RIKSBANKENS JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

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)

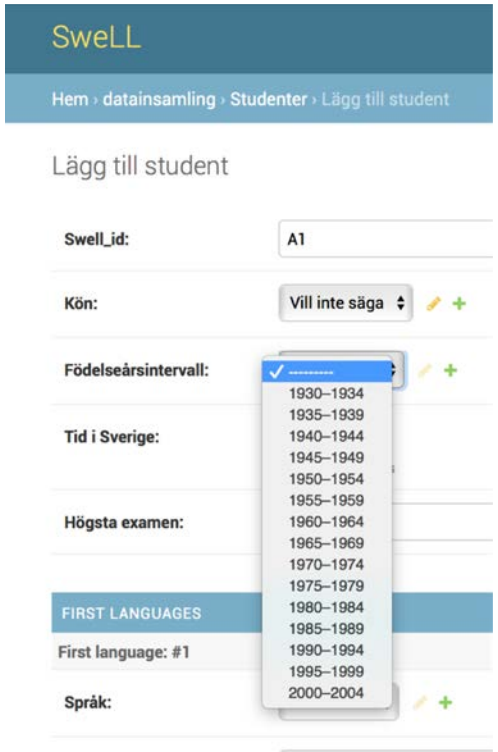
1. Deliver a well-annotated (gold standard) corpus of L2 essays
   - 600 essays, approx 100 per CEFR levels A1-C1 + 100 for control L1 learner corpus
   - Incl manual error annotation & manually checked linguistic annotation
   - Make available for research (and public?)

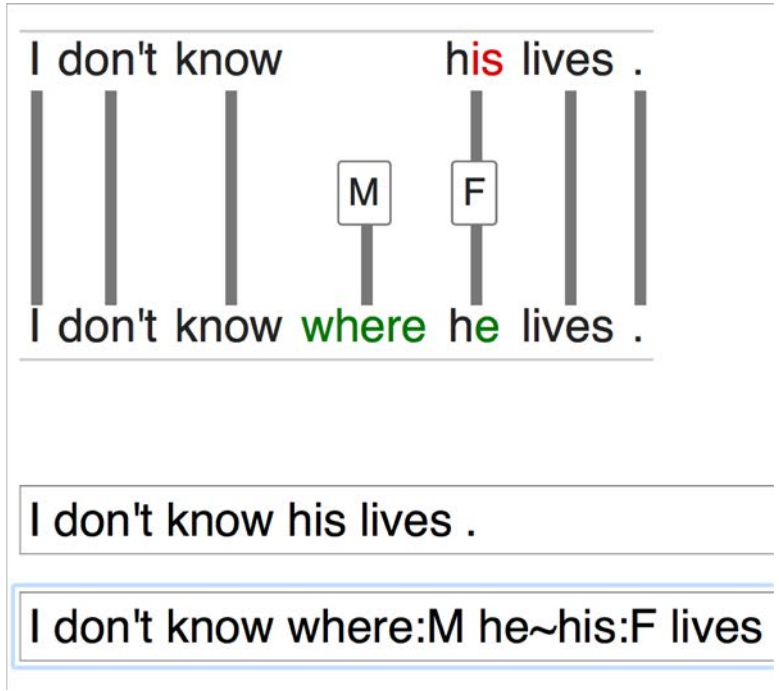
