



RIKSBANKENS
JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH
SAMHÄLLSVETENSKAPLIG FÖRSKNING

Annotation of learner corpora – first SweLL insights

Elena Volodina, Lena Granstedt, Beáta Megyesi, Julia Prentice,
Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén



RIKSBANKENS
JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH
SAMHÄLLSVETENSKAPLIG FORSKNING

SweLL -

Research infrastructure for Swedish as a Second Language

Elena Volodina, Beata Megyesi, Mats Wirén,

Lena Granstedt, Julia Prentice, Monica Reichenberg,
Gunlög Sundberg



RIKSBANKENS
JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH
SAMHÄLLSVETENSKAPLIG FORSKNING

SweLL -

Swedish Learner Language

Research infrastructure for Swedish as a Second Language

Elena Volodina, Beata Megyesi, Mats Wirén,

Lena Granstedt, Julia Prentice, Monica Reichenberg,
Gunlög Sundberg

SweLL promises (main)

1. Deliver a well-annotated (gold standard) corpus of L2 essays

- 600 essays, approx 100 per CEFR levels A1-C1 + 100 for control L1 learner corpus
- Incl manual error annotation & manually checked linguistic annotation
- Make available for research (and public?)



```
graph = {
  "source": [
    {"id": "s0", "text": "I "},
    {"id": "s1", "text": "don't "},
    {"id": "s2", "text": "know "},
    {"id": "s3", "text": "his "},
    {"id": "s4", "text": "lives "},
    {"id": "s5", "text": ". "}
  ],
  "target": [
    {"id": "t0", "text": "I "},
    {"id": "t1", "text": "don't "},
    {"id": "t2", "text": "know "},
    {"id": "t3", "text": "where "},
    {"id": "t4", "text": "he "},
    {"id": "t5", "text": "lives "},
    {"id": "t6", "text": ". "}
  ],
  "edges": {
    "e-s0-t0": {"id": "e-s0-t0", "ids":
    "e-s1-t1": {"id": "e-s1-t1", "ids":
```

SweLL promises (main)

2. Set a platform (and workflow) for

- Continuous upload of new essays
- Manual error-annotation
- Automatic linguistic annotation

SweLL

Hem › datainsamling › Studenter › Lägg till student

Lägg till student

Swell_id:

Kön:

Födelseårsintervall:

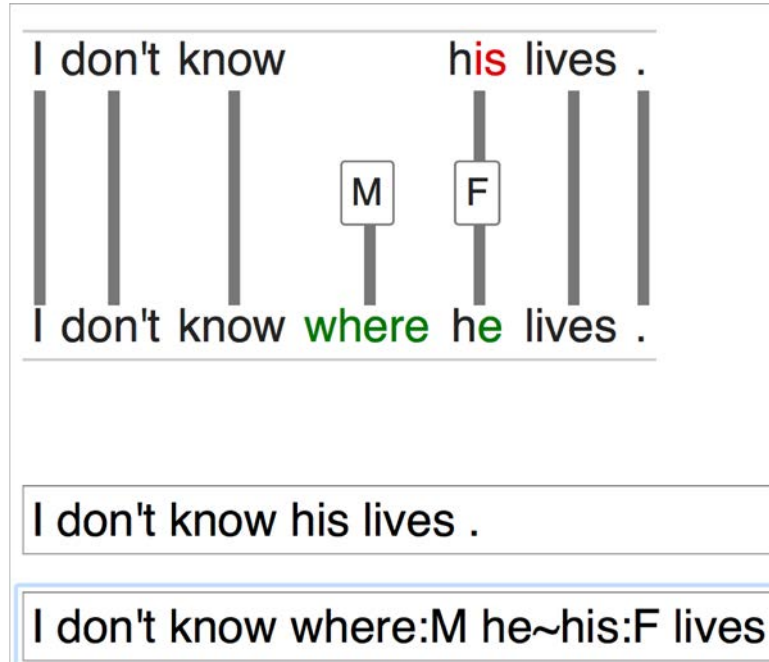
Tid i Sverige:

Högsta examen:

FIRST LANGUAGES

First language: #1

Språk:



<sentence id="8f7-8b5"> [Visa XML]

OO: Direkt objekt (ackusativobjekt)

token	msd	lemma	lex	sense
Jag	PN. UTR. SIN. DEF. SUB	jag	jag..pn.1	jag..1
vet	VB. PRS. AKT	veta	veta..vb.1	veta..1
inte	AB	inte	inte..ab.1	inte..1
var	VB. PRT. AKT	vara	vara..vb.1	vara..1
han	PN. UTR. SIN. DEF. SUB	han	han..pn.1	han..1
bor	VB. PRS. AKT	bo	bo..vb.1	bo..1
.	MAD			

</sentence>

SweLL promises (main)

- Set a platform for browsing L2 essays
 - in concordance fashion (+parallel view)
 - In full text fashion

The screenshot shows the SweLL interface with a concordance search for the word "stress". The search results are displayed in a table-like format with columns for the word and its occurrences in different contexts. The interface includes navigation elements like "KWIC", "Statistik", and "Ordbild". The search results are organized into sections for different years (2012 and 2013) and include various text excerpts where the word "stress" is used.

