (Svensk nationell datatjänst), Göteborgs universitet
- ▶ Digisam, Riksarkivet
- ▶ Tal, musik och hörsel, KTH
- ▶ Språkrådet, Institutet för språk och folkminnen
- ▶ CILTlab, Linköpings universitet
- ▶ Humanistlaboratoriet, Lunds universitet
- ▶ Datorlingvistik och teckenspråk, Stockholms universitet
- ▶ Datorlingvistik, Uppsala universitet



The Språkbanken context

GÖTEBORGS
UNIVERSITET



Our commonly mentioned infrastructures (geared toward linguists)

- ▶ Corpus infrastructure:
[<https://spraakbanken.gu.se/korp>](https://spraakbanken.gu.se/korp)
- ▶ Lexical infrastructure:
[<https://spraakbanken.gu.se/karp>](https://spraakbanken.gu.se/karp)

Översikt
Visualisations
Thematic examples
Sociograms and relations
Training and evaluation
Final Comments



Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

Korp – overview

Enkel Utökad Avancerad Jämförelse

Jura (verb) Sök ▾

även som förlad efterled och skiftlägesberoende

KWIC: träffar per sida: 25 ▾ sortera inom korpus på: förekomst ▾ Statistik: sammanställ på: ord

KWIC Statistik Ordbild Karta

Iura (verb)

Subjekt	Iura	Objekt	Adverbial
1. försök	17 ▾	1. folk	458 ▾
2. kvinna	25 ▾	2. hjärna	130 ▾
3. människa	25 ▾	3. kund	152 ▾
4. man	30 ▾	4. pengar	209 ▾
5. fara	8 ▾	5. peng	209 ▾
6. liga	8 ▾	6. stat ²	93 ▾
7. matfusket	6 ▾	7. stat	94 ▾
8. sätt	11 ▾	8. system	102 ▾
9. stat ²	10 ▾	9. människa	141 ▾
10. bedragare	7 ▾	9. under förespeglings	6 ▾
11. stat	10 ▾	10. väljare	84 ▾
12. n	9 ▾	10. ur arvsfond	4 ▾
13. sätt	7 ▾	11. konsumtent	69 ▾
14. läng	5 ▾	11. lång	20 ▾
15. läge	5 ▾	12. läng	10 ▾



The Språkbanken context (continued)

GÖTEBORGS
UNIVERSITET



Workflows and other services

- ▶ Annotation lab
- ▶ Mapservers, Mapserver and soon also Geoserver
- ▶ OCR lab (only for testing now)

For all available services and resources see
<https://spraakbanken.gu.se/>

Översikt
Visualisations
Thematic examples
Sociograms and relations
Training and evaluation
Final Comments



The Språkbanken context (continued 2)

GÖTEBORGS
UNIVERSITET



Sustainability and interoperability

- ▶ Open licenses on resources – data, tools and services
- ▶ Standard formats
- ▶ APIs, m2m, end user UI
- ▶ Support both upload to service and download and run locally
- ▶ Initialized process to become B-centre

For all available services and resources see
<https://spraakbanken.gu.se/>



The Swe-Clarin user

GÖTEBORGS
UNIVERSITET

 **Språk-**
BANKEN

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

- ▶ Workshops and specific meetups, to gather usecases/user stories, workflows and other essential information.
- ▶ Terminology alignment, important!
- ▶ Entities in time and space

A more documentcentric view in the infrastructures
– Strix.



The Swe-Clarin user stories to use cases

GÖTEBORGS
UNIVERSITET



Landing pages – FAQ + tutorial style.

- ▶ Introduction
- ▶ Hands on, on prepared material
- ▶ Further exploration
- ▶ Do it yourself, e.g. on the user's own material

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments



Blänkare för verktygslådan på Swe-Clarin-invigningen 7 oktober

Vi visar automatiska textanalysverktyg ur Swe-Clarins verktygslåda, för att till exempel:

- ▶ identifiera namn och platser
- ▶ generera ordbilder som visar ”ordens bästa vänner”
- ▶ visa tidslinjer över ords och uttrycks användning över tid
- ▶ klassificera texter utifrån olika egenskaper eller kriterier, t.ex. attityd eller tema
- ▶ hitta omnämningen av händelser av olika typer, till exempel ekonomiska transaktioner
- ▶ använda kartor för att visualisera geografiska samband eller förändringar över tid



Entrypoint example

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

- Översikt
- Visualisations
- Thematic examples
- Sociograms and relations
- Training and evaluation
- Final Comments

Entity Recognition

Tokenization

In order yo create a named entity recognition (NER) classifier model based on your own texts you need to tokenize and format the text document before you annotate it.

Upload your text in a wordprocessing document

Ingen fil är vald.



Entity markup example, obs fel bild

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

- [Översikt](#)
- [Visualisations](#)
- [Thematic examples](#)
- [Sociograms and relations](#)
- [Training and evaluation](#)
- [Final Comments](#)

Entity Recognition

Tokenization

In order yo create a named entity recognition (NER) classifier model based on your own texts you need to tokenize and format the text document before you annotate it.

