GÖTEBORGS UNIVERSITET

Språkteknologin till din tjänst

Leif-Jöran Olsson

Språkbanken, Göteborgs universitet, CLARIN ERIC, Swe-Clarin

2016-09-12

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

Översikt

- Språkteknologin till din tjänst

GÖTEBORGS UNIVERSITET

Språk-BANKEN

Acknowledgements

- 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.

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

The Språkbanken context

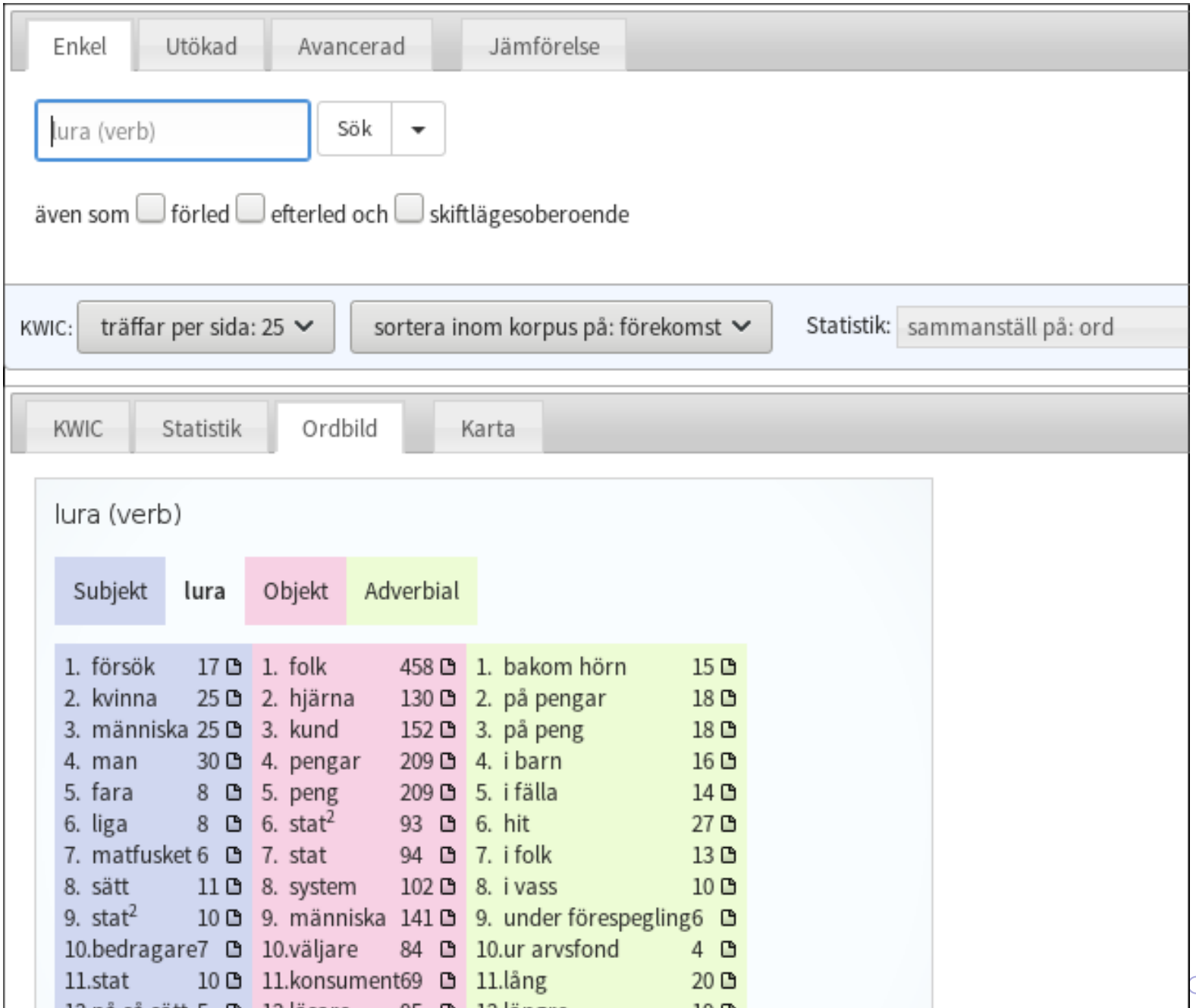
Our commonly mentionend infrastructures (geared toward linguists)

- Corpus infrastructure: <https://spraakbanken.gu.se/korp>
- Lexical infrastructure: <https://spraakbanken.gu.se/karp>

Korp – overview

GÖTEBORGS UNIVERSITET

Språkbanken (SWE-CLARIN)



GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

The Språkbanken context (continued)

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/>

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

The Språkbanken context (continued 2)

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/>

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

The Swe-Clarin user

- 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.

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

The Swe-Clarin user stories to use cases

Landing pages – FAQ + tutorial style.

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

Bläkare för verktygslådan på Swe-Clarin-invigningen 7 oktober

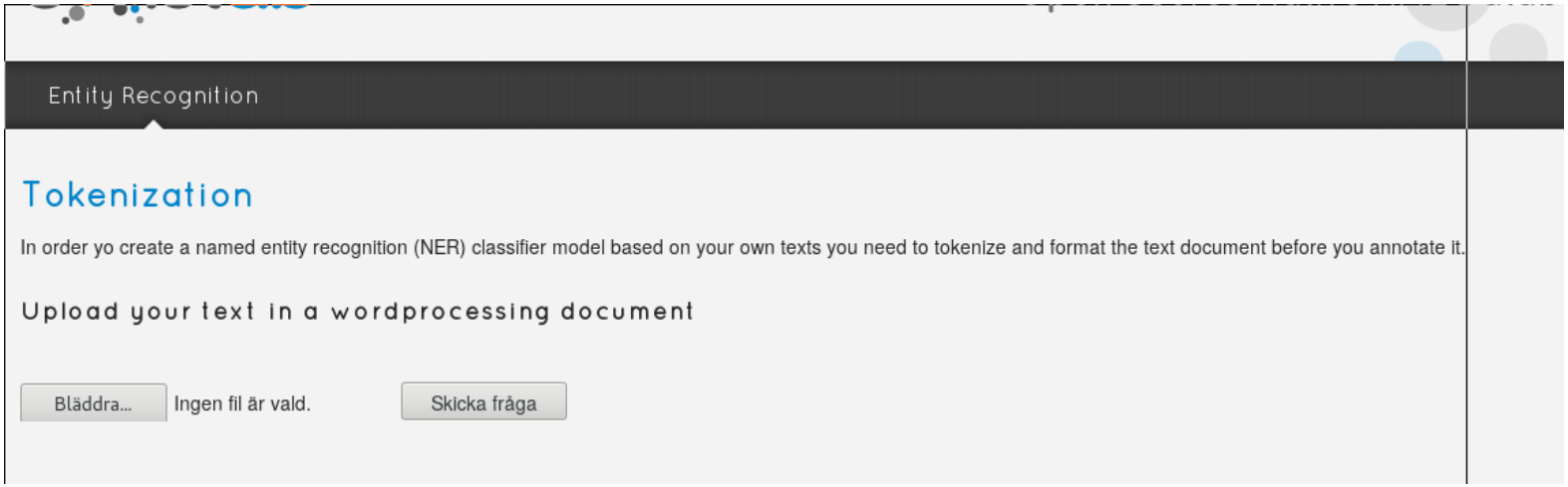
Vi visar automatiska textanalysverktyg ur Swe-Clarins verktygslåda, för att till exempel:

- 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 omnämnanden av händelser av olika typer, till exempel ekonomiska transaktioner
- använda kartor för att visualisera geografiska samband eller förändringar över tid