Context	Word
Omgiven av	stress
är det svårt för människor att hitta lycka.	
stressande	
för ådan – samlas gudingarna i flockar på tiotusental ute på öppna havet för att n	
Duktiga flickor blir sjuka av	stress
Associationssfären kring ordet " kiitorata " och flygplatsmiljön, ger också t.ex. bilder av	stress
löpande band-principer, ekorrhjul, säkerhetskontroller, trafik, affärsmän och snit	stress
Då fanns det ingen sådan	stress
, säger han och blånger ilsket på en mås som är fem före färdig att landa på hans a	stress
m allt " allt som är trevligt är bra för magen ", men jag är säker på att magen mår bättre när man inte	stress
för mycket med att försöka kontrollera allt man stoppar i sig.	stress
Det är lunchrusning i Londons metro och ungefär två miljoner hungriga Londonbor	stress
från förmiddagspalavern till Starbucks för att inhandla en halvliter cappuccino och	stress
– Det dyker alltid upp folk vid lunchdags, men det är inte samma	stress
som under läsåret.	stress
Hör du till dem som inte gillar att	stressa
runt i butiker inför jul?	stressa
Då kunde vi ta det lite lugnare i andra halvlek och inte	stressa
i anfallen som i den första.	stressa
Bort från huvudstaden och	stressen
ingen	stressar
i Hasse Ahlstrands bok och skiva för barn.	stressar
I stället för att ta	stress
över situationen där man fortfarande behövde en poäng för att ha sitt på det torr	stress
– Vi borde våga plocka bort	stressen
med resor och fruktansvärd arbetstakt.	stressar
De som jobbar i affärlivet eller exportindustrin (och hämtar in pengar till Finland)	stressande
och man kan börja må dåligt.	stressande
som ung, då man ännu söker sig själv, möta alla krav och förväntningar kan kännas frustrerande och	stress
kan ge upphov till.	stress
Clownerna konfronteras med det personliga mörker som	stress
Budskapet om fred och sinnesfrid går hand i hand med konflikfyllda känslor om materiella ting och	stress
– Det fanns gott om tid så vi har inte behövt	stressa
i onödan, konstaterar Kaikkonen som tillsammans med överstyrmannen ansvara	stressa
Filmen frågar om	stressen
och konsumtionen i väst är ett verkligt alternativ, ett alternativ till ett meningsful	stressen
och trötthet.	stress
– Det var	stressande
och jag var helt slut.	stressande
Eller med ökande brädska och	stress
, evinnerlig jämförelse och fruktan att man inte reder sig i tävlingen, att man lever	stress
isom ligger bakom, kanske det att jag mest åkt för mig själv och med frun Nives, och inte tagit någon	stress
Osäkerhet,	stress
och knappa resurser är vardag för många kommunalt anställda.	stress

SW1203-UPPSATSER

I Sverige lever många människor det goda livet tycker jag. Det är inte så i många andra land. Människor bryr sig om att äta, träna och sova ordentligt. De vill också ha ett rikt socialt liv vilket är viktigt för psykosocial hälsa. Att ha ett gott liv är något viktigt för mig. Det är ett ständigt jobb. Man måste alltid tänka på sin hälsa. Om man inte har några problem med hälsa måste man träna. Idag sitter vi mycket mer än förrut. Sittande arbete gör oss lat och vi har inte inspiration att börja röra på oss. En familj som ofta vandrar i fjällen, cyklar, promenerar eller lekar ute tillsammans har det goda livet enligt mig. Mat är en av viktigaste saker angoende det goda livet. Man måste välja väldigt noga sin mat. Det är lätt att vara nöjd med halverdig mat vilket man lagga snabbt. Att lagga riktig nytig mat tar mycket tid och man bör förbereda sig. Jag menar att jag måste köpa färska grönsaker om jag ska lagga någon nytig mat. Jag menar med mat försöker man att äta hälsosamt och undvika fetma, diabetes, hög blodtryck och hjärt och kärlsjukdomar. Det är bäst om man är vegetarisk och icke-rökare. Jag tror att frasen "Det goda livet " ska referera till glad familjen som lever hälsosamt liv utan stress. Å andra sida är jag inte säkert att det är möjligt i ett modernt samhälle leva detta liv. I dagens samhället ar viktigt att tjäna mycket pengar därför att pengarna betyder en hög status och vi alla vill ha hög status.

I modernt samhälle kommer tyvärr stress och många andra negativa saker. Till slutet vill jag säga att det goda livet är mitt mål. Ett foto av lycklig familjen på ett bord.

Nu tillbaka till Europa och Sverige. Här har människorna andra problem. Stress, långa

SweLL focus (main)

- Adult learners (16+ years)
- Healthy learners
- Written essays (no speech data)
- Where possible – longitudinal data

An electronic research infrastructure



- (free accessible) data in electronic format
- technical platform for exploring data, including tools and algorithms for data analysis, and visualization
- a set of tools and technical solutions for new data collection and preparation, including data processing and annotation
- a network of experts in the relevant disciplines, incl. legal and ethical questions



Data



The Not-So-Secret life of a PI

Available & reliable data



- Restrictions on use of personal information to protect “subjects”, i.e. physical people
- Important consequences for learner corpora (L2) projects –
IF you want data to be available for research!
 - Metadata precautions
 - Text de-identification and pseudonymization
 - Name-ID mapping keys handling