→

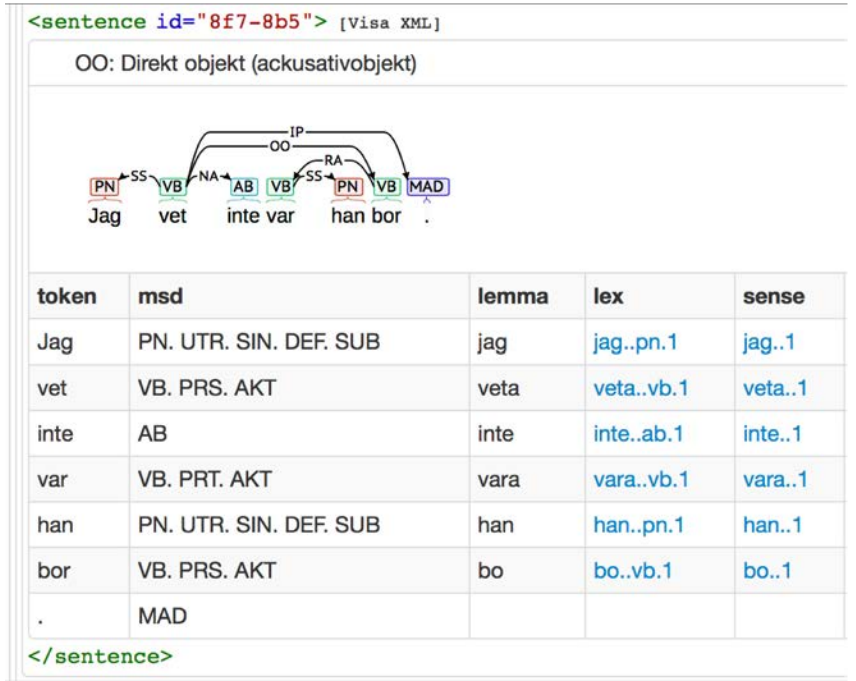
```
graph = {
  "source": [
    {"id": "s0", "text": "I "},
    {"id": "s1", "text": "don't "},
    {"id": "s2", "text": "know "},
    {"id": "s3", "text": "his "},
    {"id": "s4", "text": "lives "},
    {"id": "s5", "text": ". "}
  ],
  "target": [
    {"id": "t0", "text": "I "},
    {"id": "t1", "text": "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

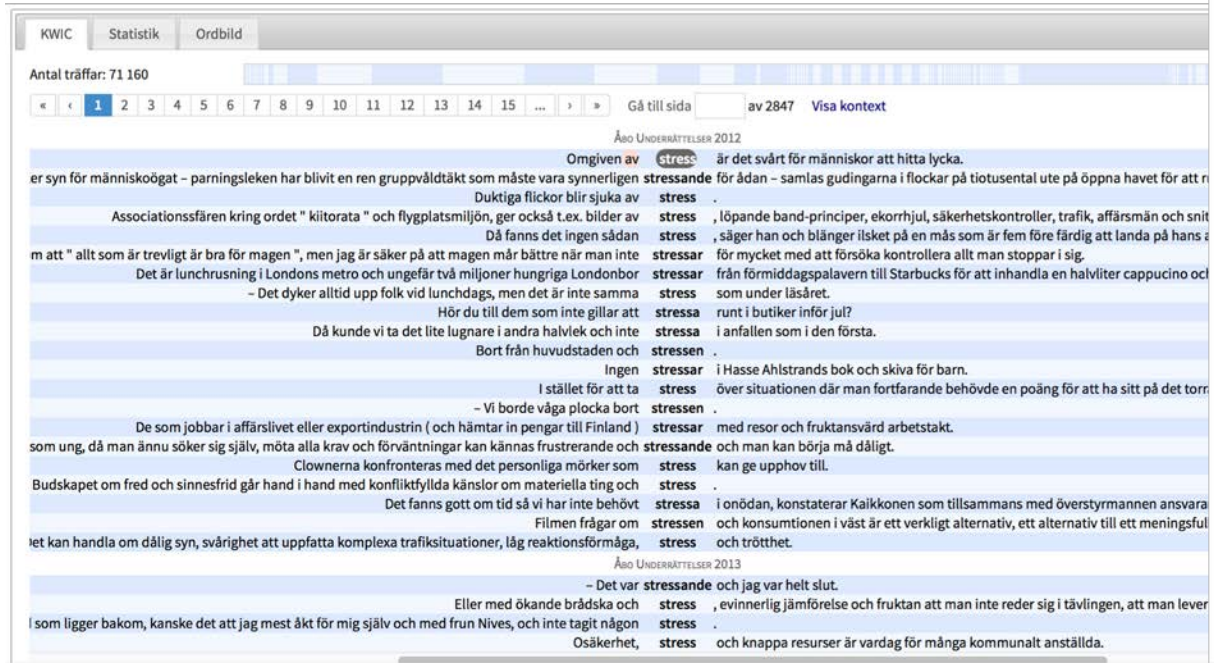