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

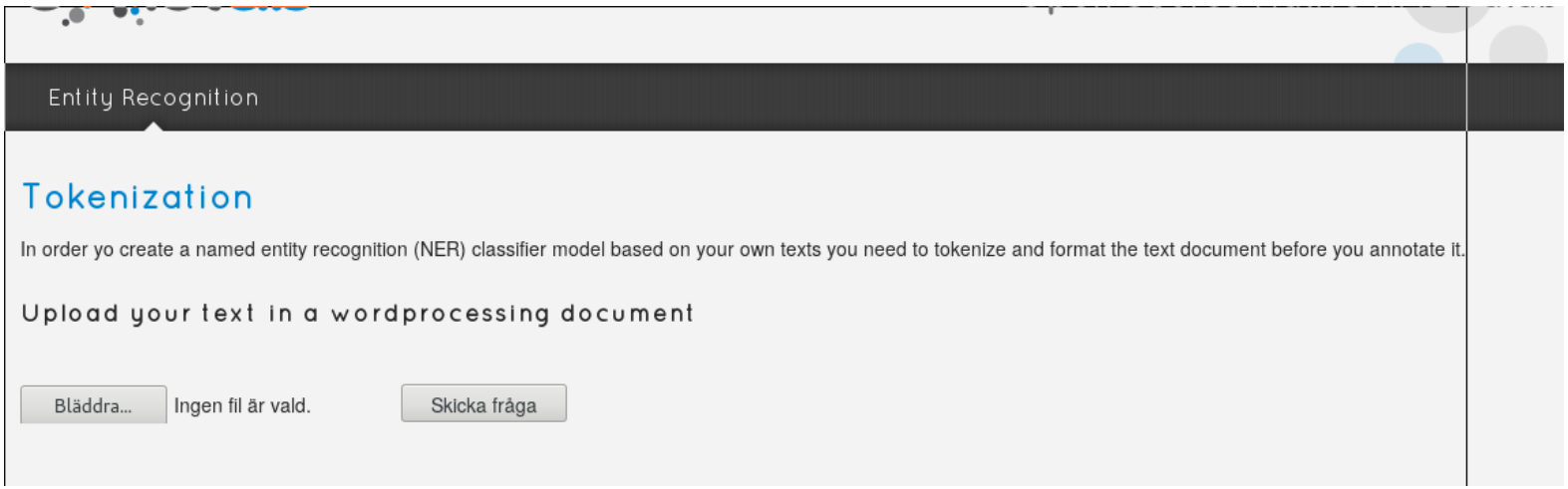
Entrypoint example



GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

Entity markup example, obs fel bild



GÖTEBORGS UNIVERSITET

Språk-BANKEN

Visualisations

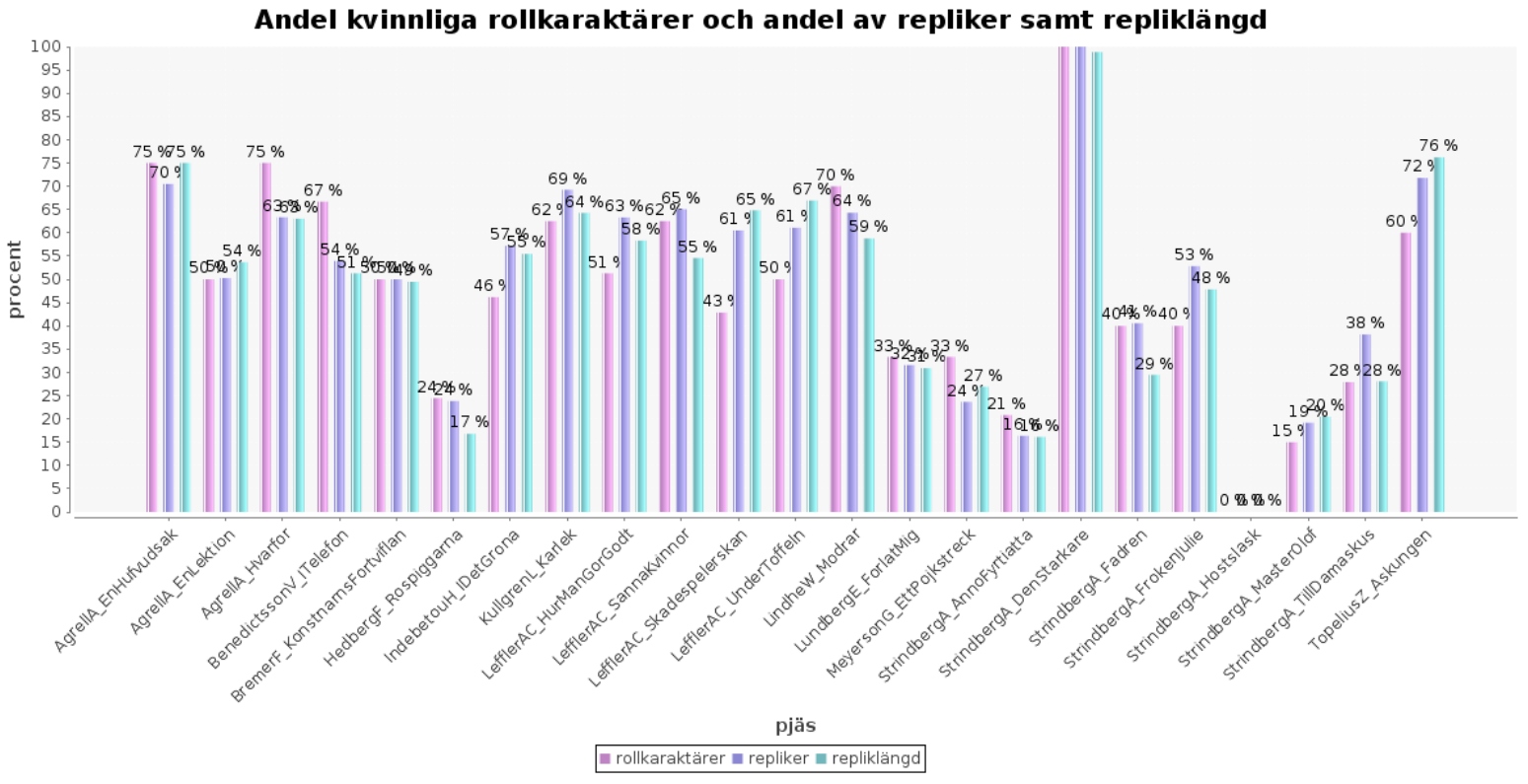
We produce several visualizations automatically by reusable components and eXist-db apps.