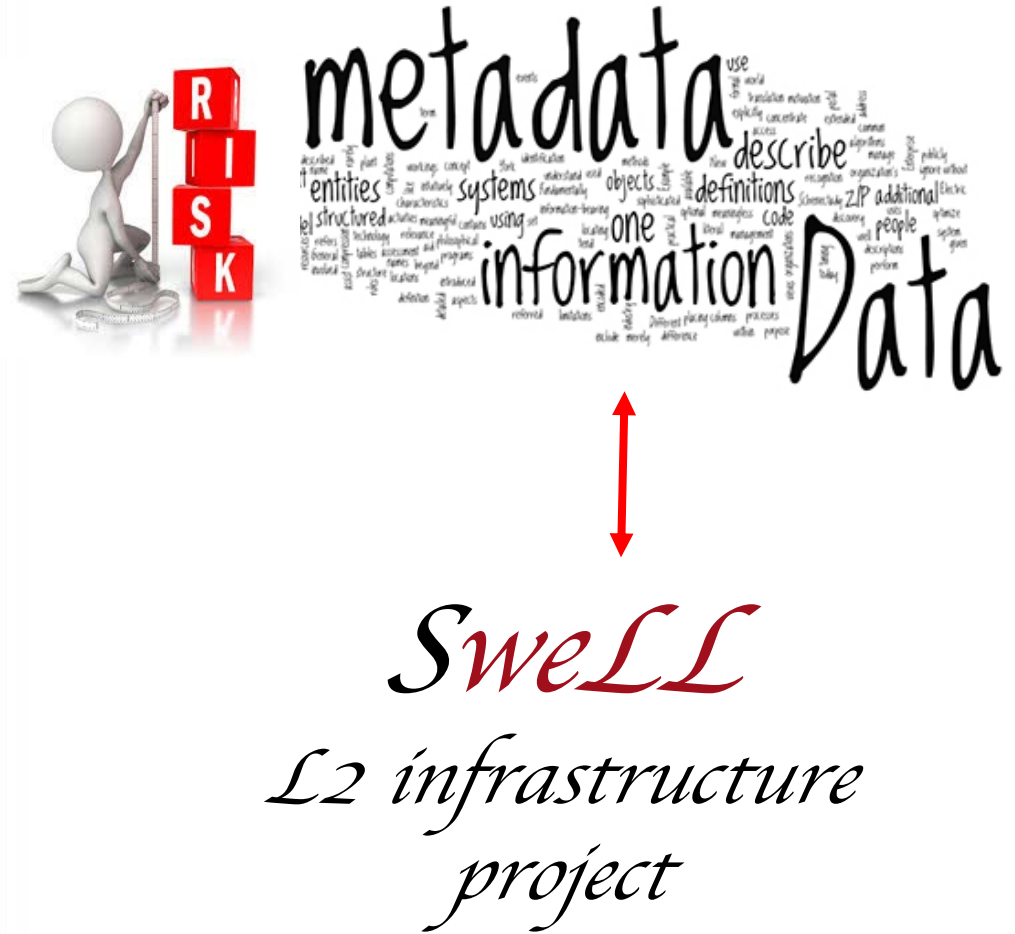
SOCIO-DEMOGRAPHIC METADATA

- L1: Romansh, German, Korean
- Year of birth: 2001
- Gender: male
- Education / highest degree: high school
- Time in L2 country: 1 year
- Other languages: Russian, French

TASK METADATA:

- Date: April 2018
- CEFR level: A2

TEXT: "My name is Ali and I live in Växjö. I am 17 years. I moved to Sweden one year ago. I like Växjö. I am web developer."



SOCIO-DEMOGRAPHIC METADATA

- L1: Romansh, German, Korean
- Year of birth: 2001
- Gender: male
- Education / highest degree: high school
- Time in L2 country: 1 year
- Other languages: Russian, French

TASK METADATA:

- Date: April 2018
- CEFR level: A2

TEXT: "My name is Ali and I live in Växjö. I am 17 years. I moved to Sweden one year ago. I like Växjö. I am web developer."

No information on the country of birth

Birthyear: 5-year spans, e.g. 2000-2004

No exact date for entering the L2 country



No information on school or teacher

Pseudonymization of text data: names, cities, ages, professions, etc.

SVALA pseudonym. tool

[Demo](#)



undo redo

previous next
prev mod next mod

Filter / numeric label

2 Morphology
gen
def
Errors
ort
Names
firstname:male
firstname:female
firstname:unknown
surname
middlename
initials
Geographic data
country_of_origin
country
zip_code
region
city-SWE
city
area
place
geo
street_nr

show options

1

• Ali

2

• Borlänge.
• Borlänges

3

• 18

4

1 Jag heter Ali och bor i Borlänge. Jag är 18 år. Jag

2 Jag heter Peter och bor i Guntorp Jag är 19 år. Jag

3 flyttade till Sverige för 3 år sedan. Jag gillar Borlänges gator.

4 flyttade till Sverige för 2 år sedan. Jag gillar Guntorp-gen gator.

firstname:male 1
city-SWE 2
age_digits 3
year 4
city-SWE 2 gen

1. ORIGINAL TEXT → @PLACEHOLDER → RENDERING

2. ORIGINAL TEXT → @PLACEHOLDER → REPLACEMENT

3. ORIGINAL TEXT → @PLACEHOLDER → ORIGINAL

```
e-s6-t49": {  
  "id": "e-s6-t49",  
  "ids": ["s6", "t49"],  
  "labels": ["city-SWE", "2"],  
  "manual": true  
}
```

Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish

Beáta Megyesi¹, Lena Granstedt², Sofia Johansson³, Julia Prentice⁴, Dan Rosén⁴,
Carl-Johan Schenström⁴, Gunlög Sundberg³, Mats Wirén³ & Elena Volodina⁴

¹Uppsala University, ²Umeå University, ³Stockholm University, ⁴University of Gothenburg, Sweden

