UPPSALA UNIVERSITET

RIKSBANKENS JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

SWE-CLARIN

2017-2019 (2020)

Pseudonymization in SweLL

Beáta Megyesi
Uppsala University

August 26, 2021

UPPSALA UNIVERSITET

Outline

- SweLL in a nutshell
- GDPR
- Anonymisation vs pseudonymisation
- Choices of learner metadata
- Pseudonymisation of learner text
- Conclusion

[Thanks Elena Volodina for some of the slides]

UPPSALA UNIVERSITET

SweLL

1. Deliver an annotated (gold standard) corpus of L2 essays:
    - 694 essays on various levels in the Swedish school system
    - 502 annotated and released in a Gold corpus
    - available as open source for research

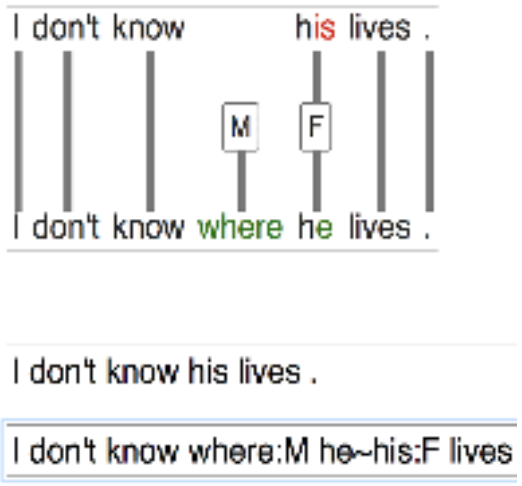


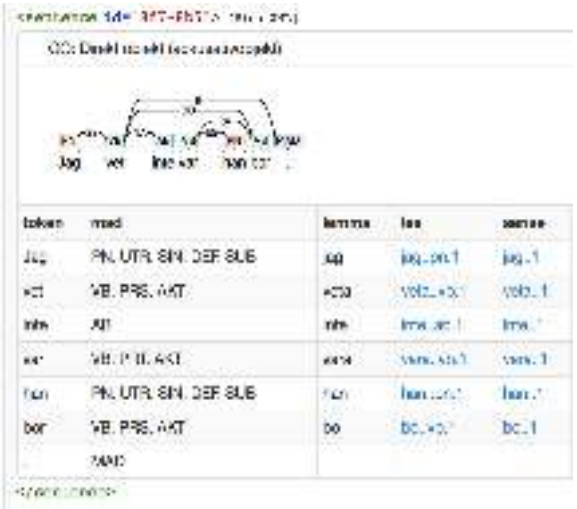


SweLL promises

2. Set a platform (and workflow) for uploading of new essays with metadata, error correction and annotation







SweLL promises

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





SweLL focus

- Adult healthy learners (16+ years)
  - no consent from caregivers required
- Written essays - no speech data
- Various proficiency levels of Swedish as L2
- Collected in schools, adult education, and universities (SFI, Sva, Behörighetsgivande kurs, TISUS, CEFR)
- Where possible – longitudinal data



The Not So Secret life of a PI

UPPSALA UNIVERSITET

Data Accessibility

Data needs to be open outside the original project **a challenge**!

Need to reduce the risks of illegal or unethical use of the data before it can be made accessible…

1. legal and ethical aspects of essay collection;
2. data handling flow;
3. principles of de-identification and pseudonymization;
4. tools and platforms for ensuring the previous steps.

General Data Protection Regulation (GDPR) announced



- Restrictions on use of personal information to protect ”subjects”, i.e. physical persons
- Important consequences for learner corpora (L2) projects – IF we want data to be available for research…



UPPSALA UNIVERSITET



GDPR

- Restricts use of digital data containing personal information
- Hands back ownership of personal data to data subjects
- Imposes data protection **by design and by default** (Art 25)

UPPSALA UNIVERSITET



GDPR

Regulation EU 2016/6791

- Restrictions on use of personal information to protect ”subjects”, i.e. physical persons:

  Personal information: “Any information relating to an identified or identifiable natural person (’data subject’)”;

  ”An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person”.

