Towards Privacy by Design in Learner Corpora Research: A Case of On-the-fly Pseudonymization of Swedish Learner Essays

Elena Volodina, Yousuf Ali Mohammed
Arild Matsson, Sandra Derbring
University of Gothenburg, Sweden
name.surname@gu.se

Beáta Megyesi
Uppsala University, Sweden
name.surname@lingfil.uu.se

Abstract

This article reports on an ongoing project aiming at automatization of pseudonymization of learner essays. The process includes three steps: identification of personal information in an unstructured text, labeling for a category, and pseudonymization. We experiment with rule-based methods for detection of 15 categories out of the suggested 19 (Megyesi et al., 2018) that we deem important and/or doable with automatic approaches. For the detection and labeling steps, we use resources covering personal names, geographic names, company and university names and others. For the pseudonymization step, we replace the item using another item of the same type from the above-mentioned resources. Evaluation of the detection and labeling steps are made on a set of manually anonymized essays. The results are promising and show that 89% of the personal information can be successfully identified in learner data, and annotated correctly with an inter-annotator agreement of 86% measured as Fleiss kappa and Krippendorff's alpha.

1 Introduction

Access to language data is an obvious prerequisite for research in digital humanities in general, and the development of NLP-based tools in particular. However, accessible data becomes a challenging target where personal data is involved. This is very true of language learner data where tasks are often phrased so that they — directly or indirectly — elicit explicit personal information, e.g. "Describe your school" or "Introduce yourself".

The recent public debate — starting with Edward Snowden's revelations of US government's abuse of personal integrity — has led to important changes in European legislation (Encinas et al., 2015; ENISA, 2017; ENISA, 2018) as well as in attitudes towards sharing and collection of data online.

The General Data Protection Regulation (GDPR) has come timely in relation to this debate. GDPR is a legal European regulation restricting the use of digital data containing personal information. Article 4 in GDPR defines personal information as *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* (EU Commission, 2016, Art.4).

Among others, the GDPR focuses on handing back the ownership of personal data from software providers to private people (data subjects). However, the risks of misuse may linger despite the legislation, and the best protection would be either *not* to provide any personal information at all (which often is impossible), or to make sure that the software implementation encrypts, masks, hides or totally prevents any personal information to enter servers (thus preventing its potential unauthorized exploitation). Technology that could safeguard data subjects in that respect — that is, various de-identification/pseudonymization techniques, alongside with encryption, authorization, data minimization etc. — are recommended to be built into the software from the start, ensuring that the software complies with the requirement of *data protection by design and by default* (EU Commission, 2016, Art.25).

Proceedings of the 28th International Conference on Computational Linguistics, pages 357–369
Barcelona, Spain (Online), December 8-13, 2020

GDPR states that *anonymous data*, that is, data that is de-identified in such a way that no re-identification is possible, falls outside the scope of the GDPR (EU Commission, 2016, Recital 26). However, no data is truly anonymous and is said to be an unattainable target (Rocher et al., 2019). Pseudonymization, on the other hand, is recognized by the GDPR as one of the ways (and a requirement) to reduce risks of re-identification of a data subject (EU Commission, 2016, Recital 28). Pseudonymization is effective if data cannot be attributed to a specific data subject without use of any additional information. The additional information should, then, be kept separately from the rest of the data.

While the potential landscape of privacy-protecting techniques is huge (e.g., see an overview in Danezis et al. (2014)), our experiment focuses on a limited domain of learner corpora, zooming into one technique, namely, pseudonymization. In the absence of labeled pseudonymized data to apply data-intensive machine learning approaches, we choose to experiment with rule-based approaches to detect, label and pseudonymize information that we define as personal on a set of L2 Swedish essays. We use resources for names, geographic names, work and study places, etc. Some linguistic, common knowledge and certain task-related constraints are handled, such as polysemy between personal names and geographic names, and some morphological markers. Below we present the reasoning around the pseudonymization process, the first experiments, results and analysis. We view our experimentation with rule-based pseudonymization techniques as the first building block of a "privacy by design" platform for online essay collection.