- 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

GÖTEBORGS UNIVERSITET

Språkbanken

Speaker gender and speaker division



GÖTEBORGS UNIVERSITET

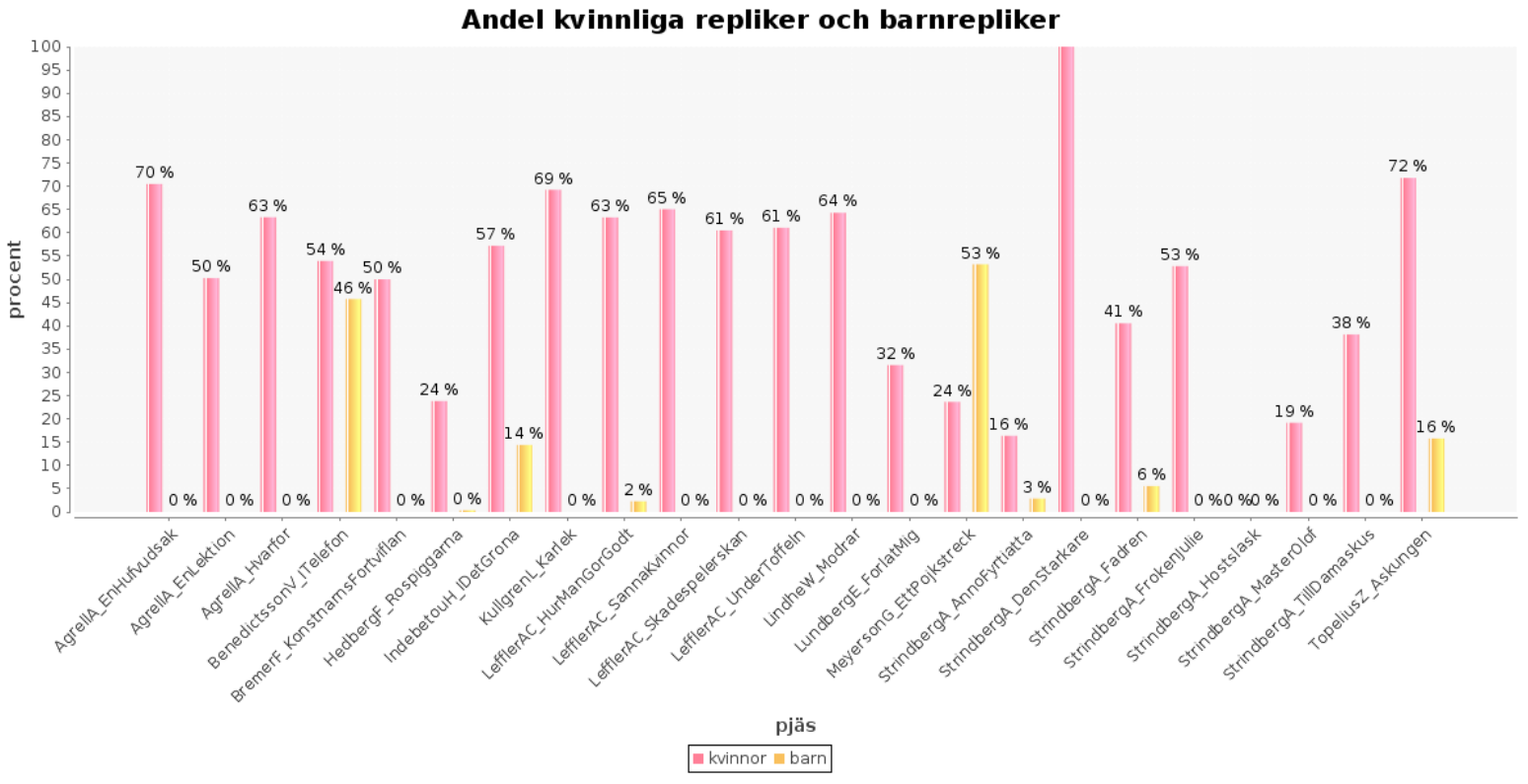
SWE-CLARIN Språk-BANKEN

Speaker gender division per play

GÖTEBORGS UNIVERSITET

Språkbanken

Female and child speeches division per play



GÖTEBORGS UNIVERSITET

Språkbanken

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

Rollutforskning

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

Pjäs

I det gröna (1896)
Indebetou, Hedvig

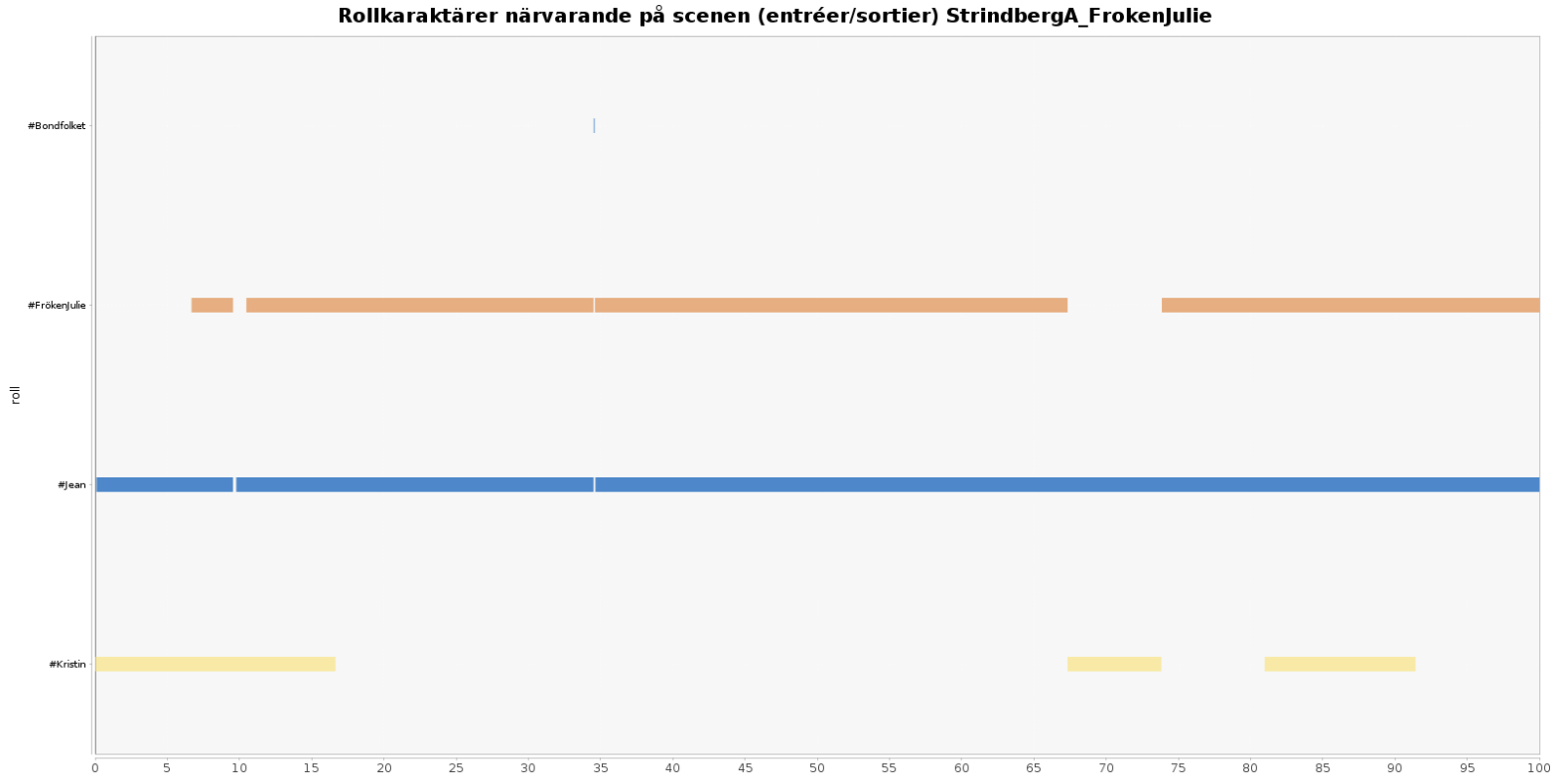
Skådespelerskan (1883)
Leffler, Anne Charlotte

Modrar (1887)
Lindhé, Wilma

GÖTEBORGS UNIVERSITET

Språkbanken

Presence)



GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

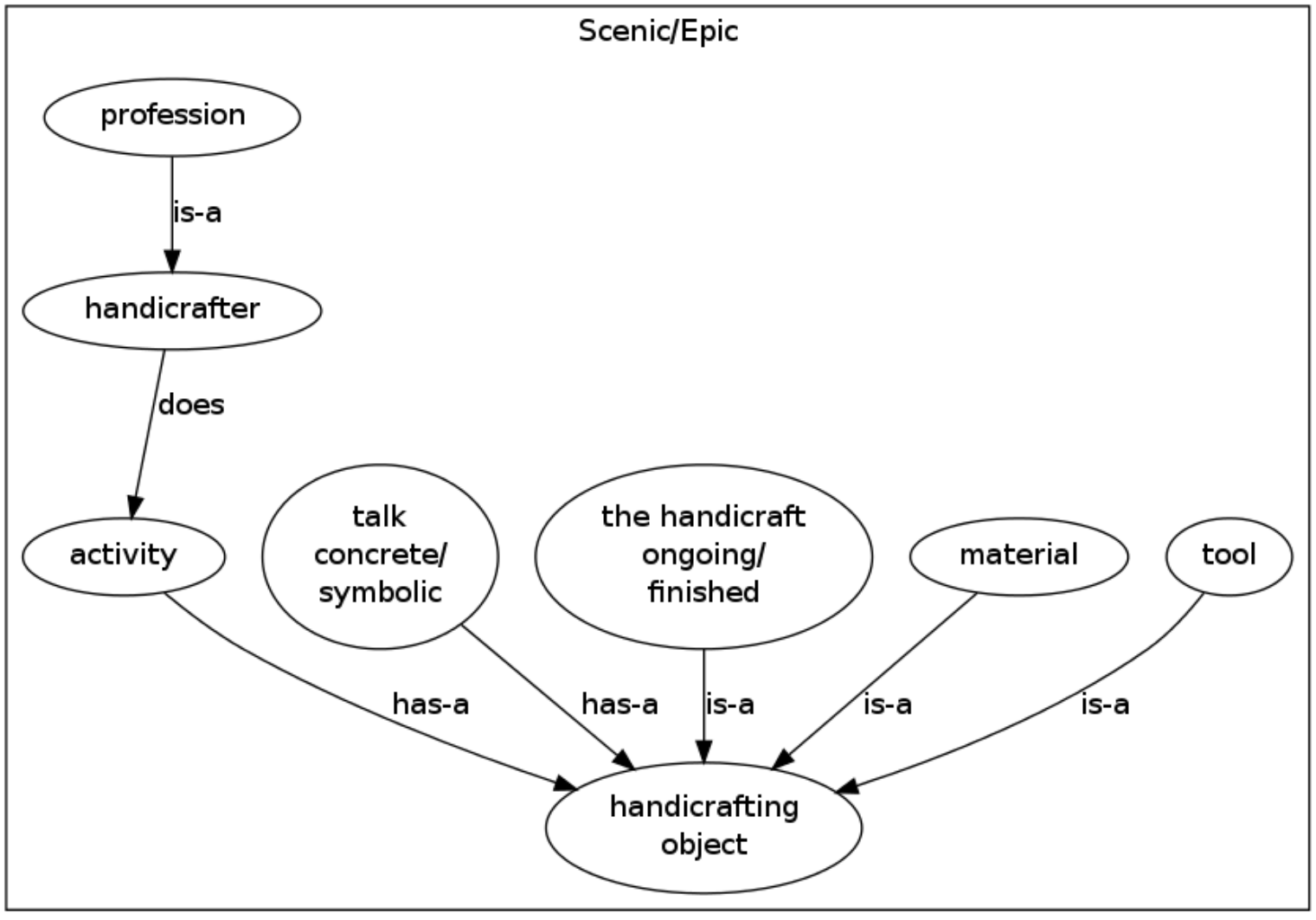
Handicraft

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 stucture elements, with key-value pairs
- They can be tied to anchors to make them discontinuous

GÖTEBORGS UNIVERSITET

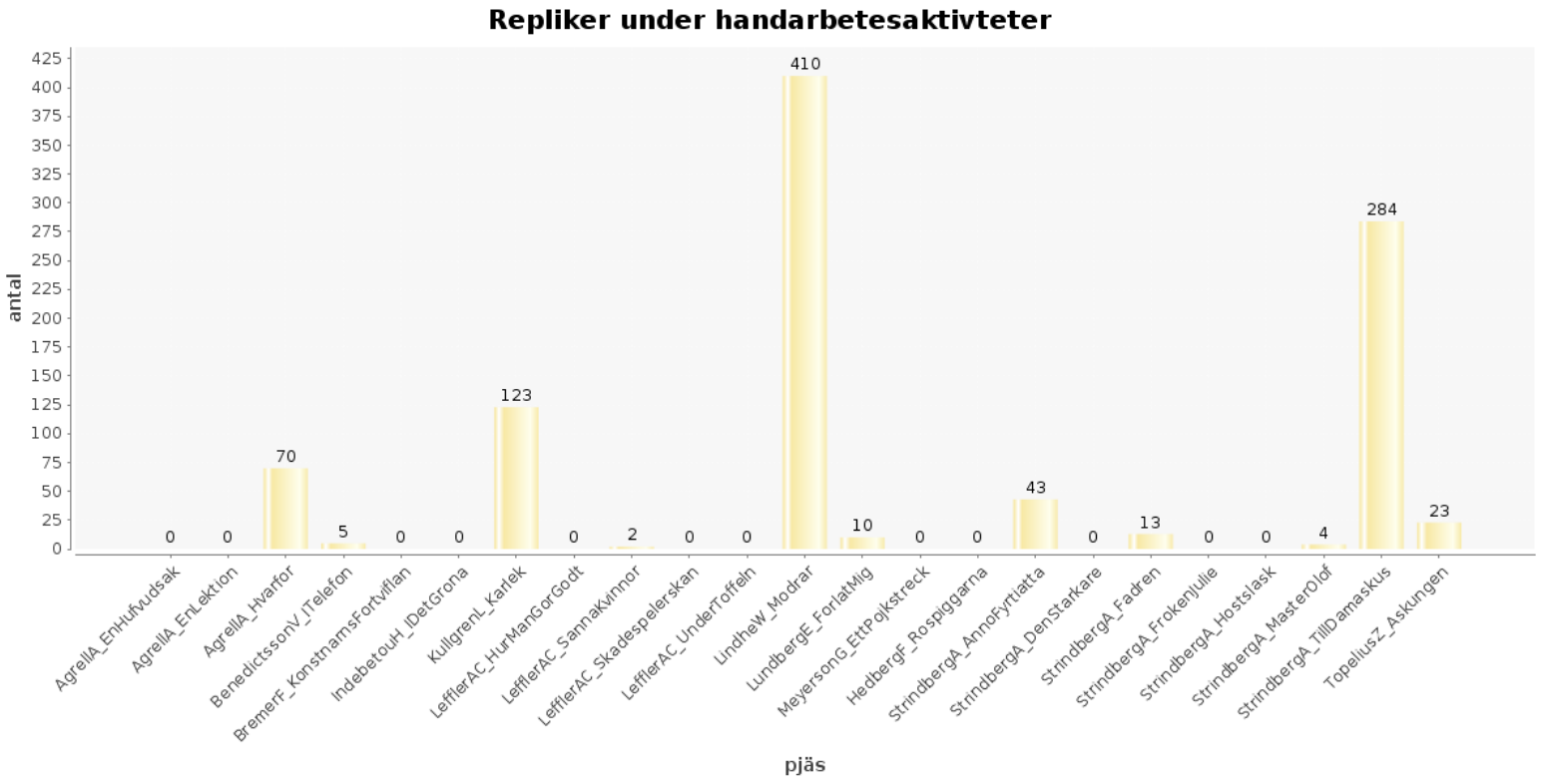
The model (simplified)