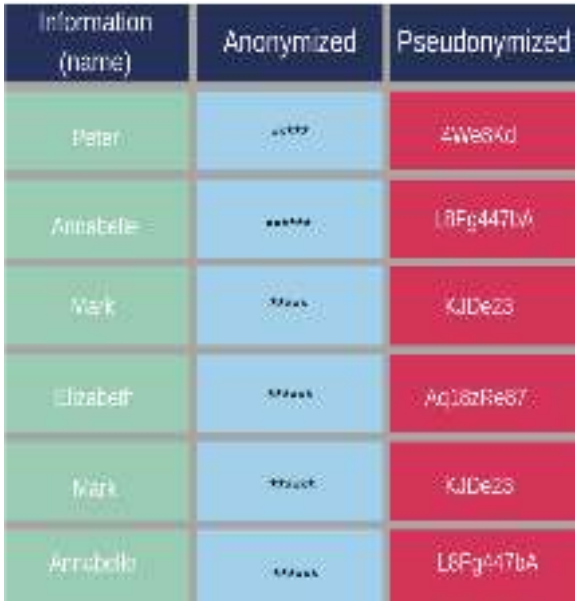
  (Article 4, EU GDPR)

UPPSALA UNIVERSITET

General Data Protection Regulation

GDPR

- Anonymization is not possible – subjects' right to withdraw participation
- Pseudonymization is recognized as a way to reduce risks of re-identification of data subjects (Recital 28)



https://www.linkedin.com/pulse/anonymization-does-work-big-data-due-lack-protection-direct-lafever

UPPSALA UNIVERSITET

Pseudonymization

- detection & neutralization of personally identifiable information (PII)
- predominantly manually handled
- automatic approaches are based on:
  - the recognition of named entities (NER),
  - often rule-based by dictionary lookup
- most studies in the medical domain
- no proper guidelines and lack of systematic studies for learner corpora when the project started

UPPSALA UNIVERSITET

SweLL and GDPR

How can we protect our subjects?

What are we supposed to pseudonymize?

SweLL & GDPR





- Collection precautions
- Signed agreements
- Name – ID mapping keys handling
- Data handling flow
- Portal development and data access
- Metadata precautions
- Text de-identification
- Pseudonymization
- Data usage scenarios

Model from *Building digital trust: The role of data ethics in the digital age,* Accenture Labs (June 13, 2016)

SweLL & GDPR





- Collection precautions
- Signed agreements
- Name – ID mapping keys handling
- Data handling flow
- Portal development and data access
- **Metadata precautions**
- **Text de-identification**
- **Pseudonymization**
- Data usage scenarios

Model from *Building digital trust: The role of data ethics in the digital age,* Accenture Labs (June 13, 2016)

De-identification through metadata

Stripp off common identifiers to a person: the learner and his/her text



Provide important information for research purposes about the learner and (the manipulated tex)

Identifiers: names, age, geographic places, dates, telephone numbers, e-mail, addresses, personal web-URLs, internet protocol addresses, and any unique identifiers such as social security numbers, account numbers, or vehicle identifiers

UPPSALA UNIVERSITET

Metadata

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

Metadata

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

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

No information about the actual school

No exact date for entering the L2 country

No information about school or teacher

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

UPPSALA UNIVERSITET

Categories PI

Hard replacement	Placeholder	Sensitive markup (keep original)
Age	Geo-data: *country, region, city, street, area*)	Education
Bank account		Profession
License numbers *(cars)*		Family members
Dates	Institution: *school, work, institution, other*	
E-mail		Ethnical info
Phone number		Political views
Social security number	Personal names: *female, male, neutral, surname*	Religious views
Url		Sexual info
Zip-code		

Pseudonymization of texts

1. **Identification**
2. **Annotation** - PLACEHOLDER
   - token-based
   - named entities of same type with same reference - running number, starting with 1.
   - If the particular word is repeated in the text, the same running number is assigned to it.
3. **Pseudonymisation**

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

Jag heter Hashem Mohammed Muhammed.

min mamma heter Ruqia

min pappa heter Muhamed.

