

SVALA:

Pseudonymizer service for Swedish L2 essays

Elena Volodina, Samir Ali Mohammed, Arild Matsson

RIKSBANKENS JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

GÖTEBORGS UNIVERSITET

SWE-CLARIN



SVALA:

SVA = SVenska som Andraspråk, Linking & Annotation

Pseudonymizer service for Swedish L2 essays

Elena Volodina, Samir Ali Mohammed, Arild Matsson



RIKSBANKENS JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING





SweLL –

research infrastructure for Swedish as a Second Language

RIKSBANKENS JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

GÖTEBORGS UNIVERSITET

SWE-CLARIN

Swedish Learner Language

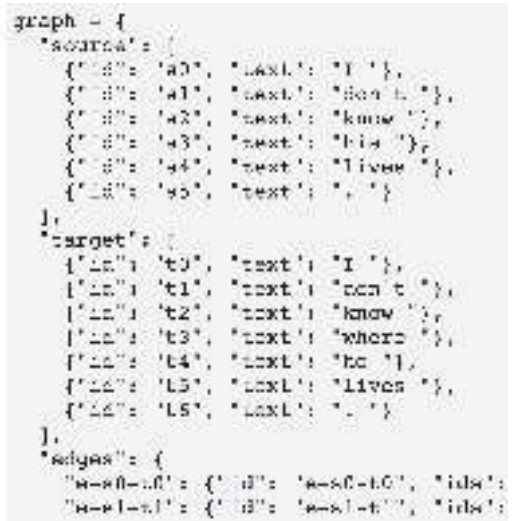
SweLL –

research infrastructure for Swedish as a Second Language

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 correction annotation & manually checked linguistic annotation
   - Make available for research (and public?)

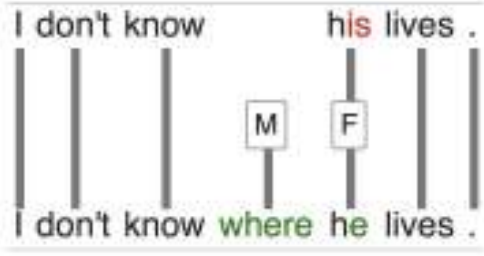




SweLL promises (main)

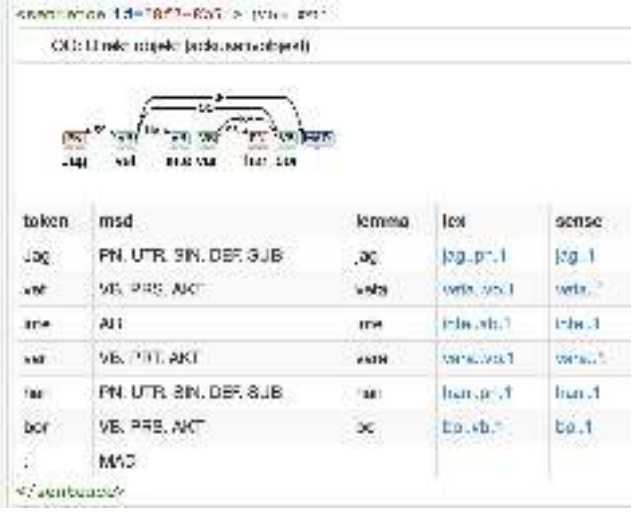
2. Set a platform (and workflow) for
   - Continuous upload of new essays
   - Manual correction annotation
   - Automatic linguistic annotation





I don't know his lives .

I don't know where:M he~his:F lives



SweLL promises (main)

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



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 förut. 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 leker ute tillsammans har det goda livet enligt mig. Mat är en av viktigaste saker angående det goda livet. Man måste välja väldig noga sin mat. Det är lätt att vara nöjd med halvledig mat vilket man lagga snabbt. Att lagga riktig nyttig mat tar mycket tid och man bör förbereda sig. Jag menar att jag måste köpa färska grönsaker om jag ska lagga någon nyttig mat. Jag menar med mat försäker man att äta hälsosamt och undvika farma, 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 halsosamt liv utan stress.** Å andra sida är jag inte säkert att det är möjligt i ett modernt samhälle leva detta liv. I dagens samhället är 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 lyckliga familjen på ett bord.

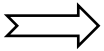
Nu tillbaka till Europa och Sverige. Har har människorna andra problem. **Stress, långa**