GÖTEBORGS UNIVERSITET

Språkbanken

Ongoing handicraft in speeches per play



GÖTEBORGS UNIVERSITET

Språk-BANKEN

Children's play and food and drink terms

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

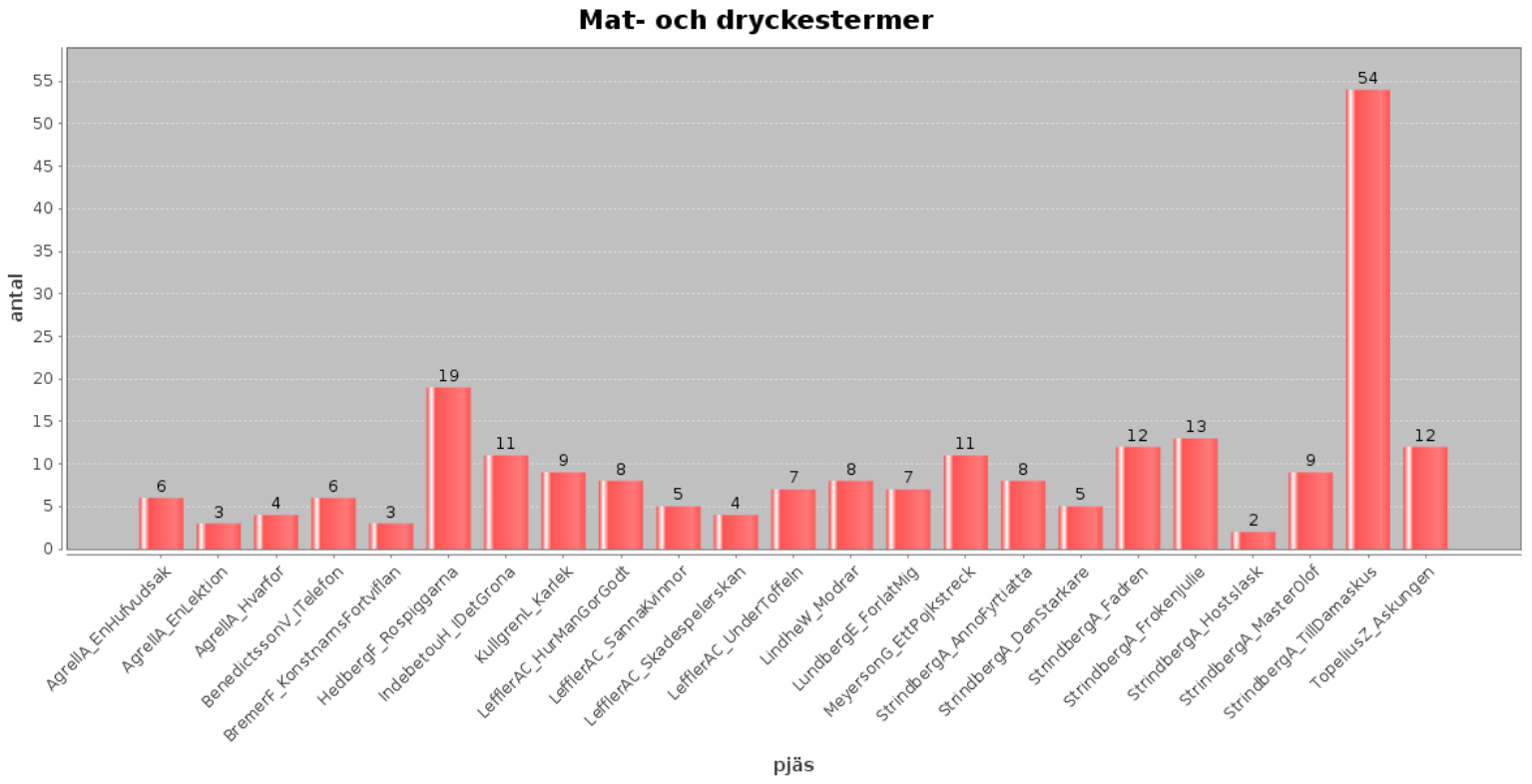
- activity,
- talk about activity,
- and play objects

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.

GÖTEBORGS UNIVERSITET

Språkbanken

Food and drink terms per play



GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

Occupations

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

Yrkesutforskning

Resultat för yrkessökning i Dramawebben

Pjäser med yrken från isco-grupp 58 i pjästexten

Hvarfor (1883)
Agrell, Alfhild

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

▾ Visa fler träffar från denna pjäs (1 st.)

Rospiggarna (1884)
Hedberg, Frans

- Nej, nu för tiden duger bara ungtuppar — nu ska de ju vara läskarlar, vet jag! Det är bara generalerna, som får vara hur gamla de vill — inte gardisterna: (ser uppåt.) Hå hå, jaja! (gäspar). 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

Språkbanken

Yrkesutforskning

Resultat för yrkessökning i Dramawebben

Pjäser med yrken från isco-grupp 2' bland rollkaraktärerna

I telefon (1887)
Benedictsson, Victoria

yrke	rollkaraktär
hushållerska	*Nilla*

Hur man gör godt (1885)
Leffler, Anne Charlotte

yrke	rollkaraktär
verkstadsägare	*kapten Artur Wulf*
statsråd	*statsråd*
kammarherre	*Kammarherren*
marskalk	*marskalkar*