Upload your text in a wordprocessing document

Ingen fil är vald.



Visualisations

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

We produce several visualizations automatically by reusable components and eXist-db apps.

- ▶ Focus on what information is important, not decorations
- ▶ From simple to more complex
- ▶ Only sometimes numbers and divisions are considered simple, so don't assume you can skip explaining

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

Speaker gender and speaker division

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Översikt

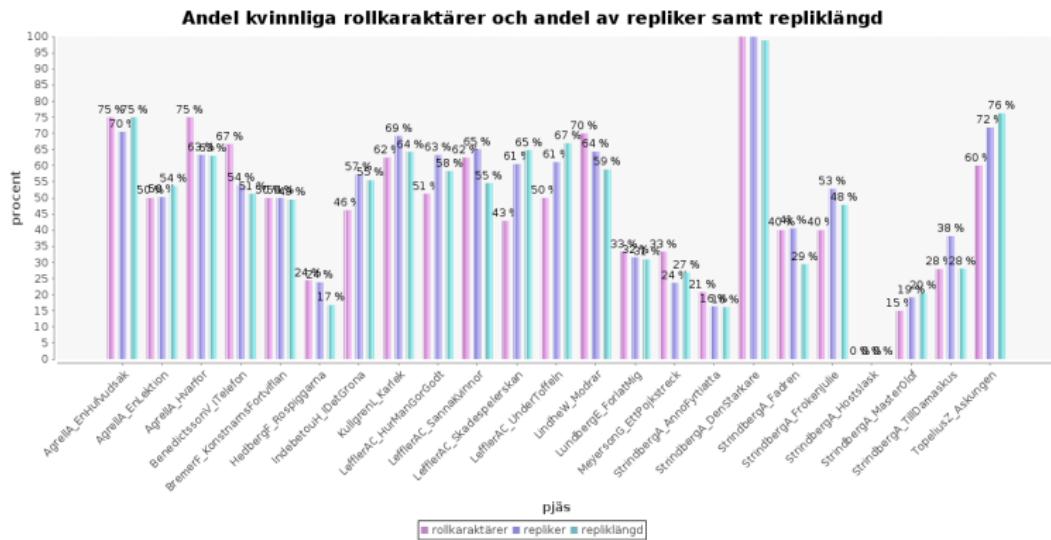
Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments



Speaker gender division per play

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Översikt

Visualisations

Thematic examples

Sociograms and relations

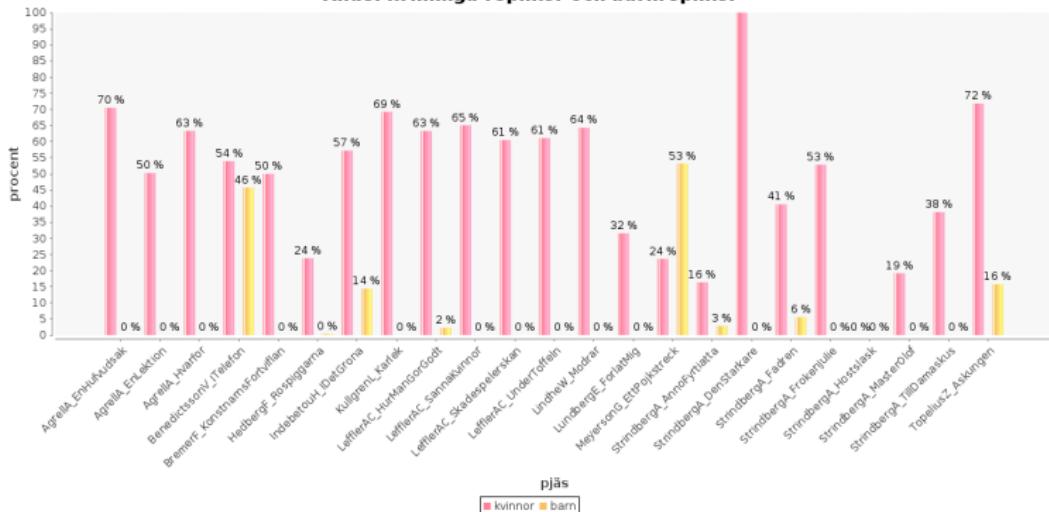
Training and evaluation

Final Comments



[Översikt](#)[Visualisations](#)[Thematic examples](#)[Sociograms and relations](#)[Training and evaluation](#)[Final Comments](#)

Female and child speeches division per play

Andel kvinnliga repliker och barnrepliker



Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

Result of searching for plays with a specific number of roles (excerpt)

Rollutforskning

Resultat för rollsökning i Dramawebben
Pjäser med mellan 10 och 15 rollkaraktärer

Pjäs



I det gröna (1896)
Indebetou, Hedvig



Skädespelerskan (1883)
Leffler, Anne Charlotte



Modrar (1887)
Lindhé, Wilma



Presence)

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Rollkaraktärer närvarande på scenen (entréer/sortier) StrindbergA_FrokenJulie



Översikt

Visualisations

Thematic examples

Sociograms and relations

Training and evaluation

Final Comments



Handicraft

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Översikt
Visualisations
Thematic examples
Sociograms and relations
Training and evaluation
Final Comments

As a thematic coding we used the example of textile handicraft since we believe that it can generate exciting issues and serve as an instructive example for other forms of semantic encoding.