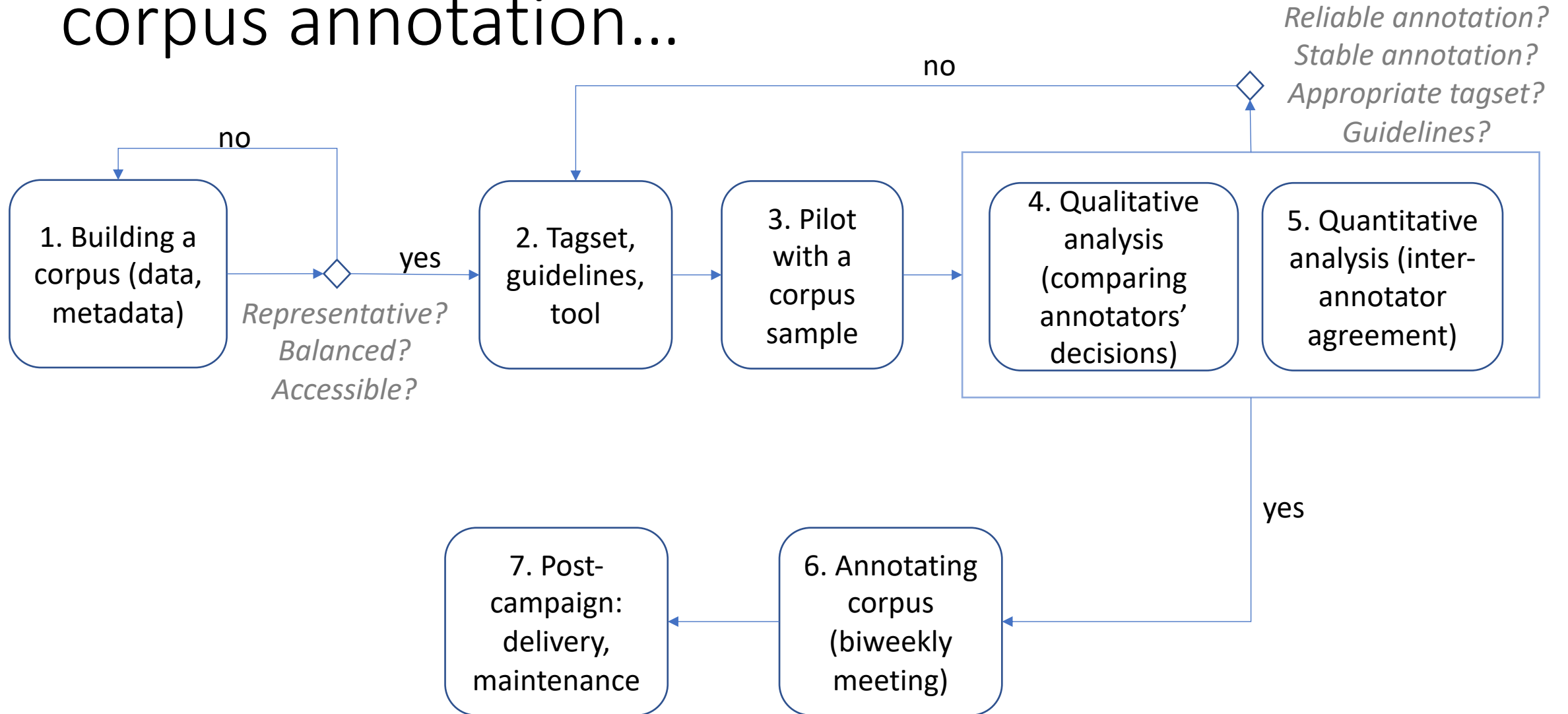
Annotation makes data interesting/useful
(you get what you annotate)

