SVALA - automatic pseudonymizer service

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

SVALA - automatic pseudonymizer service

- 8 head categories, 40 subcategories, morph. markers
  - names, geo names, institutions, transportation, age, date, miscellaneous, mark/sensitive

1. ORIGINAL TEXT →@PLACEHOLDER →RENDERING
2. ORIGINAL TEXT →@PLACEHOLDER →REPLACEMENT
3. ORIGINAL TEXT →@PLACEHOLDER →ORIGINAL

Guidelines: https://spraakbanken.github.io/swell-project/Anonymization_guidelines



Samir Ali Mohammed,
Systems developer



Arild Matsson,
research engineer

SVALA - pseudonymizer: facts

- Rule-based
  - tokenization, regular expressions
  - minor spelling correction (Levenstein distance)
- Git Repo: https://github.com/SamirYousuf/LR_project
- Python
- IN: text; OUT: json / text

- Three steps:
  - detection
  - placeholder labelling
  - pseudonymization



SVALA – hands-on demo

https://spraakbanken.gu.se/swell/dev/

Example essay (translation into English + mocking errors)

- I live in Stockholm on apartement . Jag är 29 . I live with my boyfriend . His name is Cezary . Apartement mine has a pattio and tree room . I enjoy there in Stockhulm but a lot of time to goto shop , fortifive minut . I have the buss and the Stockholm train . I lived in Danmark bifore , in Odense . It was less than Stockholm . I enjoy their too becaus I had more friends . I think it is hard to have friends here . But I enjoy better job here . In Odense jobbe I only on one website . In Stockholm I work on many website . I am webdevelooper . But Stockholm is closser to Luxembourg than Odense . It is important how one lives because I am not in my country . I mess my mother and my father but I live her with my boyfriend .

Future work

- Test on assistants → learn to incorporate automatically their corrections
- Integrate into Lärka and test on learners → collect their corrections into “reports”
- Adapt to English?
- Shared task?

Future work (for SVALA)

- Add (automatic) linguistic annotation and visualization
- Develop support for automatic error detection and labeling
- Make it available/modifiable for other languages/users (tagsets, versioning, server…)

GitHub repository

https://github.com/spraakbanken/swell-editor

A link to an article on SVALA

Wirén Mats, Arild Matsson, Dan Rosén, Elena Volodina. 2019. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. *CLARIN-2018 post-conference volume*. LiUP Press.

http://www.ep.liu.se/ecp/159/023/ecp18159023.pdf

A link to an article on Anonymization

Beáta Megyesi et al. (2018). Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish. *Proceedings of the 7th NLP4CALL workshop.* [pdf]

http://www.ep.liu.se/ecp/152/006/ecp18152006.pdf

A link to this presentation

http://tiny.cc/gximdz

Thank you!

- Questions?
- Comments?

undo redo

SVALA pseudonymization

show options

previous next

prev mod next mod

Filter / numeric label

1

Morphology
- gen
- def
- pl

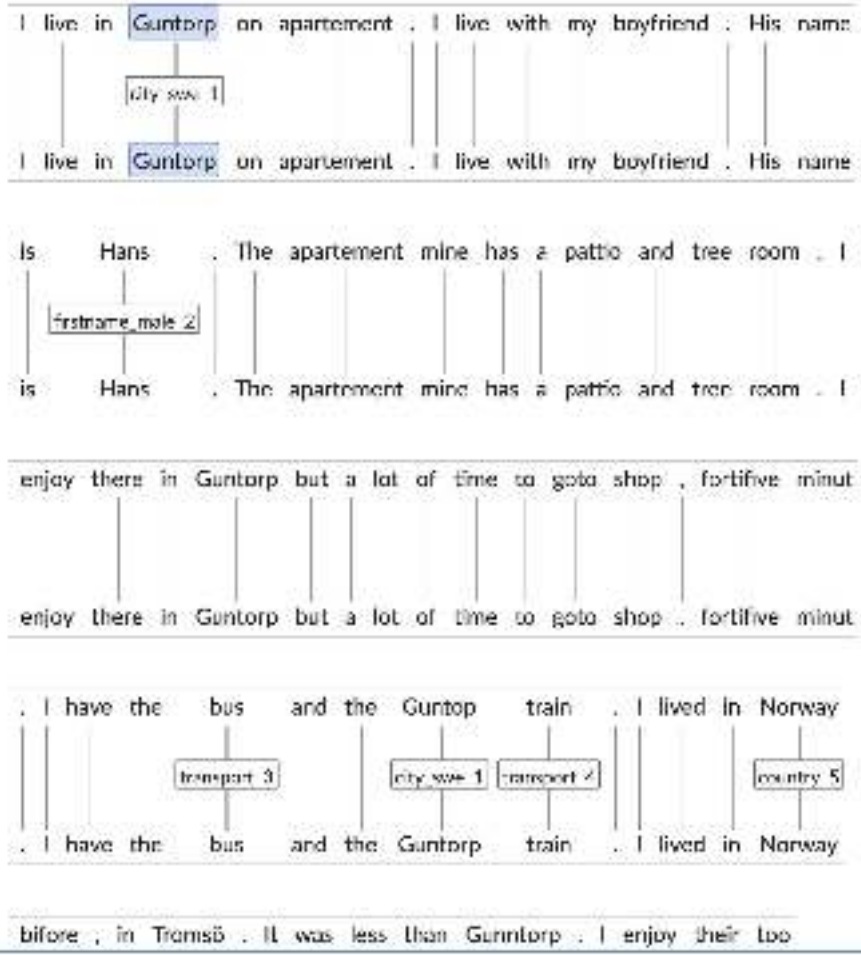
Names
- firstname_male
- firstname_female
- firstname_unknown
- surname
- middlename
- initials