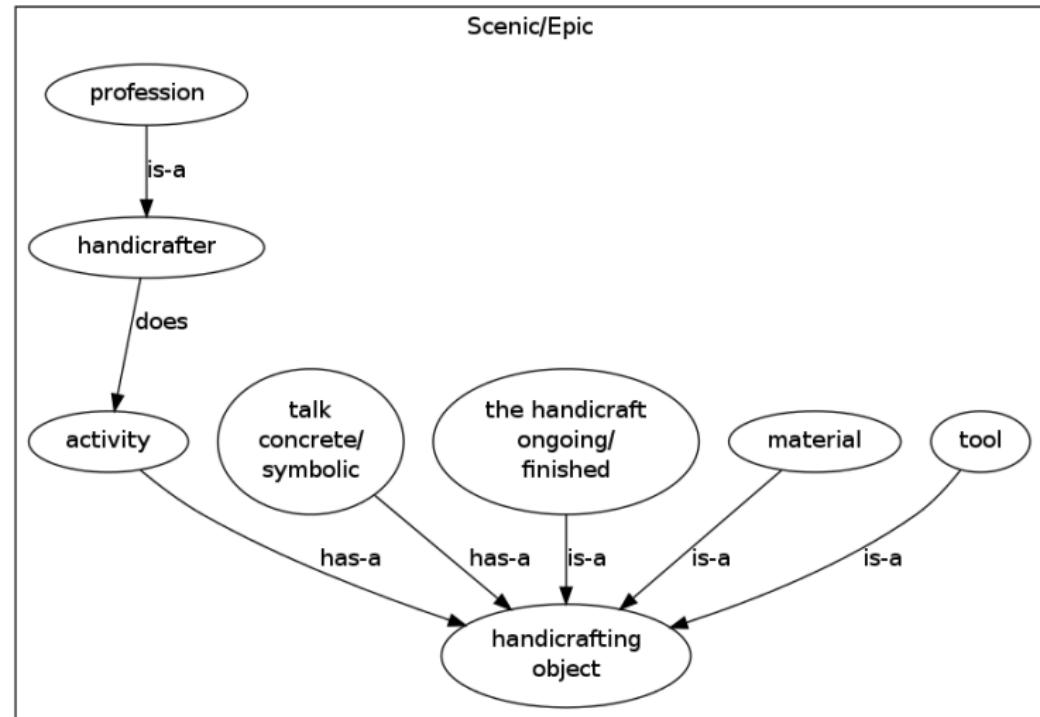
- ▶ Using feature structure elements, with key-value pairs
- ▶ They can be tied to anchors to make them discontinuous

The model (simplified)

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

Översikt
Visualisations
Thematic examples
Sociograms and relations
Training and evaluation
Final Comments

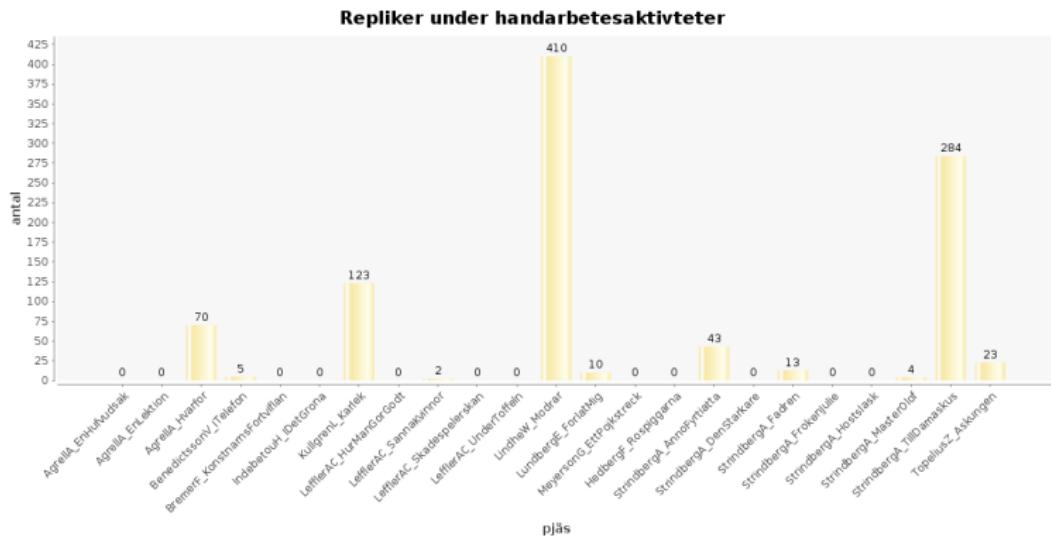


Ongoing handicraft in speeches per play

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

- Översikt
- Visualisations
- Thematic examples
- Sociograms and relations
- Training and evaluation
- Final Comments