Annotation should better be *good*,
i.e. reliable

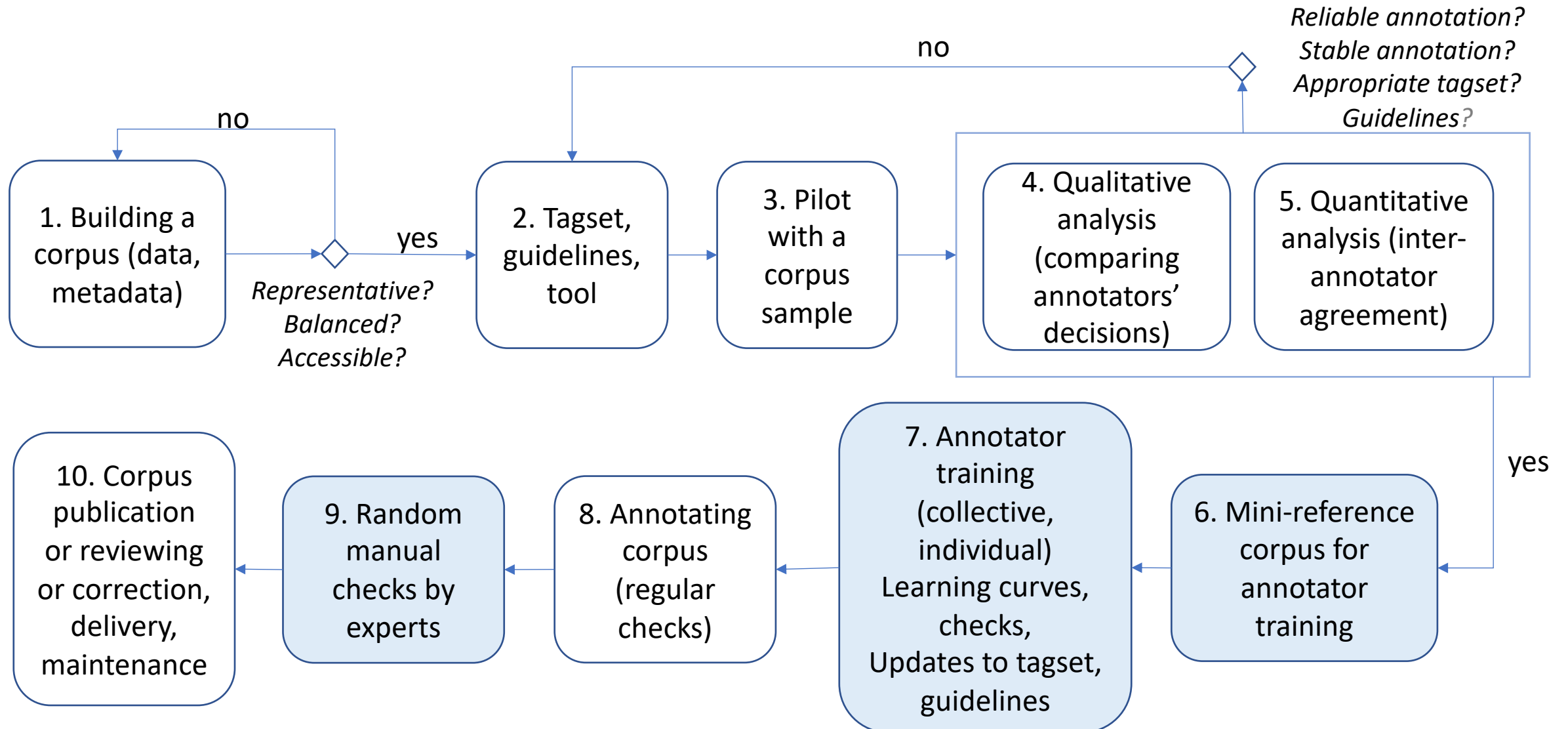


Gold standard corpus

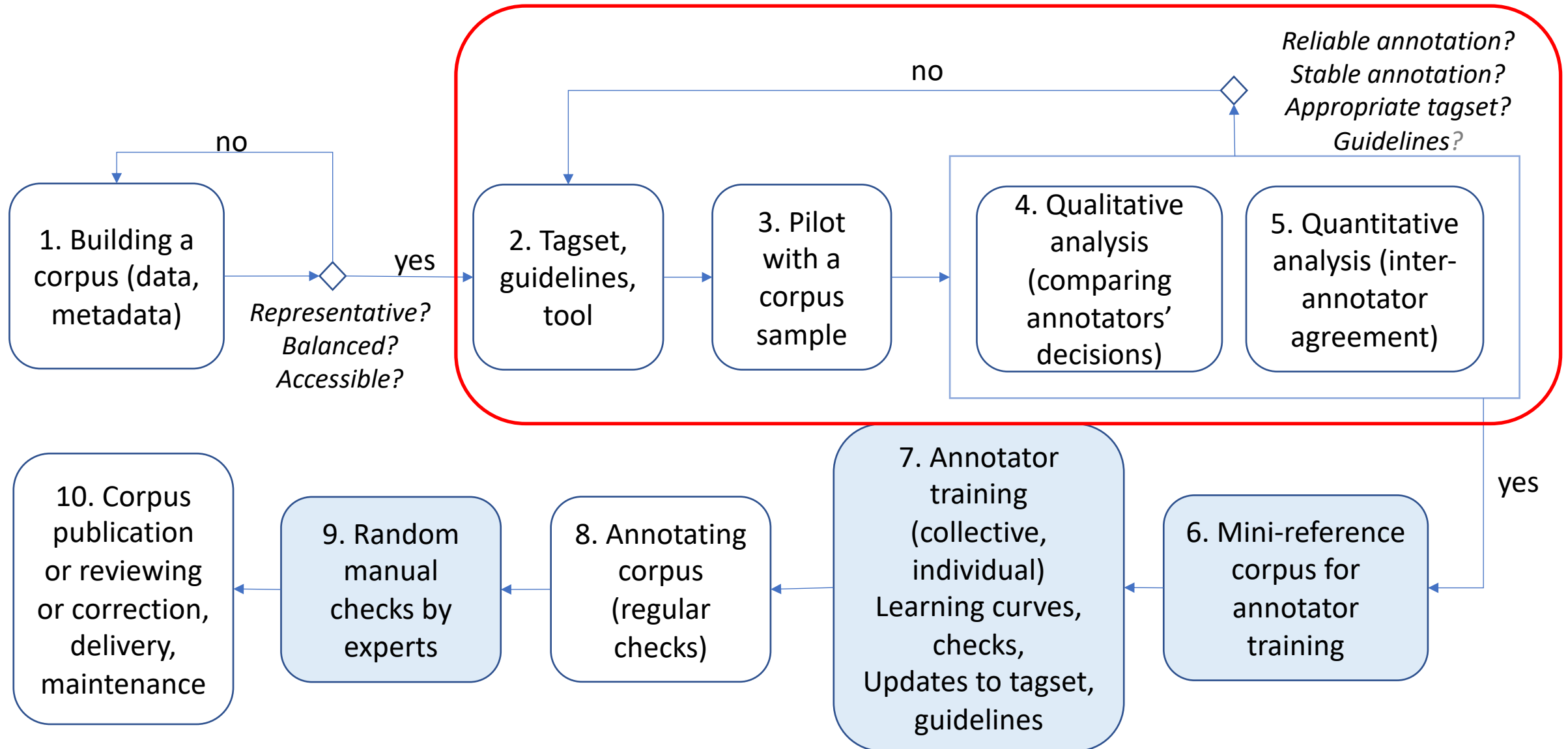
Hovy et al. 2010. Towards a "Science" of corpus annotation...



Fort. 2016. Collaborative annotation...



Fort. 2016. Collaborative annotation...