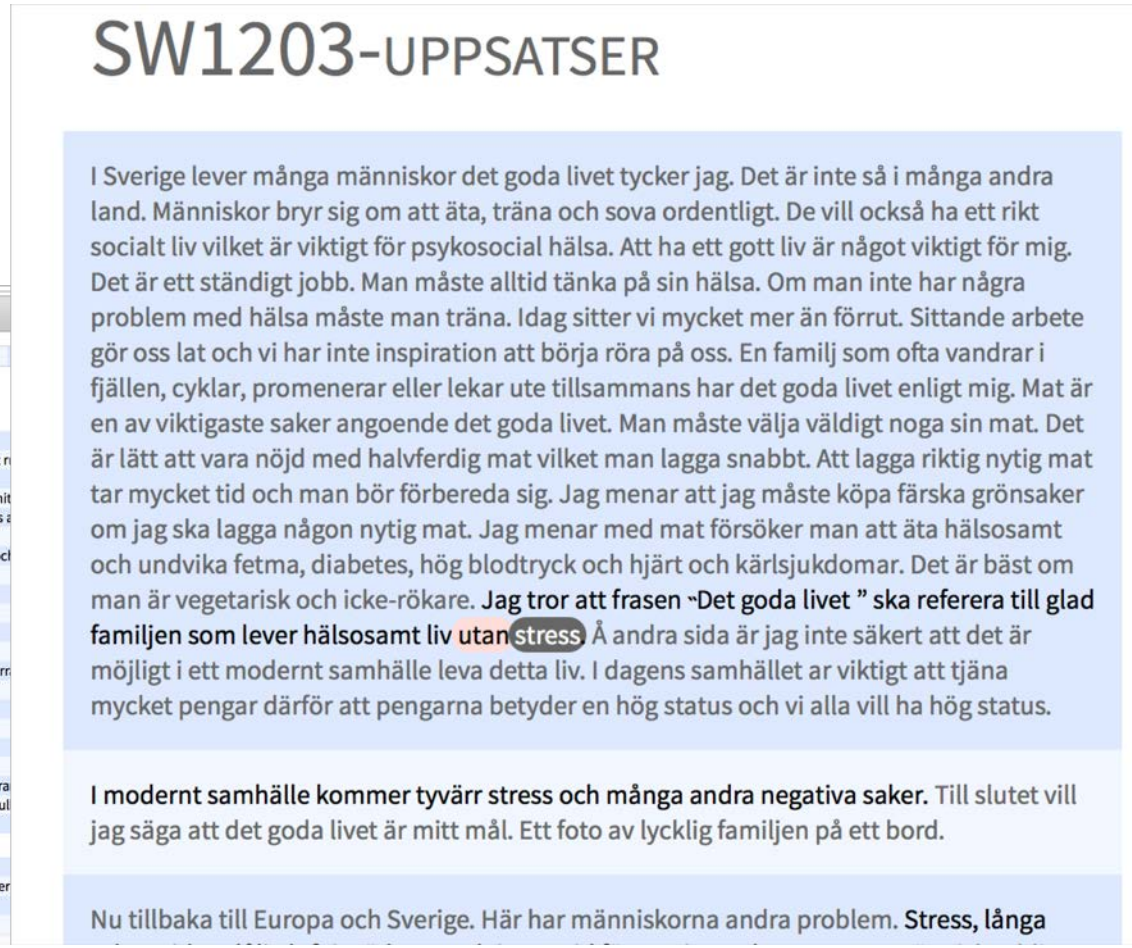




SweLL promises (main)

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





SweLL focus (main)

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

An electronic research infrastructure

"WARNING. You have violated HIPAA! An email notification has been sent to a federal agency, your supervisor, and your mother."

- (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."

SweLL

L2 infrastructure project