Geographic data
- country_of_origin
- country
- zip_code
- region
- city_swe
- city
- area
- place
- geo
- street nr

Institutions
- institution
- school
- work
- other_institution

Transportation
- transport
- transport line

Age
- age digits



Document comment:

1
- Guntorp
- Gunton
- Gunntorp
- Guntorp
- Guntorp

2
- Hans

3
- bus

4
- train

5
- Norway

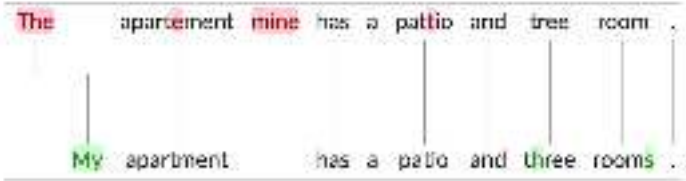
6
- Tromsö

SVALA repo issues

target text:

I live in B-svensk-stad in an apartment . I live with my boyfriend . His name is Peter . My apartment has a patio and three rooms . I enjoy it there in B-svensk-stad but it takes a lot of time to go to the shop . forty-five minutes . I have the D-transport and the B-svensk-stad C-transport . I lived in C-land before , in D-stad . It was smaller than B-svensk-stad . I enjoyed it there too because I had more friends . I think it is hard to find friends here . But I enjoy a better job here . In D-stad I worked only on one website . In B-svensk-stad I work on many websites . I am a web developer . But B-svensk-stad is closer to B-land than D-stad . It is important how one lives because I am not in my country . I miss my mother and my father but I live here with my boyfriend .

Document comment:



undo redo

SVALA correction annotation

show options

previous next
prev mod next mod
group orphan
auto

Enter filter text

- L-M
- L-R
- L-REF
- L-W

Morphological

- M-ADJ/ADV
- M-CASE
- M-DEF
- M-F
- M-GEND
- M-NUM
- M-Other
- M-VERB

Punctuation

- P-M
- P-R
- P-W
- Sent Segmentation

Syntactical

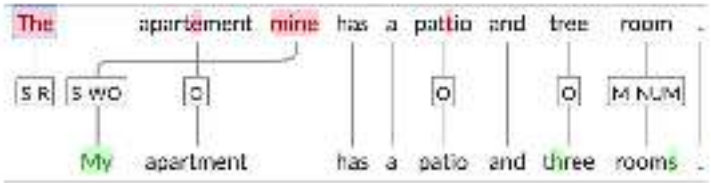
- S-adv
- S-CON
- S-finV
- S-M
- S-Msubj
- S-R
- S-W
- S-WO

Unidentified

- OBS!
- Uni

target text:

I live in B-svensk-stad in an apartment . I live with my boyfriend . His name is Peter . My apartment has a patio and three rooms . I enjoy it there in B-svensk-stad but it takes a lot of time to go to the shop , forty-five minutes . I have the D-transport and the B-svensk-stad C-transport . I lived in C-land before , in D-stad . It was smaller than B-svensk-stad . I enjoyed it there too because I had more friends . I think it is hard to find friends here . But I enjoy a better job here . In D-stad I worked only on one website . In B-svensk-stad I work on many websites . I am a web developer . But B-svensk-stad is closer to B-land than D-stad . It is important how one lives because I am not in my country . I miss my mother and my father but I live here with my boyfriend .




```
{
  "source": [
    {"id": "s18", "text": "The "},
    {"id": "s19", "text": "apartement "},
    {"id": "s20", "text": "mine "},
    {"id": "s21", "text": "has "},
    {"id": "s22", "text": "a "},
    {"id": "s23", "text": "pattio "},
    {"id": "s24", "text": "and "},
    {"id": "s25", "text": "tree "},
    {"id": "s26", "text": "room "},
```


Document comment:

M-NUM

- room—rooms

O

- apartement—apartment
- pattio—patio
- tree—three

S-R

- The—

S-WO

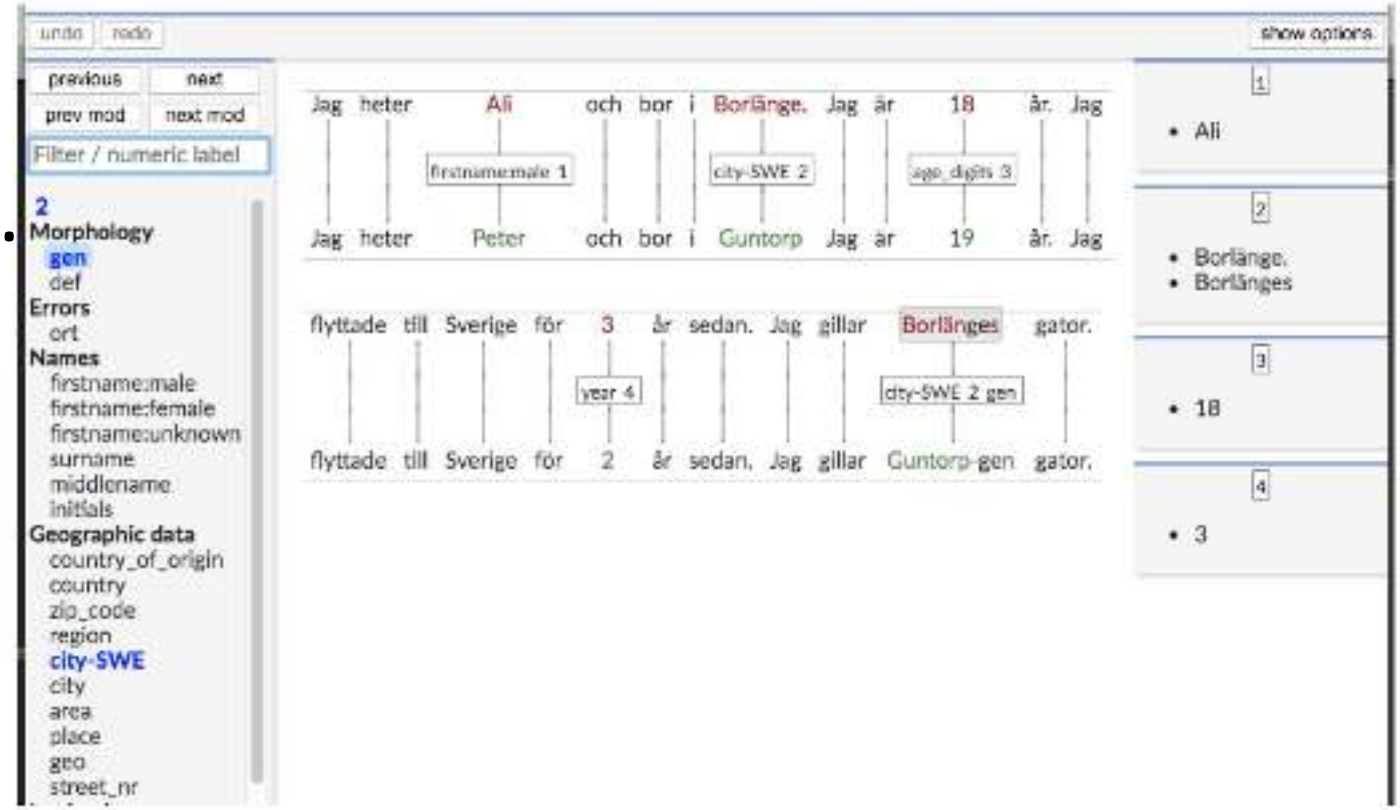
- mine—My

SVALA repo issues

SVALA pseudonym. mode

[Demo]





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
,
```