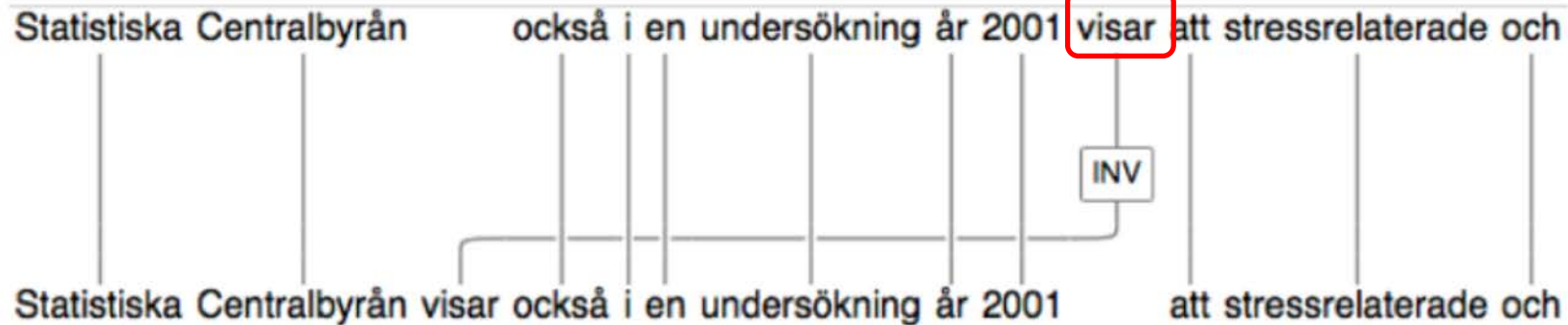
Annotation - choosing a taxonomy



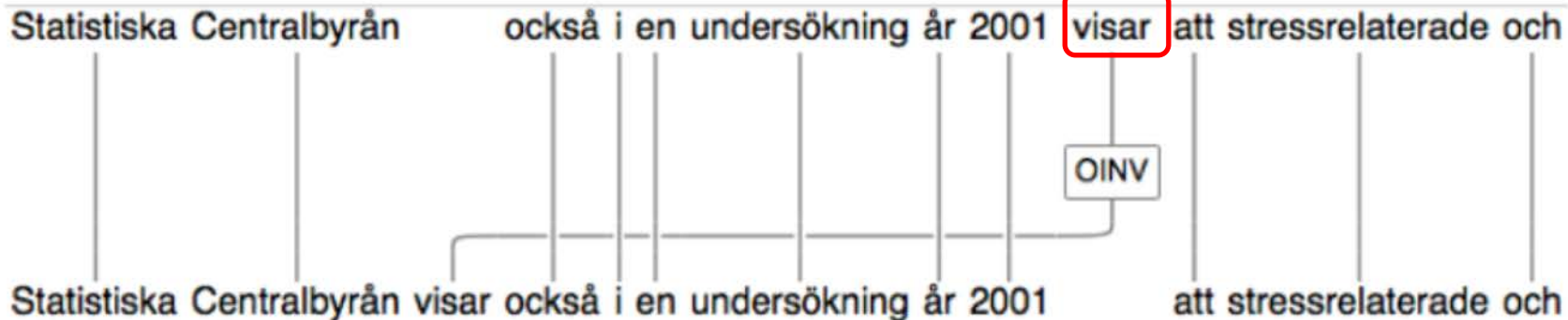
- Other taxonomies :
 - ASK : 23 Error types
 - Lexical (8), morphological (3), syntactical (7), punctuation (4), unidentified (1)
 - MERLIN: 64 error types
 - grammar (21), orthografic (8), intelligibility (8), vocabulary (10), coherence (4), sociolinguistic (10), pragmatics (3)
- How detailed should the taxonomy be?
- How important is the target language?
 - similarity between Norwegian and Swedish
 - Comparability between ASK and SweLL wanted