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 show options

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

Jag heter Ali och bor i Borlänge. Jag är 18 år. Jag
firstname:male 1 city-SWE 2 age_digits 3
Jag heter Peter och bor i Guntorp Jag är 19 år. Jag

flyttade till Sverige för 3 år sedan. Jag gillar Borlänges gator.
year 4 city-SWE 2 gen
flyttade till Sverige för 2 år sedan. Jag gillar Guntorp-gen gator.

1
- Ali

2
- Borlänge.
- Borlänges

3
- 18

4
- 3

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 Megyesi[1], Lena Granstedt[2], Sofia Johansson[3], Julia Prentice[4], Dan Rosén[4],
Carl-Johan Schenström[4], Gunlög Sundberg[3], Mats Wirén[3] & Elena Volodina[4]
[1]Uppsala University, [2]Umeå University, [3]Stockholm University, [4]University of Gothenburg, Sweden

Annotation makes data interesting/useful

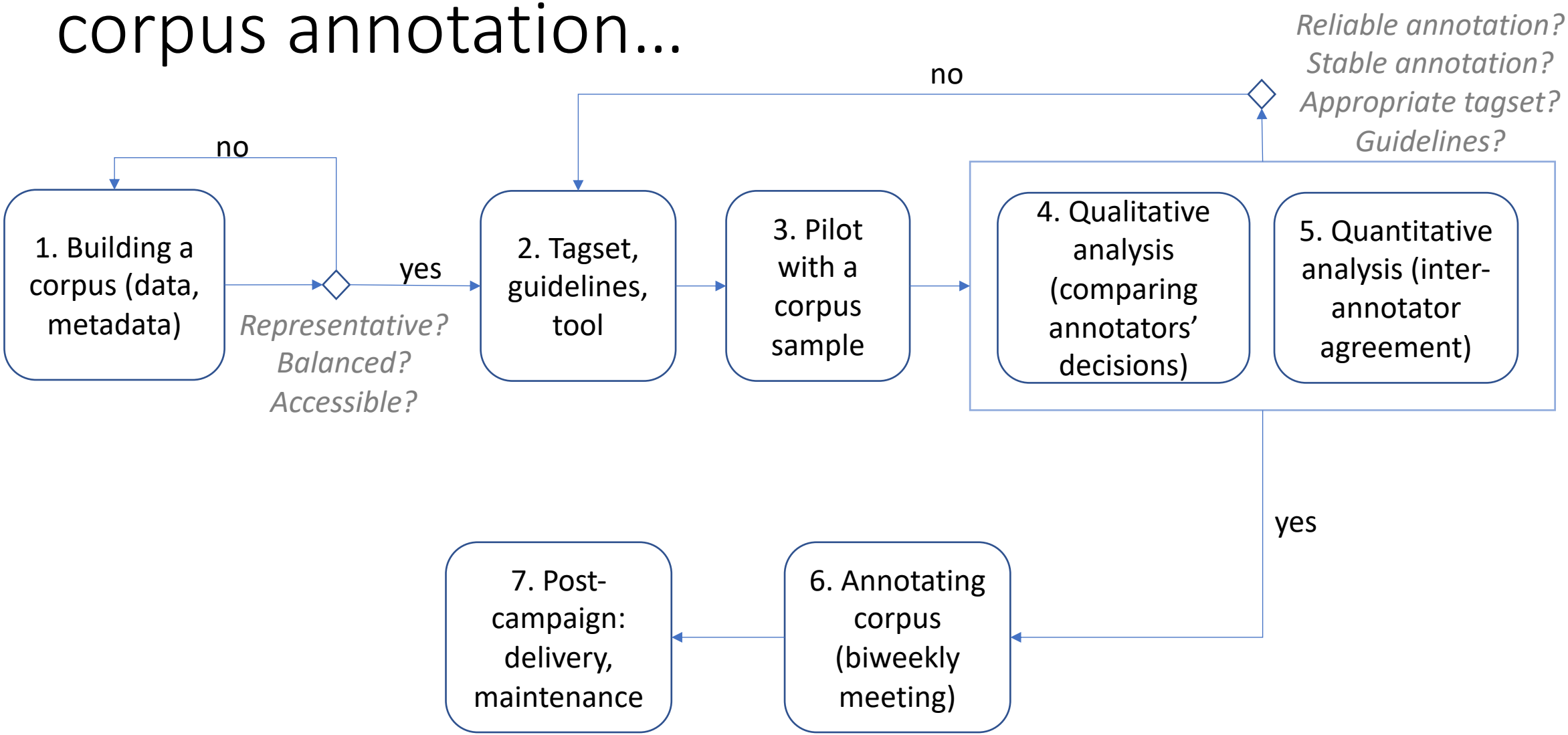
(you get what you annotate)