->

Jag heter <firstname1-m> <middlename1> <surname1>.

min mamma heter <firstname2-f>

min pappa heter <firstname3-m>.

->

Jag heter Muhammed A. Khalil.

min mamma heter Leila

min pappa heter Ahmed.



Guidelines: https://github.com/spraakbanken/swell-project/blob/master/Anonymization_guidelines.md

Rendering

- Applied to the information that can be collected from general resource lists

E.g.: personal names and surnames; city and country names, nationalities and languages; geographic names (lakes, mountains, regions, etc.); street names, names of schools, institutions, work places; etc.



Efternamn och antal bärare för den folkbokförda befolkningen 31 december 2017

Efternamn med minst tio bärare	Antal bärare
Andersson	234 094
Johansson	233 859
Karlsson	125 958
Nilsson	162 100
Eriksson	125 888
Larsson	117 130
Olsson	100 692
Persson	95 607
Svensson	95 594
Gustafsson	57 290
Pettersson	50 998

Country	Continent
Afghanistan	Asia
Albania	Europe
Algeria	Africa
Angola	Africa
Argentina	South America
Australia	Oceania
Austria	Europe
Azerbaijan	Europe
Bahrain	Asia
Bangladesh	Asia
Belarus	Europe
Belgium	Europe
Bolivia	South America
Bosnia and Herzegovina	Europe
Botswana	Africa

Replacement



- Used for entity types where general resource lists do not exist
- Numerical expressions, dates
  - pattern used in the original, preserving all delimiters
  - date: 2018/01/01 -> 1111/11/11
  - phone numbers: 089-777-654-22 -> 000-000-000-00;
- Middle names/initials M/B/C… -> A
- Age: random number from the range of +/- 2
  - Digits: 18 16-20 -> 17

[ORIGINAL] My older sister is **THIRTY** and my younger sister is ***EITY**.
[CORRECTION] My older sister is **THIRTY** and my younger sister is **EIGHTEEN** (or **EIGHT** ?).
[@PLACEHOLDER + RANGE] My older sister is **@AGE STRING(28-32)** and my younger sister is **@AGE STRING(16-20)**.
[RANDOM REPLACEMENT] My older sister is **@AGE STRING(28)** and my younger sister is **@AGE STRING(20)**

UPPSALA UNIVERSITET

Original

No clearcut whether the information may be considered risky to keep or not.

- professions: I am a **web developer**.
- education: I am taking courses in **Linguistics**.
- number of siblings or family members: I have **five** sisters and **three** brothers.
- political or religious views:

  We were happy to participate in a demonstration against **Erdogan**.

Mark up the segment but do not take any actions until further notice.

previous | next

prev mod | next mod

Filter / numeric label

Morphology
- gen
- def
- pl

Names
- firstname_male
- firstname_female
- firstname_unknowr
- initials
- middlename
- surname

Geographic data
- foreign
- area
- city
- country
- geo
- place
- region
- street_nr
- zip_code

Institutions
- school
- work
- other_institution

Transportation
- transport_name
- transport_nr

Age
- age_digits
- age_string

Dates
- date_digits
- day
- month_digit
- month_word
- year

Misc
- account_nr

Source text:

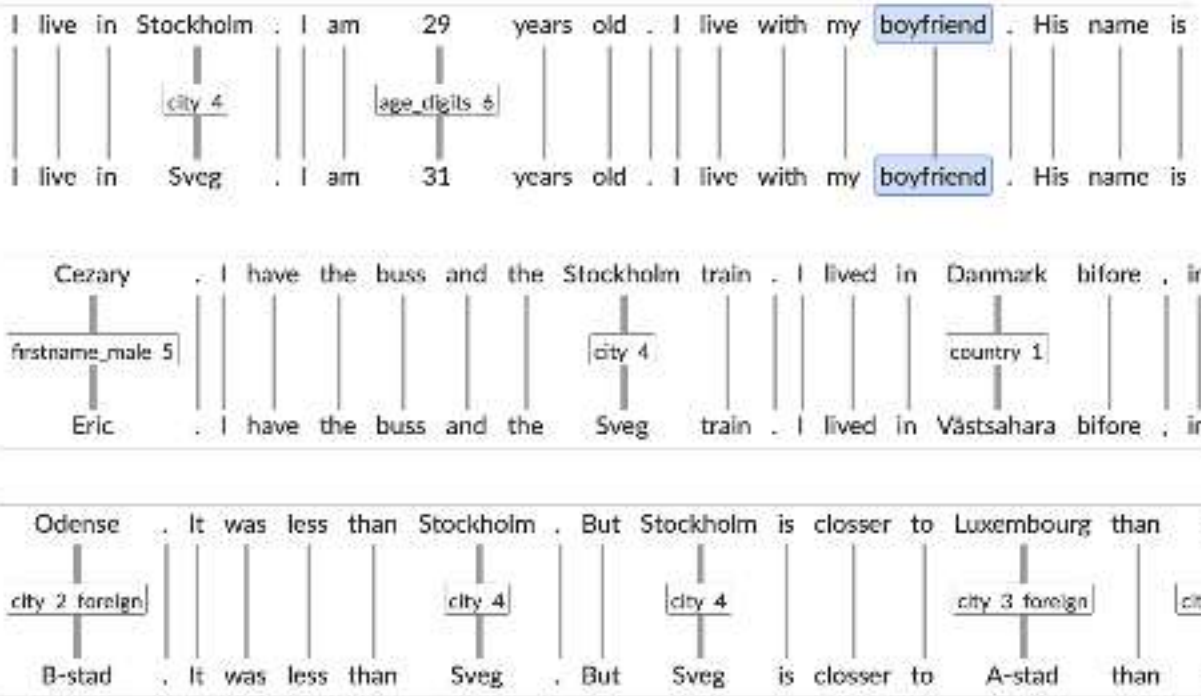
I live in Stockholm . I am 29 years old . I live with my boyfriend . His name is Cezary . I have the buss and the Stockholm train . I lived in Danmark bifore , in Odense . It was less than Stockholm . But Stockholm is closser to Luxembourg than Odense .

copy to target

Target text:

I live in Sveg . I am 31 years old . I live with my boyfriend . His name is Eric . I have the buss and the Sveg train . I lived in Västsahara bifore , in B-stad . It was less than Sveg . But Sveg is closser to A-stad than B-stad .

Spaghetti mode disable



Document comment:

1
- Danmark

2
- Odense
- Odense

3
- Luxembourg

4
- Stockholm
- Stockholm
- Stockholm
- Stockholm

5
- Cezary

6
- 29

UPPSALA UNIVERSITET

SVALA - SweLL annotation tool

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



Dan Rosén, research engineer



Arild Matsson, research engineer



Samir Ali Mohammed, systems developer

Automatic pseudonymizer

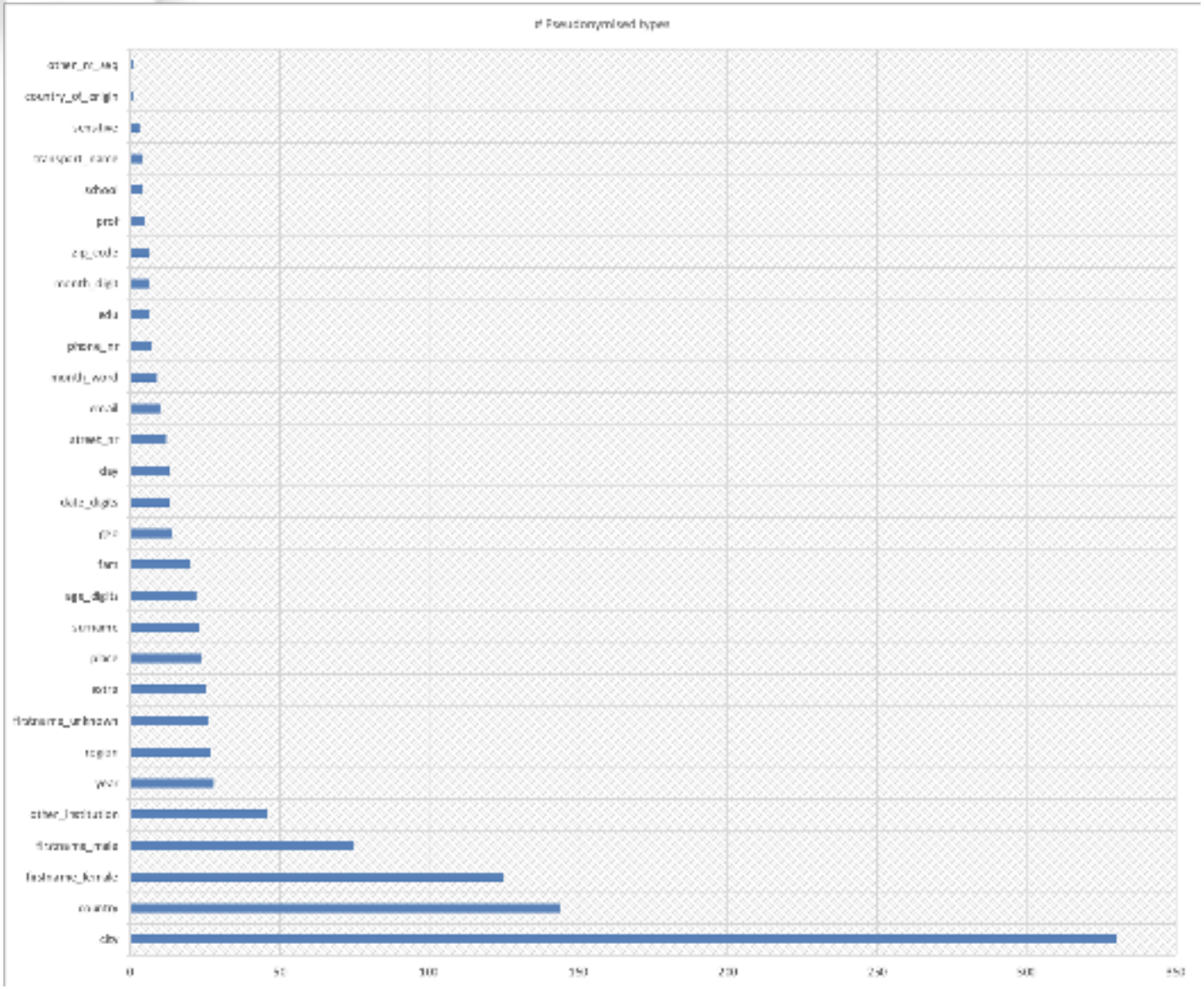
- SweLL automatic pseudonymizer service in SVALA
  - To speed up annotation work
  - For consistent pseudonymization
- Rule-based
  - regular expressions
  - minor spelling correction (Levenshtein distance)
- Python
- IN: text; OUT: json / text
- Three steps:
  - detection
  - placeholder labelling
  - pseudonymization