Children's play and food and drink terms

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

For children's play we used the same basic parts of the model as for handicraft, that is:

- ▶ activity,
- ▶ talk about activity,
- ▶ and play objects

As an other example of potential thematic coding we extract food and drink terms. For this I created a simple hierarchical lexical resource with cooking and serving utensils, ingredients, dishes, procedures etcetera. These concept words were expanded morphologically by other lexical resources to cover all forms and some spelling variation over time.

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

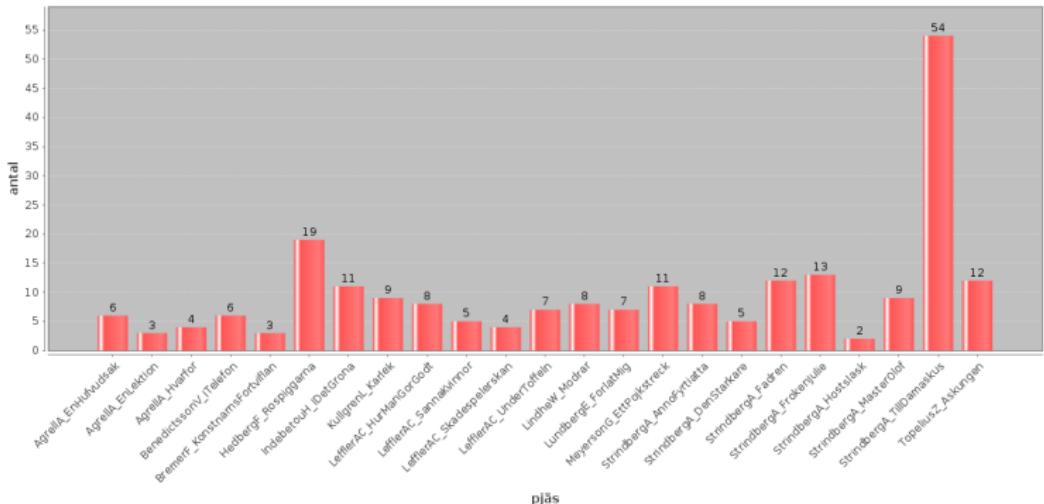
Food and drink terms per play

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

- Översikt
- Visualisations
- Thematic examples
- Sociograms and relations
- Training and evaluation
- Final Comments

Mat- och dryckestermner





Occupations

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

As an other example of thematic coding we use the Historical variant of the international standard classification of occupations (ISCO) called HISCO.

- ▶ 10 top level categories 1–0 and five levels of subcategories.
- ▶ The SCB also adapt/align its svensk standard för yrkesklassificering (SSYK) to the ISCO standard
- ▶ This makes it possible to compare occupations in an international context and link (LOD) to other datasets and implementations

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments



Occupations in play text (difference hisco 5-8, isco 0)

Översikt

Visualisations

Thematic examples

Sociograms and relations

Training and evaluation

Final Comments

Yrkesutforskning

Resultat för yrkesräkning i Dramawebben
Pjäser med yrken från isco-grupp 58 i pjästexten

 Hvarfor (1883)
Agrell, Alfhild

- God dag, kara Esther! Skulle titta in i forbigående och höra hur ni hafva det. God qvall, herr löjtnant .
- Menar? Hvem menar något nu för tiden. Man låter orden springa som en flock gass utan **vaktare**, får någon tag i dem är det tur.

[Visa fler träffar från denna pjäs \(1 st.\)](#)

 Rospiiggarna (1884)
Hedberg, Frans

- Nej, nu för tiden duger bara ungtupper — nu ska de ju vara laskarlar, vet jag! Det är bara generalerna, som får vara hur gamla de vill — inte gardisterna (ser uppåt). Hå hå, jaja! (gaspar). Skönt väder på morgonqvisten! (hastigt i det han tittar ut åt höger.) Jag tror fördubbla mina ransoner att der kom mer en poliskonstapel?
- Jo si, ångbåten ligger på slipen och jag har fått permissioner af **kapten** te fara hem och hälsa på under midsommar.