Sanna kvinnor (1883)
Leffler, Anne Charlotte

yrke	rollkaraktär
f.d. hushållerska	*Lovisa*

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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
```

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

Link to images for occupation

Occupation title source	**Kanslist**
Occupation title (english)	Clerk, office
Hiscode	**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	Åhlén 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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

Link to history of work DB

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

Search for Images

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

Sociograms and relations

- 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.

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

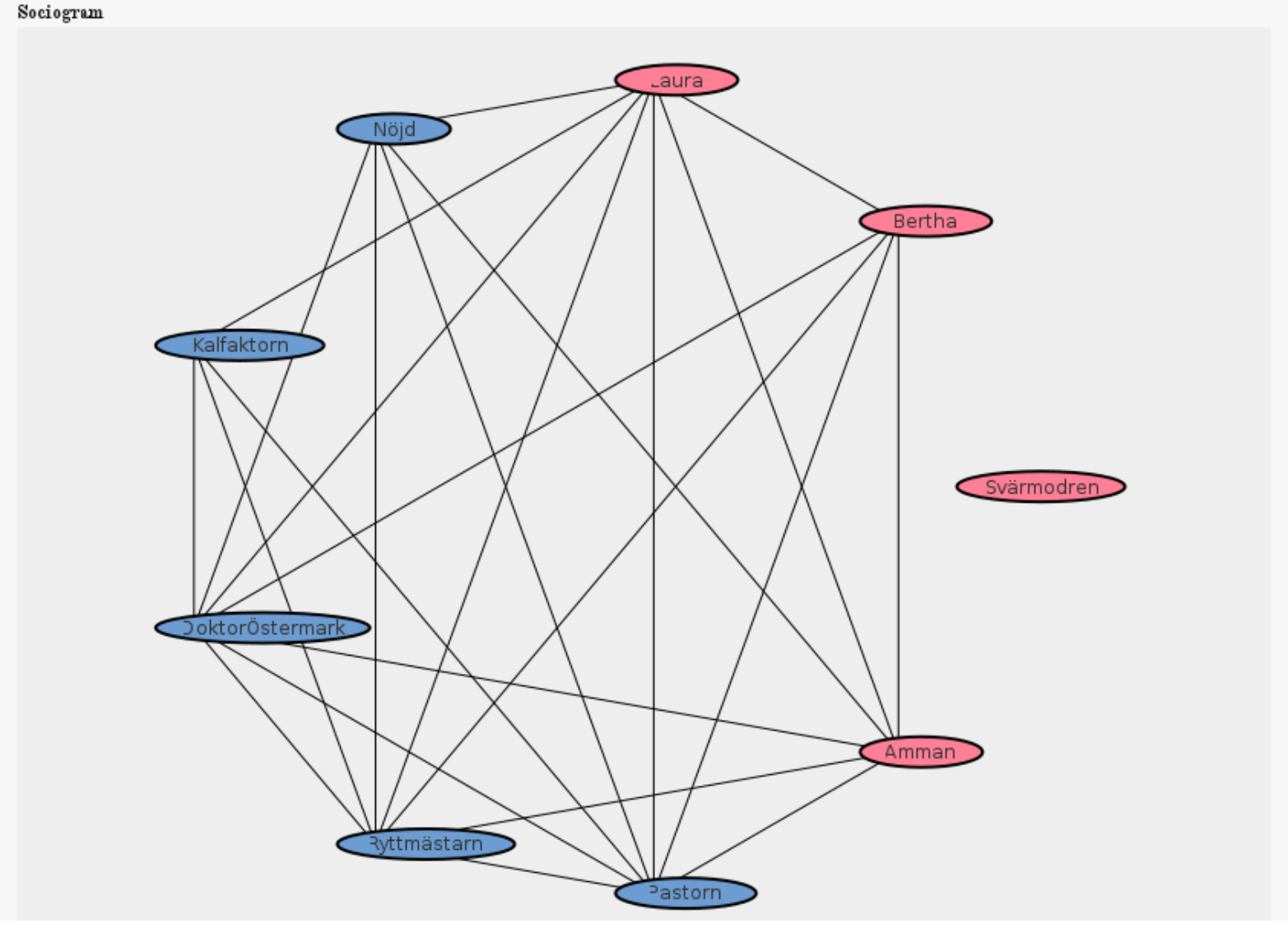
Sociograms

- 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

Språkbanken

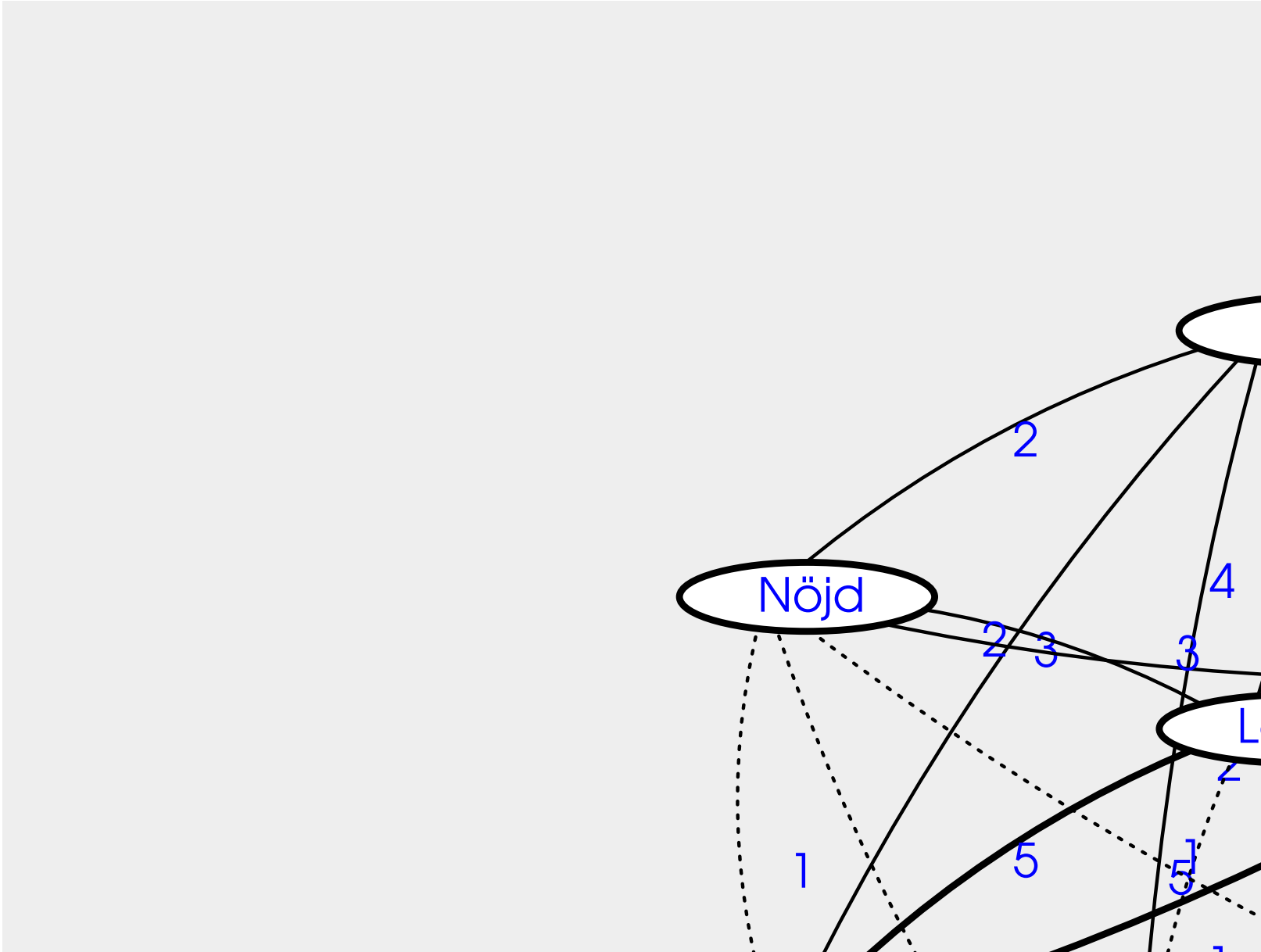
Sociogram for "The father"



GÖTEBORGS UNIVERSITET

Språkbanken

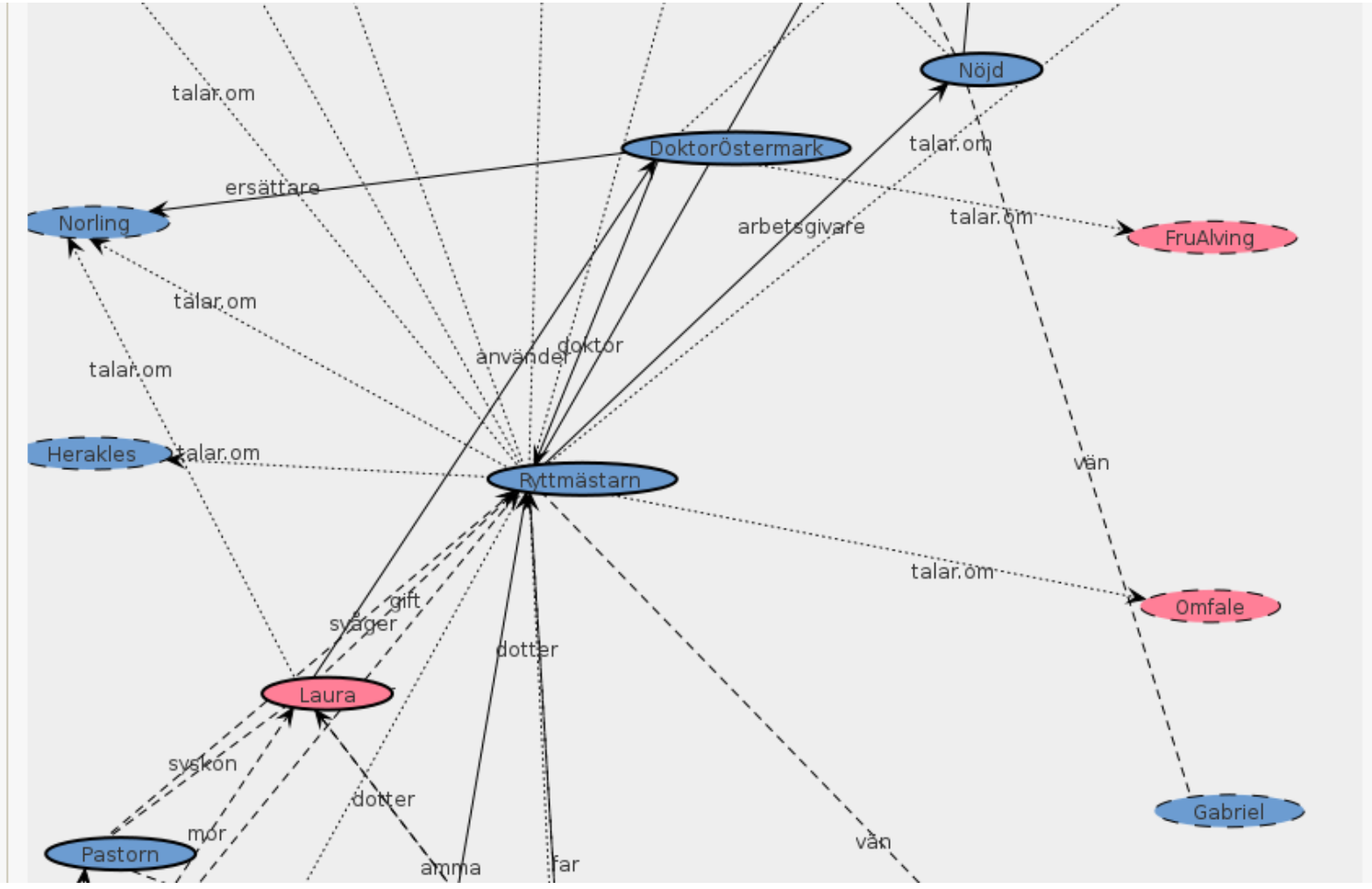
Sociogram for "The father" alternative view



GÖTEBORGS UNIVERSITET

Språkbanken

Person relations in "The Father"



GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

Multilingual document topics