Taxonomy ambiguity

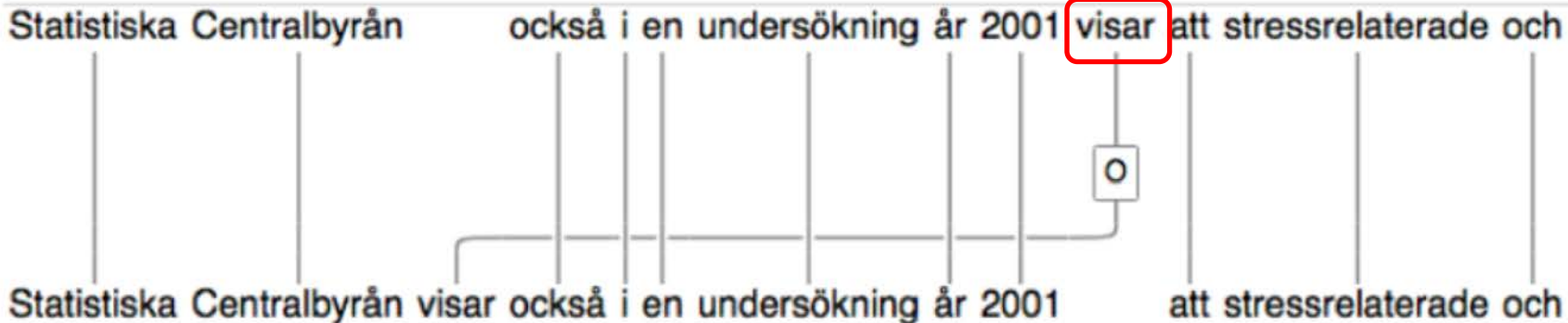
Beata



Elena

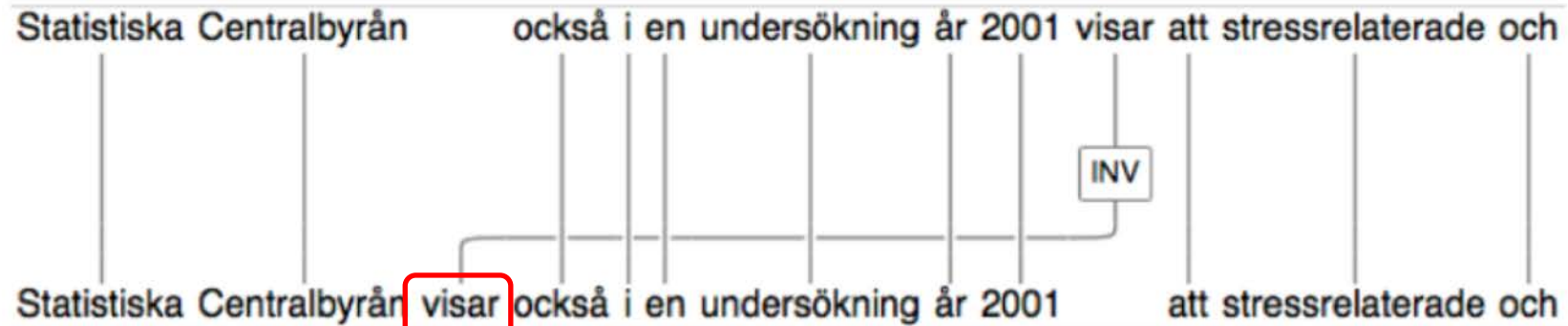


Julia

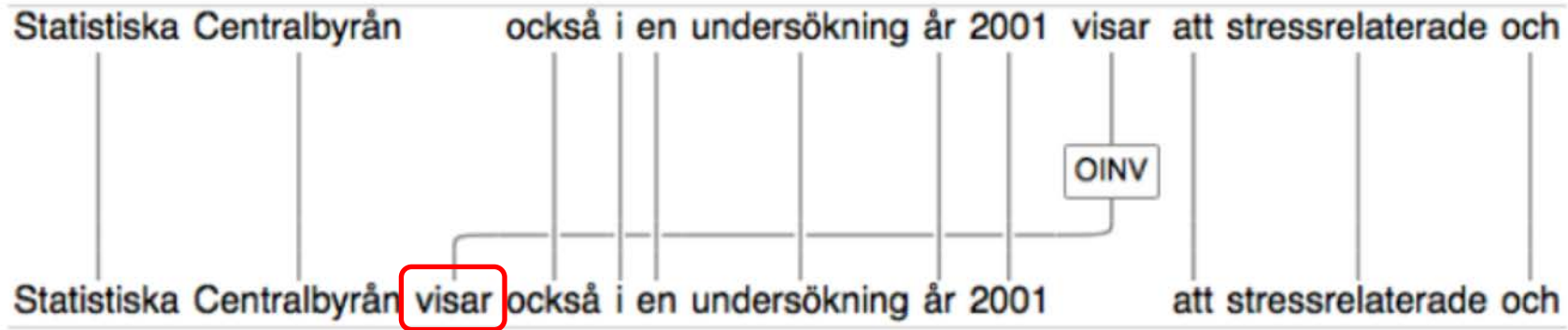


Taxonomy ambiguity

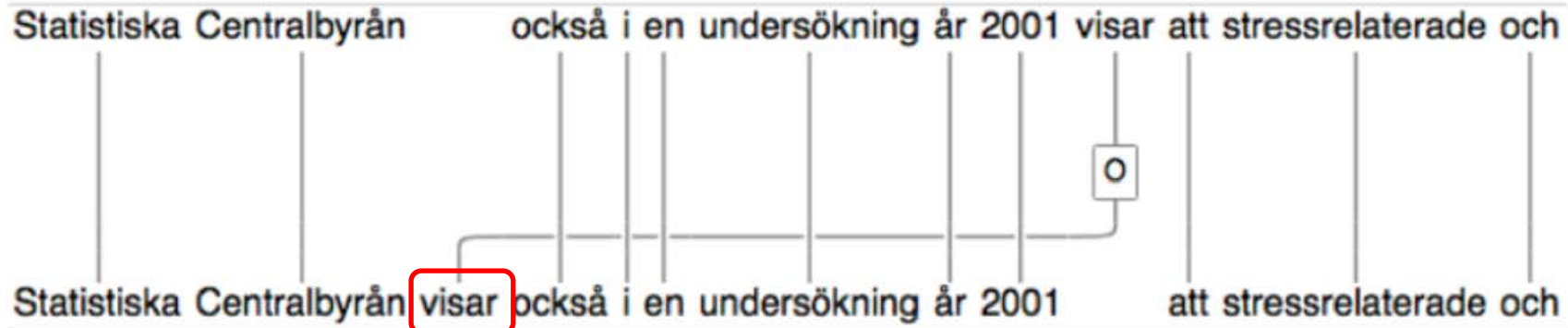
Beata



Elena

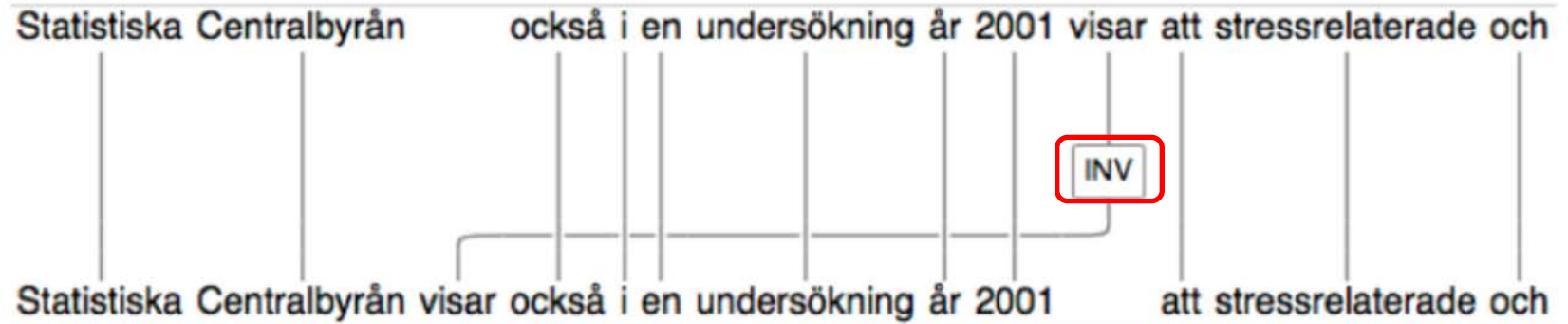


Julia

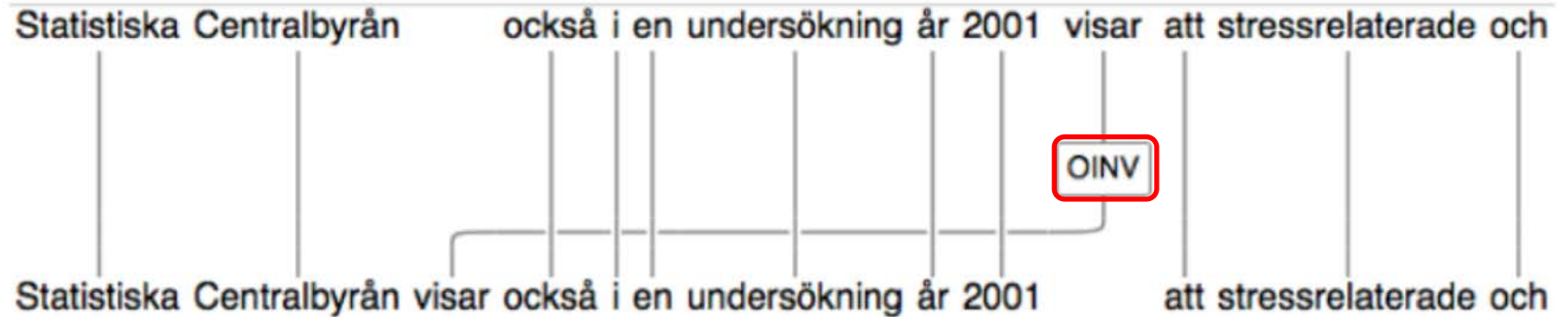


Taxonomy ambiguity

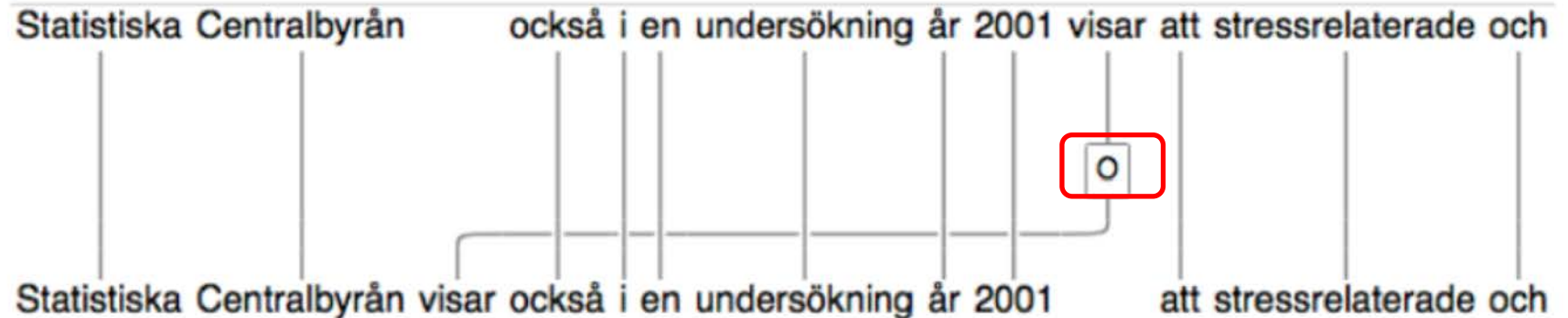
Beata



Elena



Julia



Normalization

* *I has was*

- Re-writing L2 learner original in a normative way, creating a so-called target hypothesis (Lüdeling et al., 2005)

Normalization

* *I has was* → *I have been ? I was? I had?*

Normalization: basic principles

- Minimal change
- Positive assumption
- Lexical and grammatical competence prior to functional and structural correctness

Example

* Jag trivs mycket **bor** med dem.
(Eng) I enjoy much live with them.

Potential target hypotheses:

Jag trivs mycket **bra** med dem → Minimal change (seemingly)
→ Error: wrong word / spelling?

Jag trivs mycket **med att bo** med dem → Lexical competence of *BO*, *verb*
→ Errors: idiomaticity error (trivs) +
wrong verb form (bo)

Target text:

Hej , mitt namn är firstname:female_1_ort , jag bor i area_2_ort mot area_3_ort . Min bostad är stor och har gul färg , ett fint hus . Jag bor med min mamma , pappa och ett syskon . Jag trivs mycket med att bo med dem . Allt är bra för att jag kan inte bo ensam och min mamma lagar mat så bra . Jag trivs med hennes mat och när jag har tid kan jag

Jag bor med min mamma , pappa och en syskon .
M-GEN