[Visa fler träffar från denna pjäs \(8 st.\)](#)

 I det gröna (1896)
Indebetou, Hedvig



Occupations of role characters

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

Yrkesutforskning

Resultat för yrkesökning i Dramawebben

Pjäser med yrken från isco-grupp 2' bland rollkaraktärerna



I telefon (1887)

Benedictsson, Victoria

yrke

hushållerska

rollkaraktär

Nilla



Hur man gör godt (1885)

Leffler, Anne Charlotte

yrke

verkstadsägare

rollkaraktär

statsråd

kapten Artur Wulf

kammarherre

statsråd

marskalk

Kammarherren

marskalkar



Sanna kvinnor (1883)

Leffler, Anne Charlotte

yrke

f.d. hushållerska

rollkaraktär

Lovisa



Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

Most common occupations in one selection for role characters

18	54020	arbetspiga
16	58320	amiralitetslöjtnantdotter
13	14120	adjunkt
10	14140	elisabetsyster
8	15120	f.d. författare
8	99900	arbetare
7	-1	allmosehjon
7	20210	andre legationssekreterare
7	17320	aktris
6	58220	biträdande vaktkonstapel
5	17120	kompositör
5	54010	f.d. dräng
5	55130	auktionsvaktmästare
5	20110	borgarråd
4	17140	dragspelerska



Link to images for occupation

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

Occupation title source	Kanslist
Occupation title (english)	Clerk, office
Hiscock	30000
Gender	Male
Title of the image	Överståthållarämbetets kansli, Slottsbacken 6. Interiör med t.h. Georg Henning, sittande t.v. Dahlborg.
Translation (English)	The Office of the Governor of Stockholm, Slottsbacken 6. Interior with Georg Henning (right) and Dahlborg (left).
Language source	Swedish
Country source	Sweden
Artist's name	Ahlén and Åkerlund
Date of creation	Beginning of the twentieth century
Medium	Photograph
Collection name	Stockholm City Museum
Provenance	Collection of photographs
Collection address	http://www.stadsmuseum.stockholm.se





Link to history of work DB

Occupational title	Adjunkt
Language	Swedish
Hisco code	<u>14120</u>
Provenance	<u>Demographic Data Base 1803-1900</u>
Translation	Teacher, secondary education, lutheran priest
Gender	Male
Country	Sweden

Search for Images



Sociograms and relations

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

- ▶ Sociograms
- ▶ Relations

We combine parts of the TEI namesdates module, like `<listPerson>` and `<listOrg>` with relations in `<listRelation>` elements to create graphs of relations between persons (cast and non-cast) or interaction on stage (cast only) sociograms.

Översikt
Visualisations
Thematic examples
Sociograms and relations
Training and evaluation
Final Comments



Sociograms

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

Översikt
Visualisations
Thematic examples
Sociograms and relations
Training and evaluation
Final Comments

- ▶ Sociograms are created dynamically and can be created based on any criteria of what constitutes interaction.
- ▶ These can also be weighted by giving a numeric value to the @sortKey attribute of the <relation> element.
- ▶ Of course you can also create other types of graphs based on dynamic data.

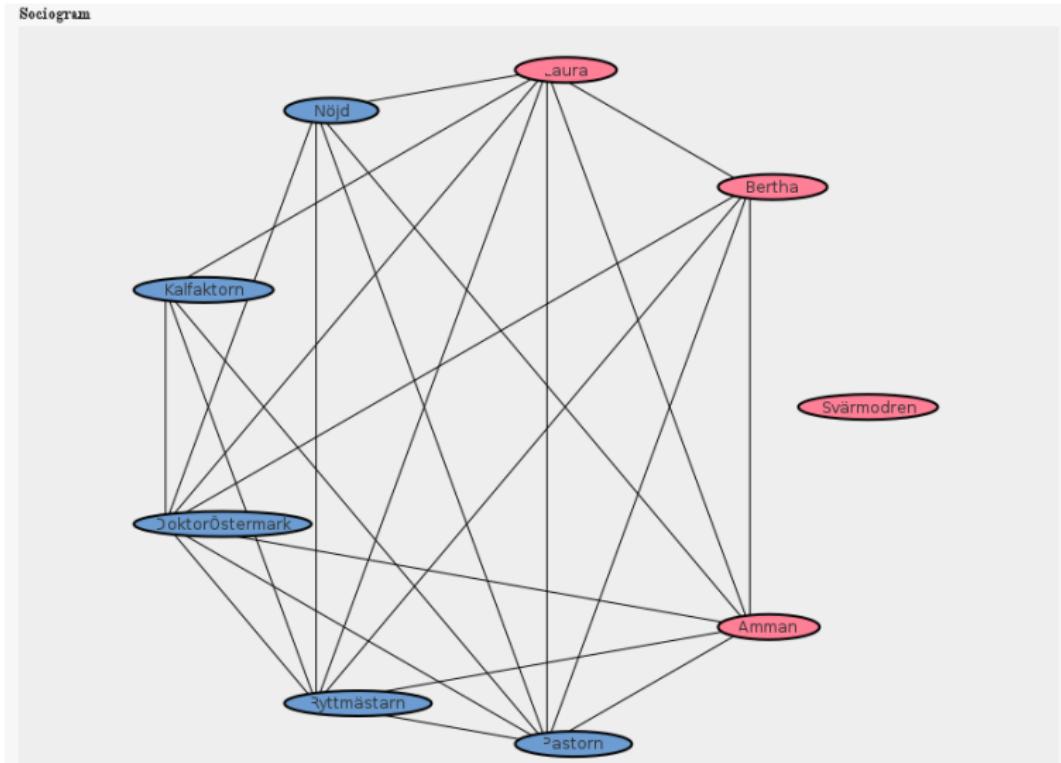


GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

- Översikt
- Visualisations
- Thematic examples
- Sociograms and relations
- Training and evaluation
- Final Comments