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

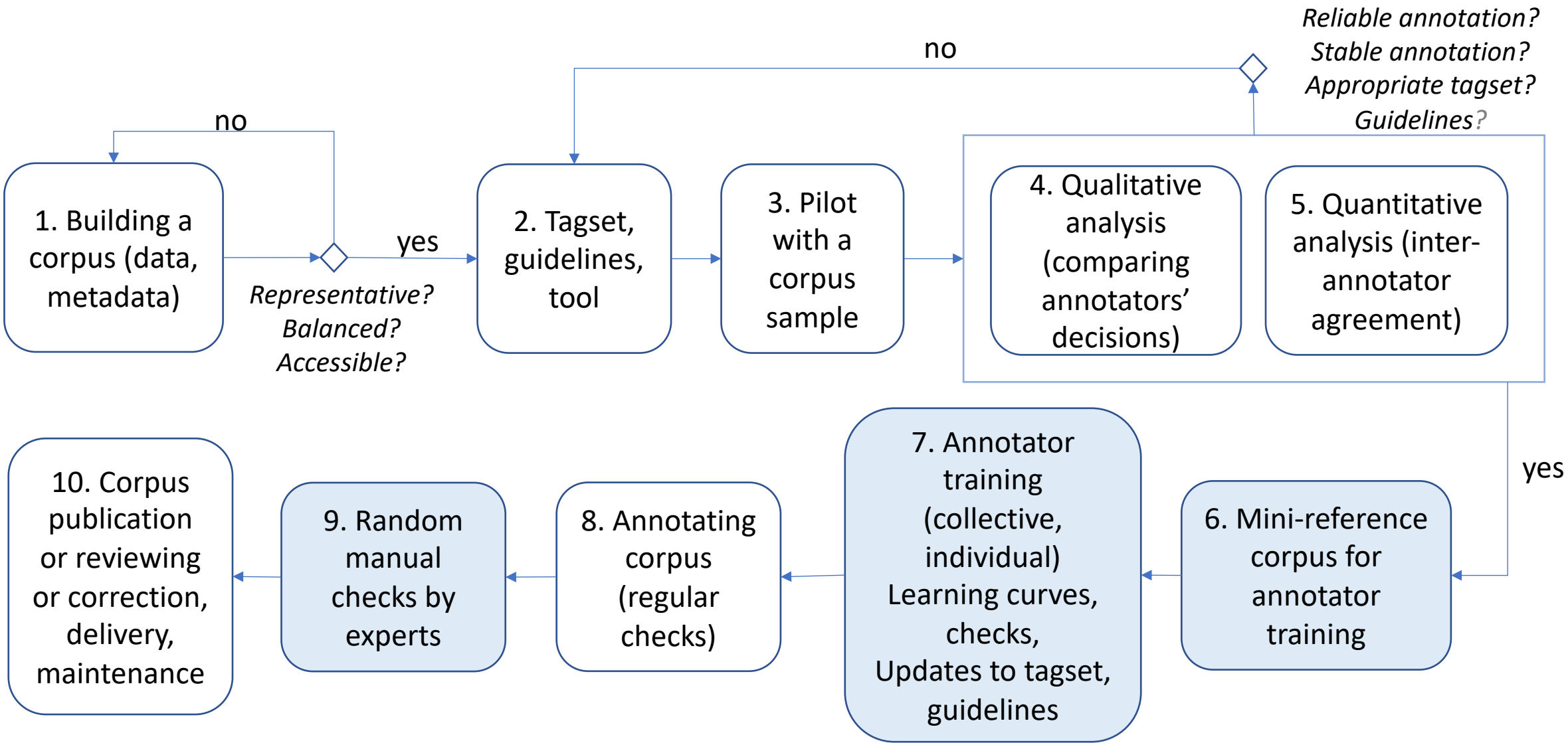
→

Gold standard corpus

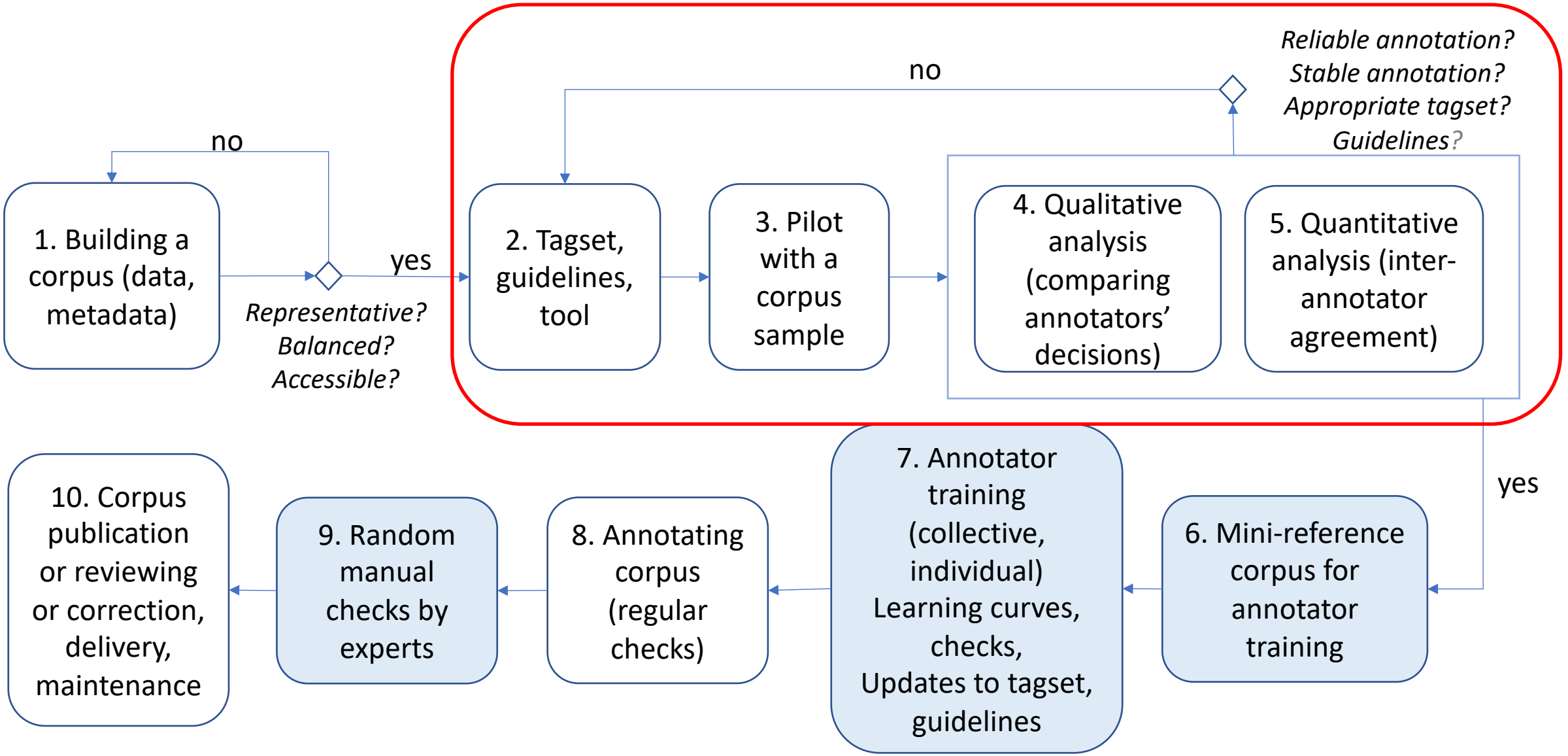
Hovy et al. 2010. Towards a ”Science” of corpus annotation…



Fort. 2016. Collaborative annotation…



Fort. 2016. Collaborative annotation…



GÖTEBORGS UNIVERSITET

Annotation - choosing a taxonomy

- Other taxonomies :
  - ASK : 23 Error types

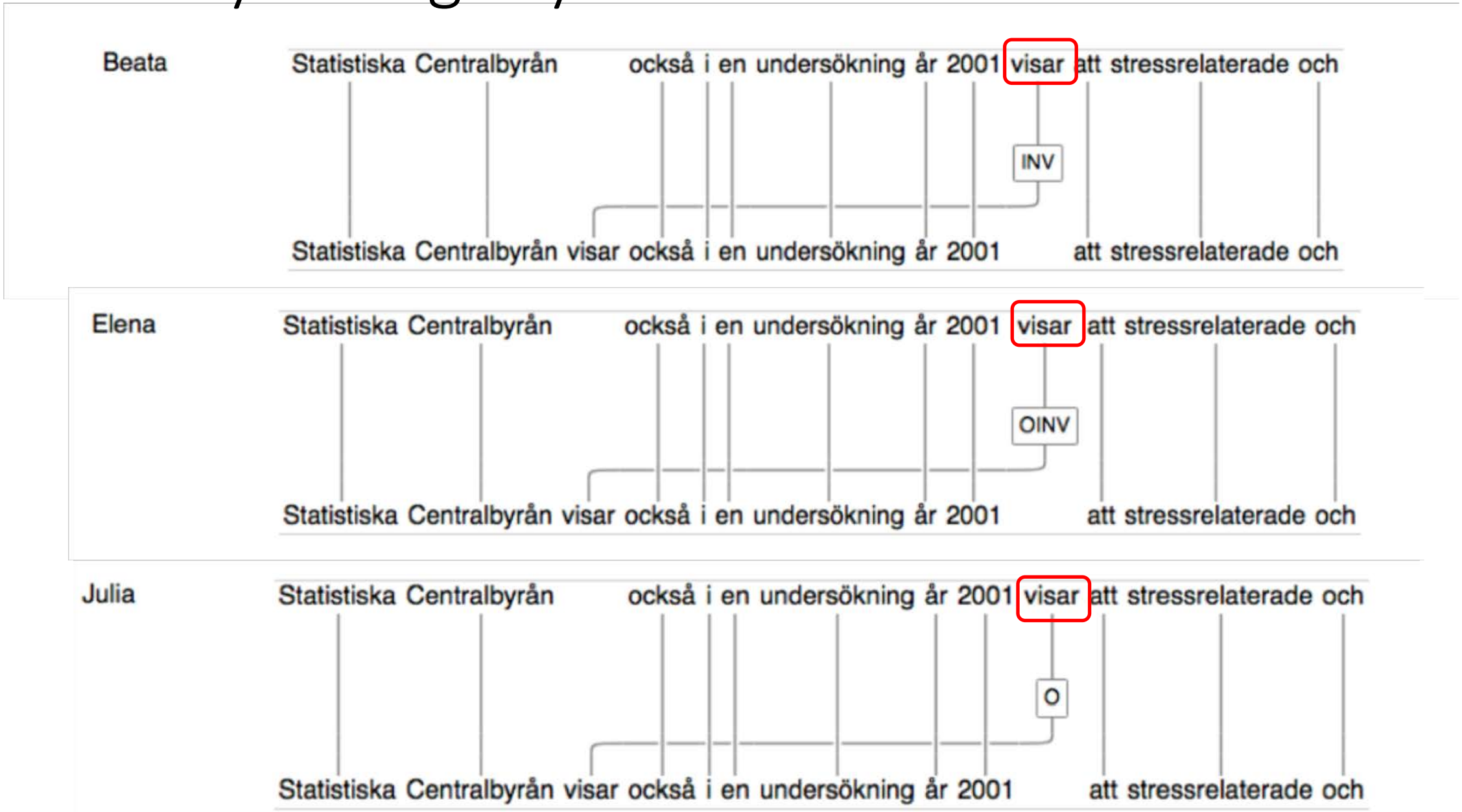