Git Repo: https://github.com/SamirYousuf/LR_project

UPPSALA UNIVERSITET

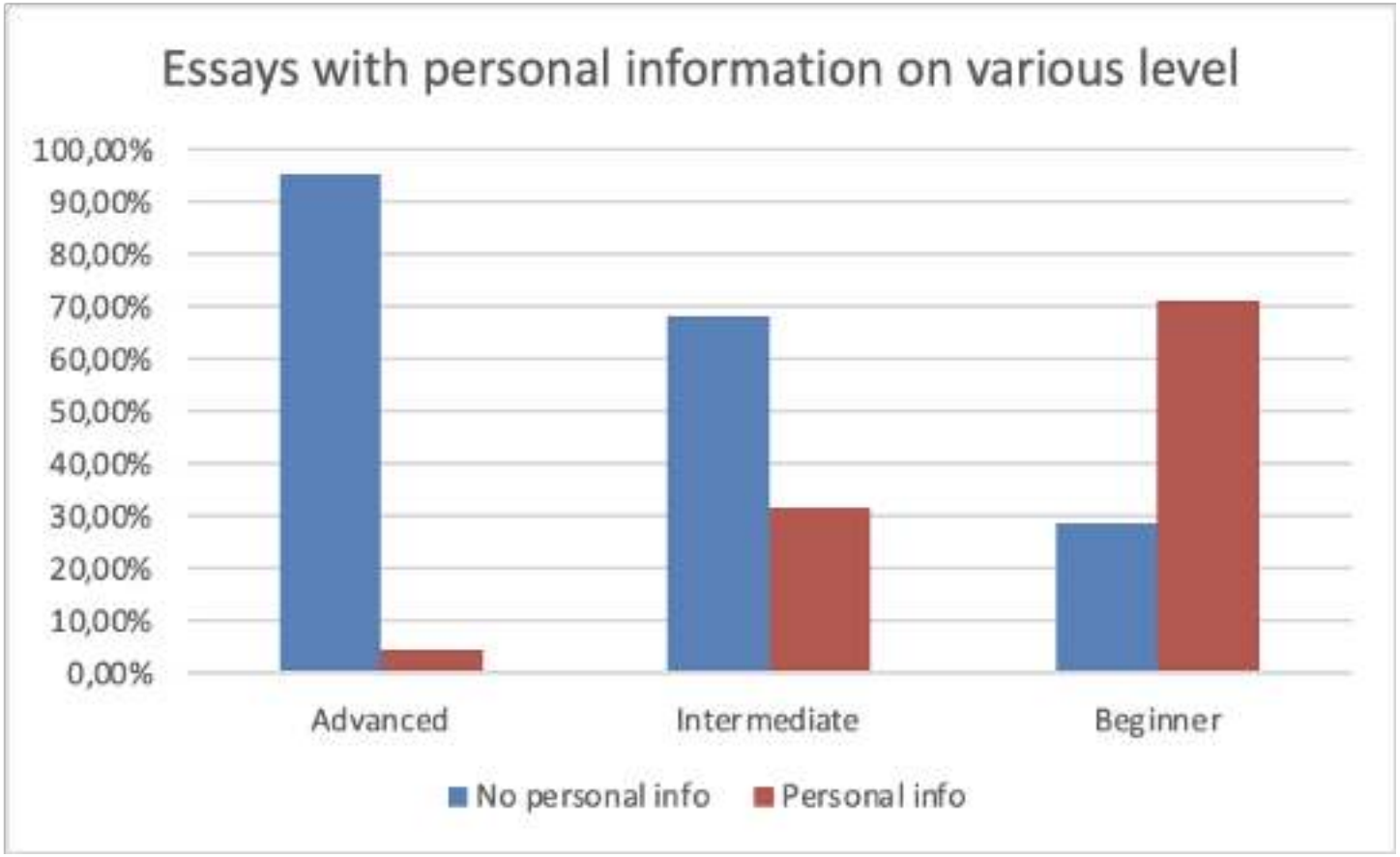
Stats



- 502 essays
- 1029 entities containing personal information

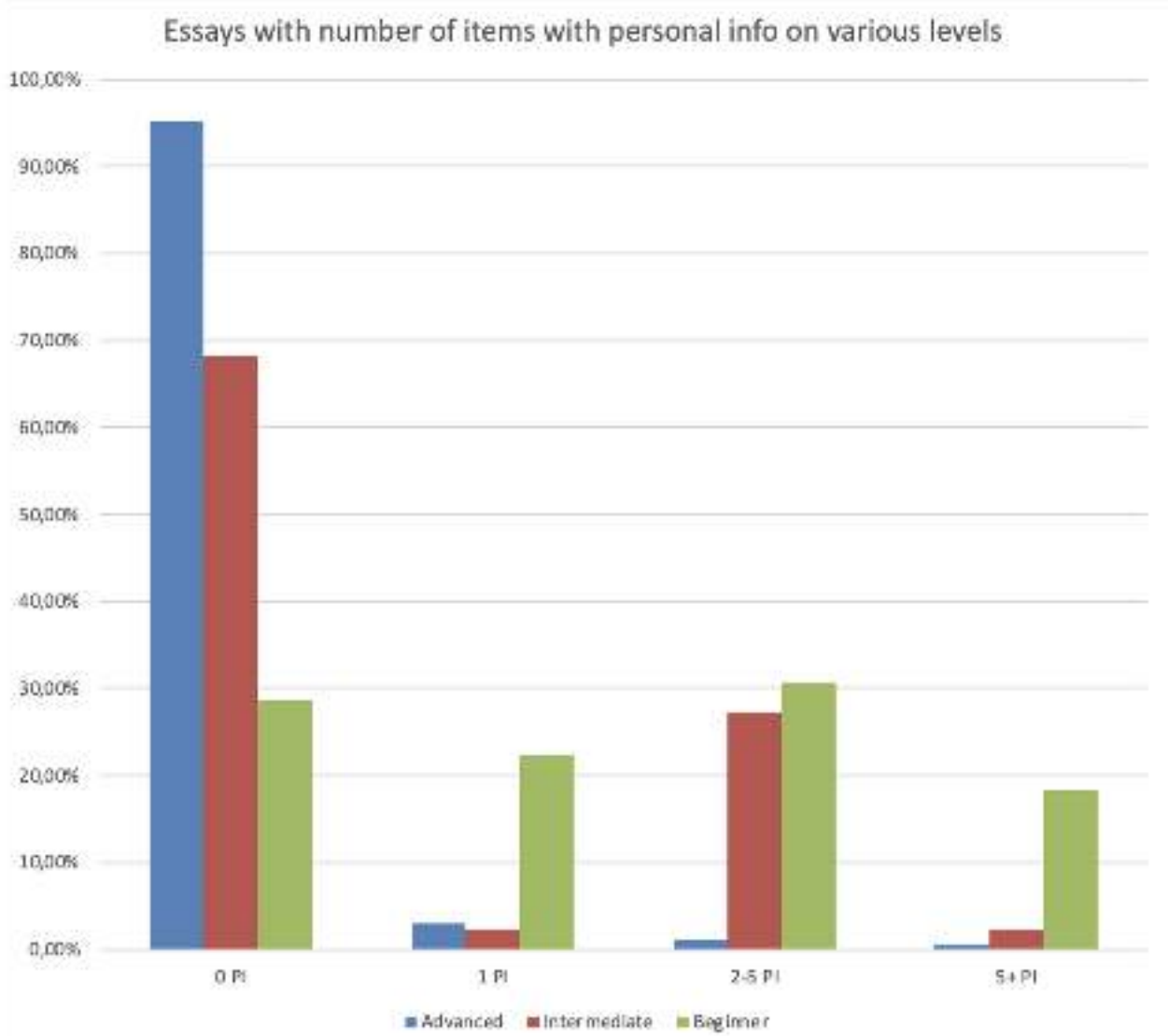
UPPSALA UNIVERSITET

Stats



UPPSALA UNIVERSITET

Stats



Some problems



- Keep the general flow of the text;
- Lost of misspellings in the replaced items;
- Wrong prepositions:

  I worked **in** Charing Cross Hospital

  I worked in @workplace

  I worked **in** Volvo
- Error propagation due to missing annotation labels:

  Alice's wallet was stolen

  @female name wallet was stolen

  Jane wallet was stolen
- Common knowledge:

  I lived in Berlin, the capital of Venezuela.

UPPSALA UNIVERSITET

Conclusion

- SweLL focuses on the pseudonymization of learner essays.
  - the legal issues influencing the way data needs to be handled and manipulated to ensure de-identification of subjects,
  - taxonomy for pseudonymization,
  - method for pseudonymization,
  - tool for pseudonymization.

UPPSALA UNIVERSITET

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