Sociogram for “The father”

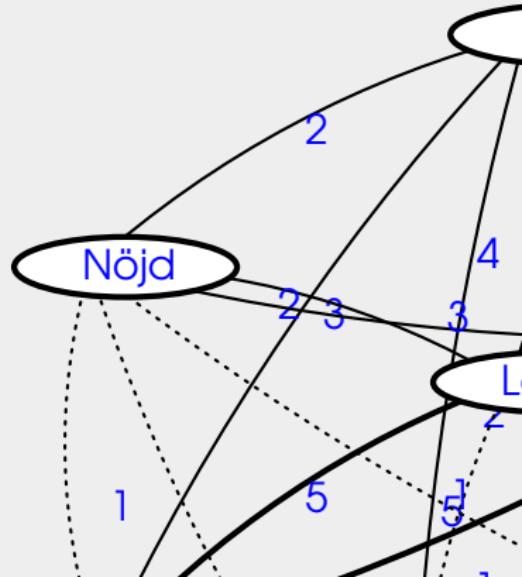




GÖTEBORGS
UNIVERSITET



- Översikt
- Visualisations
- Thematic examples
- Sociograms and relations
- Training and evaluation
- Final Comments



Person relations in "The Father"

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Översikt

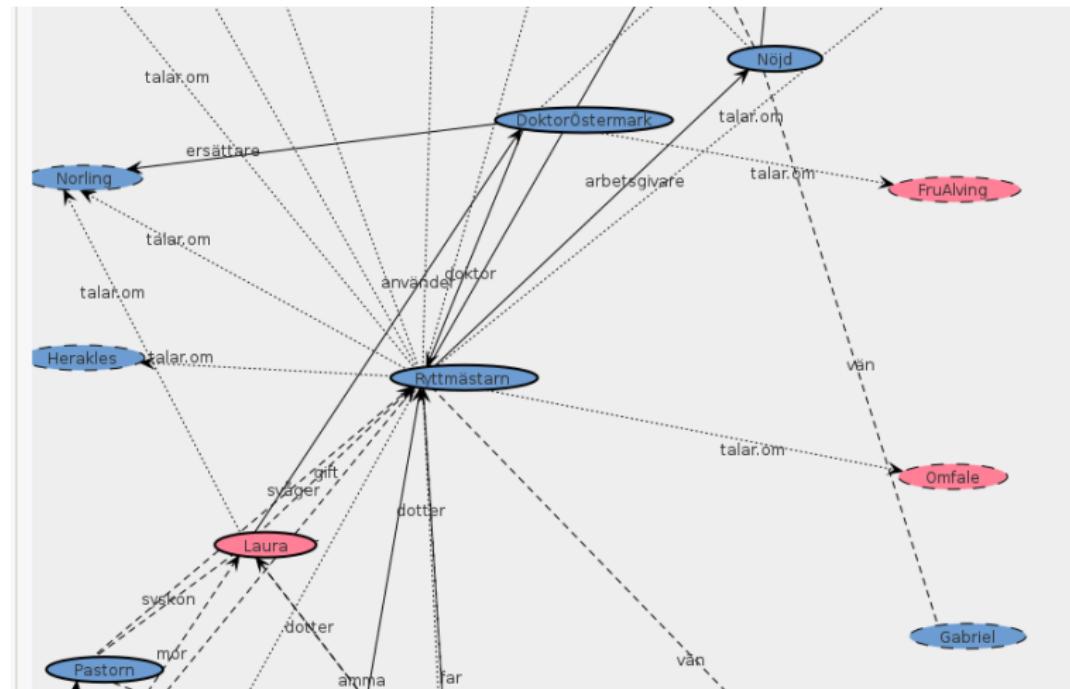
Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments





Multilingual document topics

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

- Översikt
- Visualisations
- Thematic examples
- Sociograms and relations
- Training and evaluation
- Final Comments

```
<topics:documentTopics xmlns:topics="http://exist-db.org/xquery/mallet-topic-modeling">  
    <topics:document n="0">  
        <topics:topic ref="4" weight="0.20588235557079315"/>  
        <topics:topic ref="3" weight="0.20588235557079315"/>  
        <topics:topic ref="1" weight="0.20588235557079315"/>  
        <topics:topic ref="0" weight="0.20588235557079315"/>  
        <topics:topic ref="2" weight="0.1764705926179886"/>  
    </topics:document>  
</topics:documentTopics>
```



Where did events occur?