    Lexical (8), morphological (3), syntactical (7), punctuation (4), unidentified (1)

  - MERLIN: 64 error types

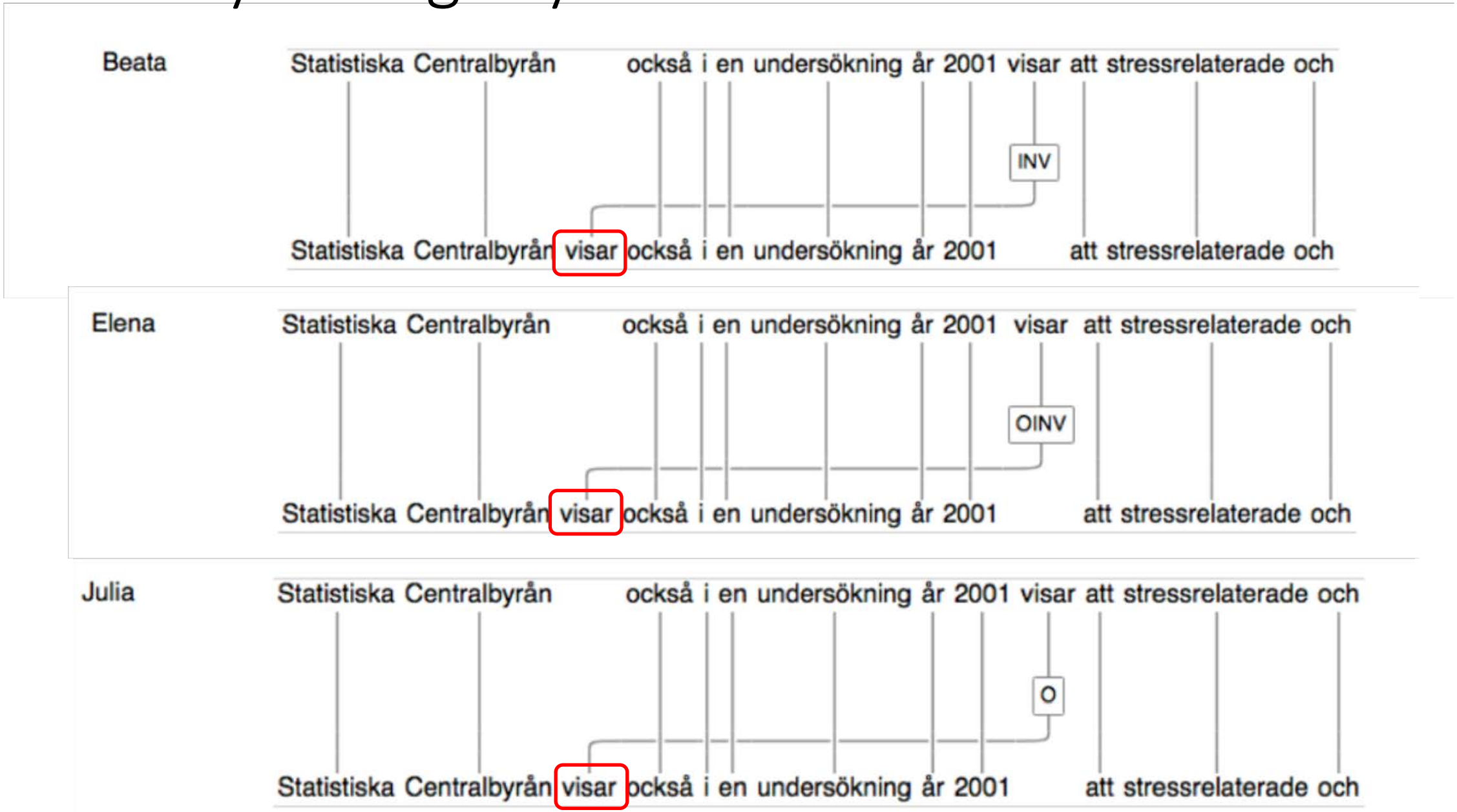
    grammar (21), orthografic (8), intelligibility (8), vocabulary (10), coherence (4), sociolinguistic (10), pragmatics (3)

- How detailed should the taxonomy be?
- How important is the target language?
  - similarity between Norwegian and Swedish
  - Comparability between ASK and SweLL wanted