2 Related work

There is not much published literature on the topic of de-identification of personal information in language data in general, and no detailed studies of anonymization/pseudonymization methods using NLP technology. De-identification of personal information through anonymization and/or pseudonymization got most attention in the medical domain where personal information is removed or masked in medical data sets to guarantee the anonymity of patients. However, few studies have been carried out on de-identification applied to other areas where language data is used.

Before the age of GDPR, one of the earliest and most comprehensive studies on anonymization was presented by Rock (2001). She gave an overview of anonymization methods, and legal rights, responsibilities, and obligations when using texts in corpora. Later Medlock (2006) presented a study on NLP and anonymization where he introduced a publicly-available benchmark corpus along with an interactive model for anonymizing data based on syntactic analysis and active learning. He defined anonymization as "the task of identifying and neutralizing sensitive references within a given document or set of documents". Following Medlock (2006), anonymization in current studies usually involves two distinct steps: first the text sequence containing personal information is identified, and then neutralized. Neutralization can be performed either by the replacement of the personal information with a placeholder, a category type of the personal information, or by another similar token belonging to the same category type.

Since the GDPR legislation, we have seen an increased interest in the NLP community to deal with automatic pseudonymization, see e.g. the recent workshop on NLP and Pseudonymisation (Ahrenberg and Megyesi, 2019). Still, most of the literature on the topic deals with medical data, e.g. Marimon et al. (2019), with available GDPR guidelines such as "Identifiability, anonymisation and pseudonymisation" published by the Medical Research Council (MRC, 2019).

3 Experiment setup

The experiment has been carried out within the SweLL project, aiming at building a digital infrastructure for research on Swedish as a second language (L2), including compilation of a digital corpus of essays (Volodina et al., 2018; Volodina et al., 2020). In connection to that, intensive work is ongoing with collection and manual annotation of essays written by adult L2 learners of Swedish, planned for release in 2021. One of the steps includes pseudonymization of essays, performed manually at this stage.

To experiment with the potential of automatic pseudonymization, we opted for a rule-based approach, which was motivated by the fact that very few manually pseudonymized essays were available at that moment (85 for the first evaluation, and 200 for the second) which excluded the application of data-

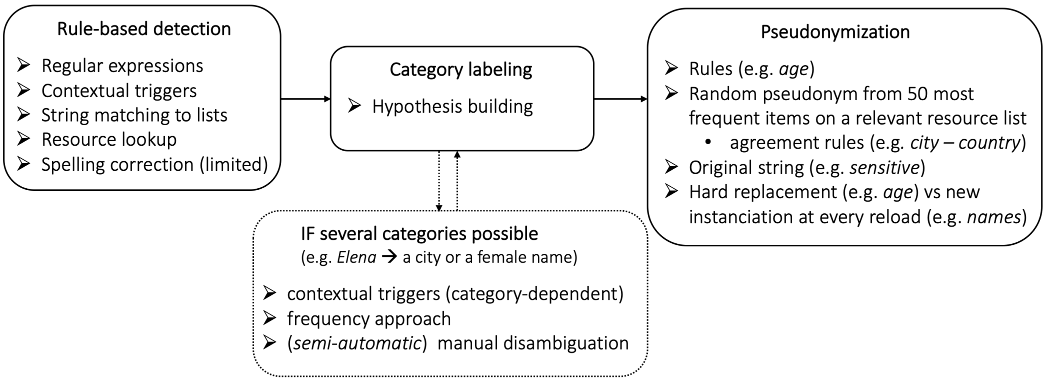

Rule-based detection
➢ Regular expressions
➢ Contextual triggers
➢ String matching to lists
➢ Resource lookup
➢ Spelling correction (limited)
Category labeling
➢ Hypothesis building
Pseudonymization
➢ Rules (e.g. *age*)
➢ Random pseudonym from 50 most frequent items on a relevant resource list
• agreement rules (e.g. *city – country*)
➢ Original string (e.g. *sensitive*)
➢ Hard replacement (e.g. *age*) vs new instanciation at every reload (e.g. *names*)
IF several categories possible
(e.g. *Elena* → a city or a female name)
➢ contextual triggers (category-dependent)
➢ frequency approach
➢ (*semi-automatic*) manual disambiguation

Figure 1: Steps in the pseudonymization process.

intensive machine learning approaches. However, we do not rule out a future possibility to get enough manually annotated data for training machine learning algorithms for this problem.