GÖTEBORGS
UNIVERSITET

Swed-Clarins
Språk-
BANKEN

If we put the annotated places, persons or other subjects on a geographic map:

- ▶ Can we find hotspot areas to explore?
- ▶ Are there clear cut boundaries for some events or subjects?

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments



OCR

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

```
30     | který se poněkud uchyloval od obyčejné její klidnosti.</p>)[xs:int($lang-map?($lang))]  
31 return  
32 ocr:initialize-font(ocr:train-language-model-string($language-model, $text), $font)  
33  
34
```

▶ /db/temp/ocular-ocr-test.xquery

XMI Output ▾ Live Preview 
'db/apps/ocular-ocr/resources/classifiers/swe.fontser



OCR and pdf problems: Korp search 'garn' in Blekingeposten

Antal träffar: 358

« < 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 > »

[Visa kontext](#)

BLEKINGPOSTEN 1850-TALET

far jag härmmed rekommendera mig till emottagning af så wäl wäfwar som garn, och an< swarar undertecknad för äterftällandet som wilja lemno mig sitt förtroende, behagade inlemlna sätwäl wäfwar som garn t^ Släd » dare-Mästare herr E. IuhlinS gärd wid L sär>ag härmmed relommenvera mig till emottagning af sa wal wäfwar som garn, och an » swarar undertecknad för återställande om will » lemma mig filt förtroend », behagade inlem » sätwäl roäfrvar som garn i. Sriid » dare-Mästare herr li.

Sä mä de äldre utaf dig fatt lär » I trones kraft undwika syndens garn!

vafnadei af hampa och linne, ett större pam fara> stoftet, silar, teytdwaror, garn, beck, tiära och

r jag smickra mig ined — ingen hade under tiden fängslat mig i kår » lekens garn: jag hade warit likgiltig för allt h » vad qwinnor o ^onao A » derson från ^stra Hastbolmon, ogaro af ^, i Ckau och 2l » st. si » garn, lomua.

garn till yttersta priser;

Blekt och oblett samt färgabt Bomulls » garn ill yttersta priler;

H. Her« tigen af Ostcrögthland har warit på 2 ull » garn.

len; 3: ne rum och tambur, handkammare ed ' stira garn rober a ben cfr.i: 2 o <* tr af trn<i

'1<f<is » «n » rtiigge » sriwillig, >>7 « >d >>>< garn » , lon >>> slld<pl » t » n.

, h » va>sö:c inlöpts Invar » jebana rudimaticci, dcraf cn större qwantitet garn, träd, wäf, strumpor, toigar m. n>, bliswit tillverke r dcrjcmtc dcta ämne liggr framför honom säsom cn smutsig och oredig garn.

Blekning af wäfwar och garn sortsättas på Marielunds Vlcka sä fort » sf » för s

Mörkblått, mellanblått och Vjnoblatl, Boul: garn af flcrc nunimer; sicre sorier fina och ordinära fä af 8nrn, 0. s., s toifirr; Vattingar af Gran, 4 dnr; Olagpr af Nio, 13 bilo; daltar, garn, d.

" Alla goda ting äro tre " säger ett garn » malt ordöprät.

mier. Loimant llfzelius, har afgätt till Tull » garn, för att vara till H.

M. Drollningen å Tull » garn.

a^ rit stark, besiprkcs af fiskare, hwilka ge » nom densamma förlorade sina garn, och att yockan war swär är en känd sak.

>va, !,gl ^>lfpao » l ejortt sor da, d. Bo », ll " garn, dlell, odllll och färgall af fkre » l >imr, » ll dlllig

!, wilja lemma mig sitt förtroende, behagade <n » lemma sätwäl wäfwar som garn i Skräddare-Mästare Hr C. Inhlins gare wid l land » vilja lemma mig sitt föltloenle, behagade in » lemma sätwäl wäfwar son » garn i Skräddare-Mästare Hr it, Iuhlin gare wib lianbb



OCR and pdf problems: textflow 1

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

346 Bilaga 6

SOU 1975:67

Det finns likväl sådana fördelar med skriftliga anteckningar att man knappast kan gå förbi dem under praktiken. För det första har man sällan tid att omedelbart diskutera med annan personer vilka lakttagelser man gör eller de sätt man väljer att lösa ett problem på. Det gör att enklare minnesanteckningar behövs inför olika konferenser och inför praktikmetodiken. För det andra innebär anteckningarna att man måste formulerera problemet för sig på ett annat sätt än om det bara förblir tankar. För det trots ger anteckningarna möjligheter till en större systematik, vilket ibland är önskvärt.

Först och främst finns dessutom ett öppet loggbookssystem, i vilket hela arbetslaget kan göra anteckningar, vinnas ytterligare fördelar. Detta kan då innehålla spontanare anteckningar av vad arbetslagsmedlemmarna upplever som viktigt just därför. De kan också röra en avgränsad problemställning som man väljer att särskilt uppmärksamma t. ex. naturvetenskapliga experiment, konflikter och konfliktlösning etc. I loggbooken kan göra anteckningar inte bara kring direkta lakttagelser utan också protokoll från konferenser, värderingar av egna och andras val av handlingsalternativ etc.