```
<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>
```

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

Where did events occur?

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?

GÖTEBORGS UNIVERSITET

Språk-BANKEN

OCR


```
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

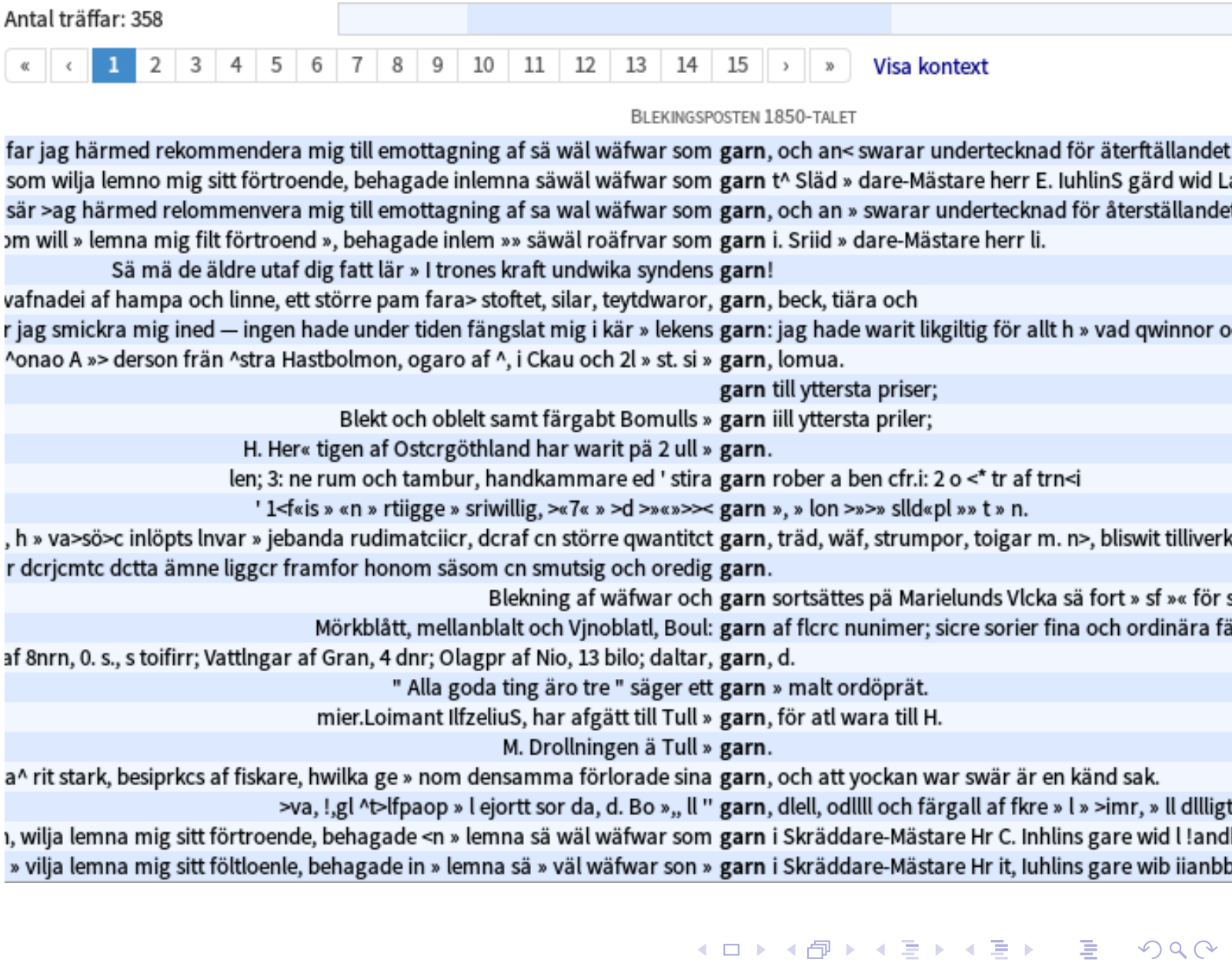
XML Output Live Preview

'db/apps/ocular-ocr/resources/classifiers/swe.fontser

GÖTEBORGS UNIVERSITET

Språk-BANKEN

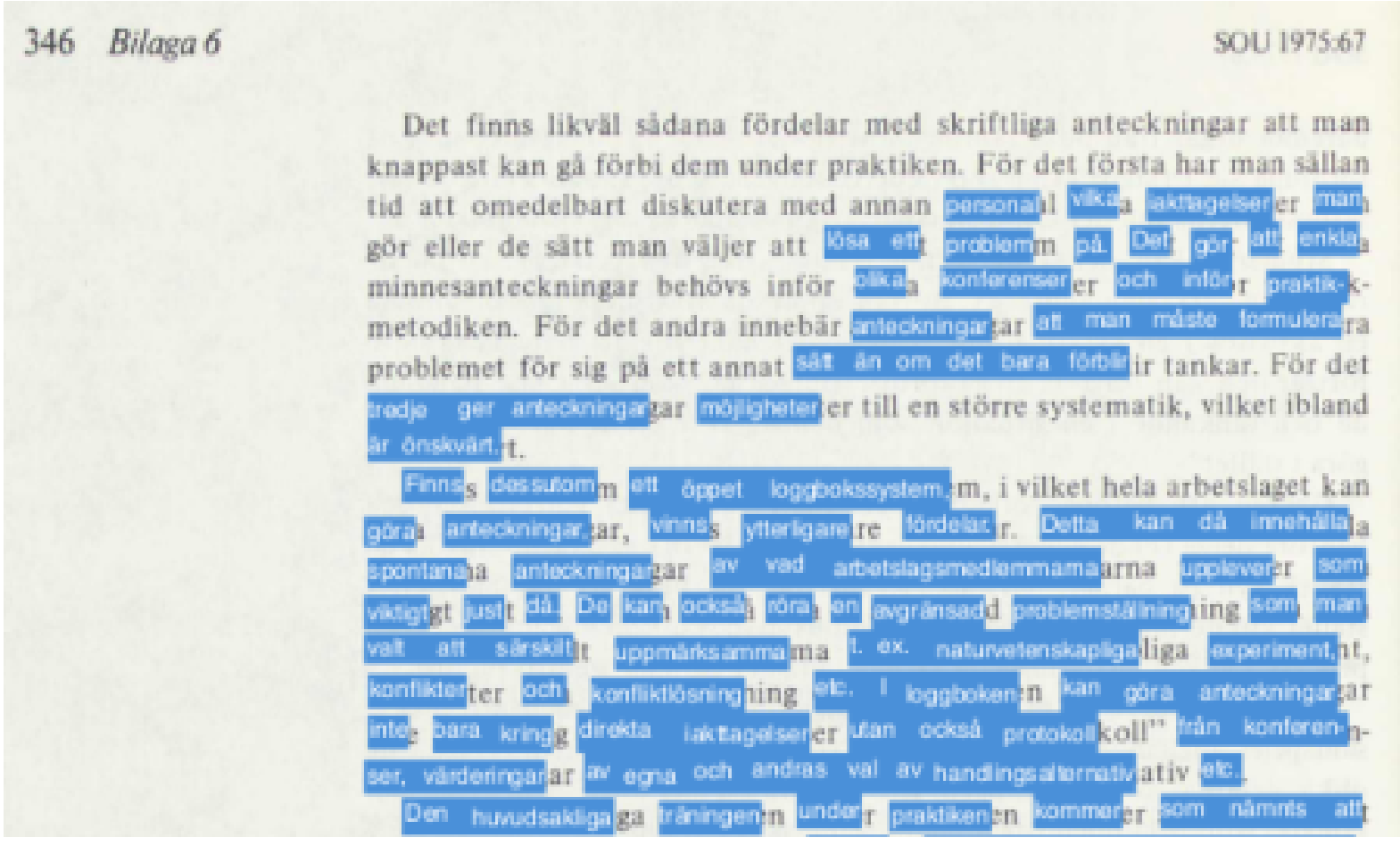
OCR and pdf problems: Korp search ‘garn’ in Blekingeposten



OCR and pdf problems: textflow 1

GÖTEBORGS UNIVERSITET

Språkbanken



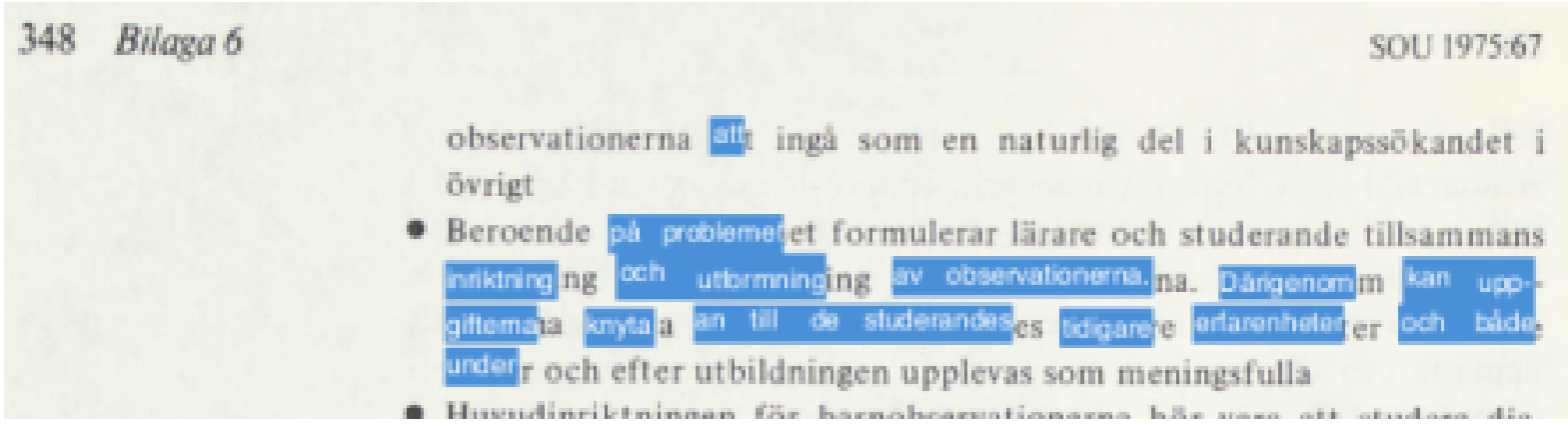
346 *Bilaga 6* SOU 1975:67

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 personal vilka uppfattningar man gör eller de sätt man väljer att lösa ett problem på. Det gör att enkla minnesanteckningar behövs inför olika konferenser och inför praktik-metodiken. För det andra innebär anteckningar att man måste formulera problemet för sig på ett annat sätt än om det bara förblir tankar. För det tredje ger anteckningar möjligheter till en större systematik, vilket ibland är önskvärt.

Finns dessutom ett öppet loggbokssystem, i vilket hela arbetslaget kan göra anteckningar, vinns ytterligare fördelar. Detta kan då innehålla spontana anteckningar av vad arbetslagsmedlemmarna upplever som viktigt just då. De kan också röra en avgränsad problemställning som man valt att särskilt uppmärksamma t. ex. naturvetenskapliga experiment, konflikter och konfliktlösning etc. I loggboken kan göra anteckningar inte bara kring direkta iakttagelser utan också protokoll" från konferenser, värderingar av egna och andras val av handlingsalternativ etc.

Den huvudsakliga träningen under praktiken kommer som nämnts att

OCR and pdf problems: textflow 2



348 *Bilaga 6* SOU 1975:67

observationerna allt ingå som en naturlig del i kunskapssökandet i övrigt