Jag bor med min mamma , pappa och ett syskon .

jag trivs mycket bor med dem .
O-CAP L-ID S-M S-M M-VERB
Jag trivs mycket med att bo med dem .

Why normalization as a separate step?

- It helps to build a better understanding of a learner's linguistic competence

Why normalization as a separate step?

- It helps to build a better understanding of a learner's linguistic competence
- It can be outsourced to SLA researchers for doing it

Why normalization as a separate step?

- It helps to build a better understanding of a learner's linguistic competence
- It can be outsourced to SLA researchers for doing it
- Error annotation depends on the change applied to the original text
→ and as such it is not ERROR annotation, but CORRECTION annotation

Why normalization as a separate step?

- It helps to build a better understanding of a learner's linguistic competence
- It can be outsourced to SLA researchers for doing it
- Error annotation depends on the change applied to the original text – and as such is not ERROR annotation, but is CORRECTION annotation
- Inter-annotator agreement with respect to error codes can be objectively measured only given that the annotators are working on the same normalized version

Next steps (2018-2020)

- Finalize mini-reference corpus
- Full scale annotation of essays
- Develop functionalities in Korp/Strix for browsing, visualizing, and statistic analysis of L2 data

I stop here...

...but have a lot of details on various aspects.

Please ask questions.

Project webpage: https://spraakbanken.gu.se/eng/swell_infra