Den huvudsakliga träningen undan praktiken kommer att nämnda att



OCR and pdf problems: textflow 2

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

348 Bilaga 6

SOU 1975:67

observationerna **atg** ingå som en naturlig del i kunskapsökandet i övrigt

- Beroende **på** **problemet** formulerar lärare och studerande tillsammans **inriktring** och **utbrmning** **av** **observationerna**. **Dågenom** **kan** **uppgifttemata** **nytta** **an** **till** **de** **studerandes** **tidigare** **erfarenheter** **och** **bäde** **under** **och** **efter** **utbildningen** **upplevas** **som** **meningsfulla**
- **Huvudlinje** **inriktningen** **för** **komplexitetsnivåen** **kör** **med** **att** **studens** **slit**

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments



- Översikt
- Visualisations
- Thematic examples
- Sociograms and relations
- Training and evaluation
- Final Comments

```
47 let $query2 := ("PREFIX myhouse: <myhouse://>
48 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
49 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
50 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
51 (: ("myhouse://table", "1", "myhouse://chair", "2") : )
52
53 let $query3 := ("PREFIX myhouse: <myhouse://>
54 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
55 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
56 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
57 (: ("myhouse://chair", "kitchen", "myhouse://table", "kitchen") : )
58
59 let $queries := ($query1, $query2, $query3)
60
61 return
62   for $query in $queries
63     return sparql:query($query)//sr:uri
64
65
```



Training and evaluation

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Översikt
Visualisations
Thematic examples
Sociograms and relations
Training and evaluation
Final Comments

- ▶ The freely available (open license) resources can be used for training and evaluation.
- ▶ Every project should contribute (could be as simple as review automatically annotated referential strings or some semantic encoding to timespecific complementary lexical resources).



Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

Named entity recognition (NER) based on dw-delkorpus1/dw-delkorpus2

```
5 let $lang := "swe"
6 let $lang-map := map {"eng": 1, "swe": 2, "ces": 3}
7 let $classifier := (xs:anyURI("/db/apps/stanford-ner/resources/classifiers/english.all.3class.distsim")
8 (: :xs:anyURI("/db/apps/stanford-ner/resources/classifiers/swe-outfile-300dpi-2class-model.ser.gz"),
9 xs:anyURI("/db/apps/stanford-ner/resources/classifiers/swe-dw12-3class-model.ser.gz"),
10 xs:anyURI("/db/apps/stanford-ner/resources/classifiers/ces-1872-outfile-djvu-2class-model.ser.gz"))[$lang])
11 let $text := (<p>The fate of Lehman Brothers, the beleaguered investment bank,
12 hung in the balance on Sunday as Federal Reserve officials and the leaders
13 of major financial institutions continued to gather in emergency meetings
14 trying to complete a plan to rescue the stricken bank. Several possible
15 plans emerged from the talks. held at the Federal Reserve Bank of New York
```

⇒ /db/temp/stanford-test.xquery

XML Output Live Preview

```
<p>När det gäller
<pers>Åke</pers>
och hans värld kommer vi långt ifrån bank- och finansväsendet.
<pers>Mimmi</pers>
och Sonja är i fjället på semester när lavinen går på
<org>Blåsjöjälet</org>
. Skärgårdens verklighet en vinterdag är inte heller så rosenskimrande. Herr
<pers>Arne</pers>
är dock i nya världen för gull och penningar. Ute på ön är oron stor för
<pers>Olagus</pers>
```



Final Comments

GÖTEBORGS
UNIVERSITET

 Språk-
BANKEN

Översikt

Visualisations

Thematic
examples

Sociograms and
relations

Training and
evaluation

Final Comments

- ▶ citations, digital and non-digital
- ▶ Resources available with open licenses:
[<https://spraakbanken.gu.se>](https://spraakbanken.gu.se)
[<https://www.dramawebben.se>](https://www.dramawebben.se)
- ▶ Tools used or mentioned with open licenses:
eXist-db apps
[<https://github.com/ljo/exist-tei-graphing>](https://github.com/ljo/exist-tei-graphing),
[<https://github.com/ljo/exist-sparql>](https://github.com/ljo/exist-sparql),
[<https://github.com/ljo/exist-mallet>](https://github.com/ljo/exist-mallet),
[<https://github.com/ljo/exist-ocular-ocr>](https://github.com/ljo/exist-ocular-ocr) and
[<https://github.com/eXist-db/jfreechart>](https://github.com/eXist-db/jfreechart), more
under [<https://github.com/ljo/>](https://github.com/ljo/) and
[<https://github.com/eXist-db/>](https://github.com/eXist-db/)
- ▶ Graphs can be used in svg, graphml and gexf
output