- Beroende på problemet et formulerar lärare och studerande tillsammans inriktning ng och utformning ing av observationerna. na. Därigenom m kan upp- giftema na knyta a an till de studerandes es tidigare e erfarenheter er och både under r och efter utbildningen upplevas som meningsfulla

GÖTEBORGS UNIVERSITET

Språkbanken

Sparql + xml-serialized RDF

```
47  let $query2 := ("PREFIX myhouse: <myhouse://>
48                   SELECT ?x ?c
49                   WHERE { ?x myhouse:count ?c ; a myhouse:furniture }
50                   ORDER BY ASC(?c)")
51  (: ("myhouse://table", "1", "myhouse://chair", "2") :)
52
53  let $query3 := ("PREFIX myhouse: <myhouse://>
54                   SELECT ?x ?r
55                   WHERE { ?x myhouse:room ?r }
56                   ORDER BY ASC(?c)")
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
```

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

Training and evaluation

- 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).

GÖTEBORGS UNIVERSITET

Språkbanken

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 $classifier := (xs:anyURI("/db/apps/stanford-ner/resources/classifiers/english.all.3class.distsim
8  (:    :xs:anyURI("/db/apps/stanford-ner/resources/classifiers/swe-outfile-300dpi-2class-model.ser.gz")
9  xs:anyURI("/db/apps/stanford-ner/resources/classifiers/swe-dw12-3class-model.ser.gz"),
10 xs:anyURI("/db/apps/stanford-ner/resources/classifiers/ces-1872-outfile-djvu-2class-model.ser.gz"))[x
   ?($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 fjällen på semester när lavinen går på
   <org>Blåsjöfjället</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>
```

GÖTEBORGS UNIVERSITET

SWE-CLARIN Språk-BANKEN

Final Comments

- citations, digital and non-digital
- Resources available with open licenses: <https://spraakbanken.gu.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-sparql>, <https://github.com/ljo/exist-mallet>, <https://github.com/ljo/exist-ocular-ocr> and <https://github.com/eXist-db/jfreechart>, more under <https://github.com/ljo/> and <https://github.com/eXist-db/>
- Graphs can be used in svg, graphml and gexf output