A truly interdisciplinary effort

- SLA
- NLP
- Software engineering

A truly interdisciplinary effort

- SLA
- NLP
- Software engineering

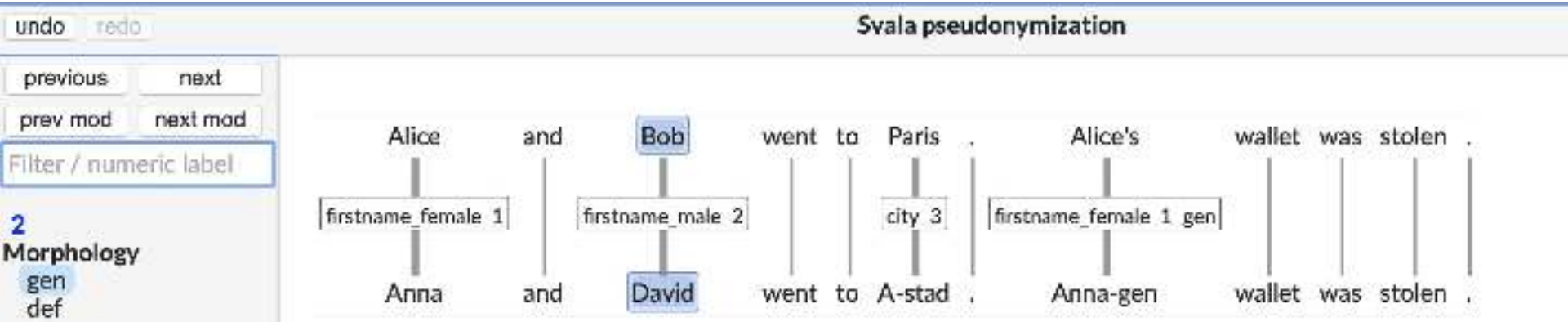


Effects on SVALA

- Methodology of annotation
- Design of SVALA
- Format
- Terminology

Methodology

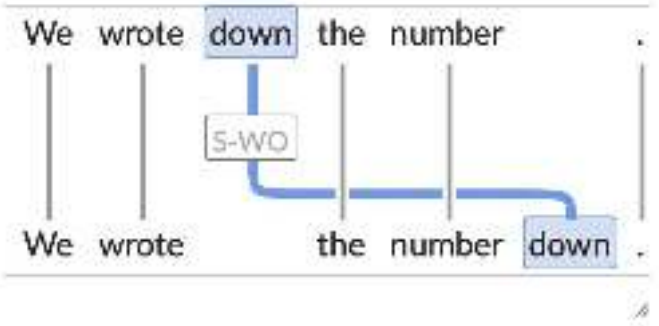
- Anonymization in 3 steps:
  - detection
  - placeholder labelling
  - pseudonymization
- Normalization separate from correction annotation



Design

- As user-friendly (for the target user group) as possible:
  - No xml tags
  - Work in a ”familiar” environment (i.e. text editor)
  - Visual support
  - Support for word-order changes





Format

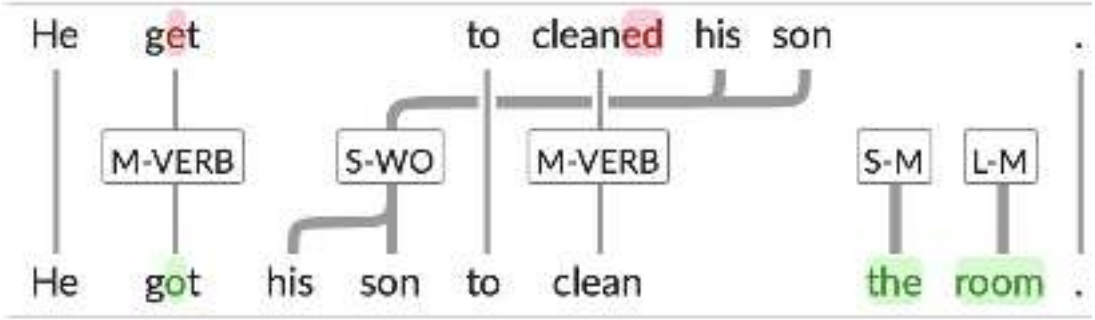
- JSON format
- 3 separate ”objects”:
  - for original text
  - normalized text
  - links with labels


```
{
  "source": [
    {"id": "s0", "text": "We "},
    {"id": "s1", "text": "wrote "},
    {"id": "s2", "text": "down "},
    {"id": "s3", "text": "the "},
    {"id": "s4", "text": "number "},
    {"id": "s5", "text": ". "}
  ],
  "target": [
    {"id": "t0", "text": "We "},
    {"id": "t1", "text": "wrote "},
    {"id": "t2", "text": "the "},
    {"id": "t3", "text": "number "},
    {"id": "t4", "text": "down "},
    {"id": "t5", "text": ". "}
  ],
  "edges": {
    "e-s2-t4": {
      "id": "e-s2-t4",
      "ids": ["s2", "t4"],
      "labels": ["S-WO"],
      "manual": true
    },
```

Terminology

- Correction annotation (not error annotation!)

(1)



(2)



SVALA - SweLL annotation tool

- Parallel text
- Visualized diff
- Semi-automatic word alignment
- Annotation on source–target links

Desired:

- Drop-down menus for error codes
- Three-tier representation (original, spell-corrected, normalized)
- Support with automatic spelling error detection



Dan Rosén,
research engineer



Arild Matsson,
research engineer

SVALA - automatic pseudonymizer service

- Why?
  - GDPR
  - Ethical reasons
  - To speed up annotation work
  - To boost essay collection