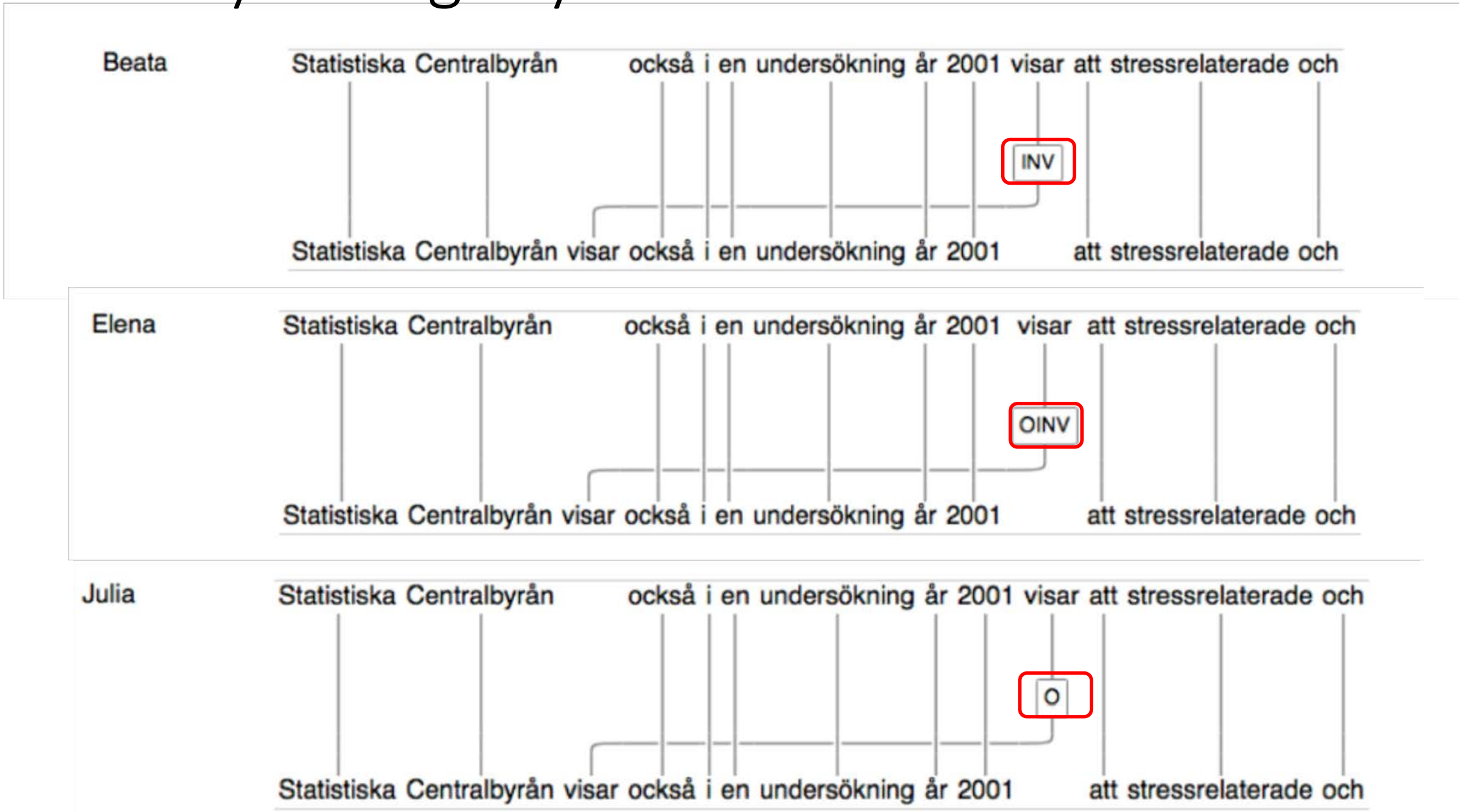
Taxonomy ambiguity



Taxonomy ambiguity



Taxonomy ambiguity



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.

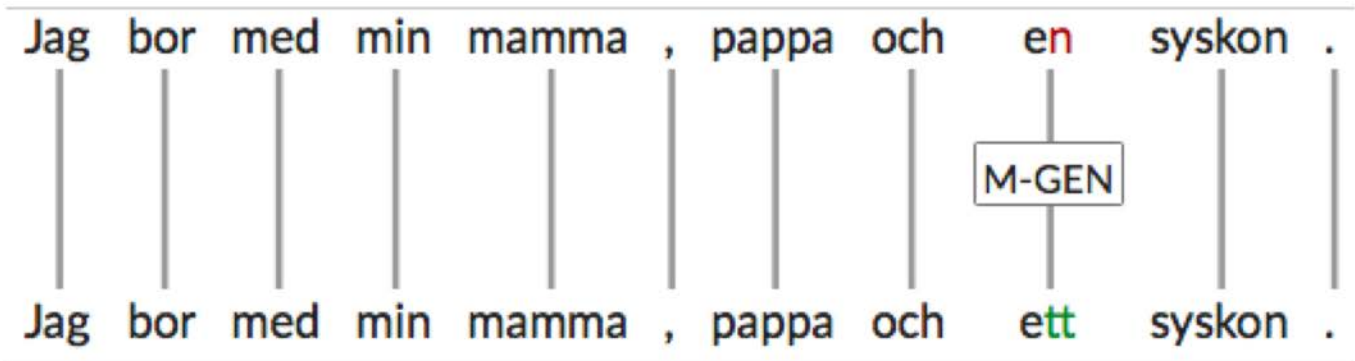
Potential target hypotheses:

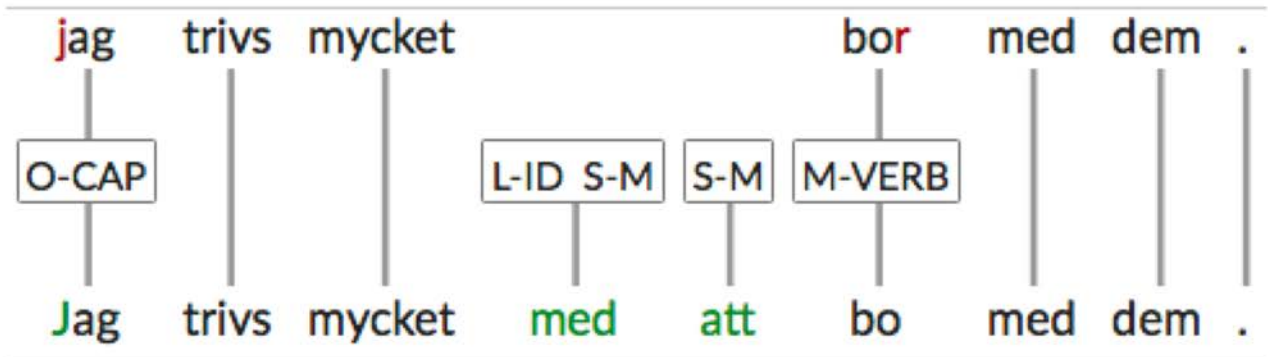
Jag trivs mycket **bra** med dem → Minimal change (seemingly)
→ Error: wrong word / spelling?

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





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