The automatic pseudonymization[1] is implemented in python and is split into three subsequent steps: (1) Detection of personal segments (Section 3.3), (2) Labeling of categories (Section 3.4) and (3) Pseudonymization (Section 3.5), see Figure 1 for an overview. During the *detection* step, regular expressions, contextual triggers, resource lookup and matching strings to lists are used. Minor spelling correction based on Levenstein distance is also applied to some categories. During the *category labeling* step a hypothesis is generated about which categories an identified segment can belong to, and where necessary, disambiguated using contextual triggers, information on part of speech, or frequency. During the *pseudonymization* step, the identified categories are either replaced with potential candidates of the same type, e.g. names – with new names; or are replaced with a standard variant representing the category, e.g. *url.com* for all urls. We associate individual rules with each of the 15 categories (and their sub-categories). To analyze, whether the rules could identify and effectively replace sensitive data points, we compare the detected segments and the labels in the manually annotated versions and ours, and report the accuracy. The pseudonymizer service has been incorporated into the demo version of SVALA, the SweLL annotation tool used in the project (Wirén et al., 2019; Rosén et al., 2018), for better visualization and manual inspection.

3.1 Pseudonymization categories

Earlier, the concept of anonymization was seen as binary: either data was seen as anonymous or not (Sweeney, 2000). Removing explicit identifiers, for example name, date of birth and telephone number, was considered sufficient. However, identifiability is relative and contextual, and seemingly impersonal data points – that on their own apply to many people – can, if taken together, identify people uniquely. Sweeney (2000) has shown that by providing date of birth, zip code and gender, 87% of the US population could be uniquely identified. When these data points appear in unstructured texts, e.g. in personal stories, blogs or language learner essays, the task of detecting and masking them becomes even more challenging.

While the GDPR lists pseudonymization as a method to mask a real person behind the data, as well as to show GDPR compliance with the requirements such as privacy by design, the major concern is which data points need to be manipulated, masked or removed to achieve an acceptable level of pseudonymization; as well as how the risks can be estimated and acceptable levels of protection ensured.

As a starting point, we focus on categories identified in the SweLL project (Megyesi et al., 2018) as summarized in Table 1. The categories build upon the previous work in the medical domain, namely *Health Insurance Portability and Accountability Act (HIPAA)* [2], and have been further enriched and modified within the SweLL project to fit into the context of (adult) second language learning.

[1] referred to *pseudonymizer service* within the project

[2] https://cphs.berkeley.edu/hipaa/hipaa18.html/18identifiers

Hard replacement (9 categories)
Age; Bank account; License numbers; Dates; E-mail; Phone number; Personal identity number; Url; Zip code
Placeholder (3 head categories)
Geo-data: country, region, city, street, area Institution: school, work, institution, other Personal names: female, male, neutral, surname
Sensitive markup (7 categories)
Education; Profession; Family members (Ethnical info; Political views; Religious views; Sexual info)

Table 1: Categories for pseudonymization.

Evaluation 1	**A level**	**B level**	**C level**	**Total**
No. of essays	59	10	16	85
No. of tokens	7 332	7 610	8 331	23 273
Avg. tokens/essay	124	761	521	274
No. of tags (manual)	281	0	8	289
Avg. tags/essay	4.8	0	2.0	3.4

Evaluation 2	**A level**	**B level**	**C level**	**Total**
No. of essays	189	0	11	200
No. of tokens	27 510	–	4 359	31 869
Avg. tokens/essay	146	–	396	159
No. of tags (manual)	627	–	93	720
Avg. tags/essay	3.3	–	8.5	3.6

Table 2: Overview of evaluation dataset.

Table 1 shows three groups of categories: hard replacement, placeholder and sensitive markup. The difference between the groups lies partially in the degree to which these data points are structurally predictable. The first group, *hard replacement*, can be detected using regular expressions. In the case of the 9 categories in this group, we replace the original segment with a standardized representation, e.g. all *bank accounts*, despite their formats, are represented in the pseudonymized version by *0000-00 000 00*.

The second group, *placeholder*, contains three large heterogeneous categories that exhibit more varied linguistic and contextual behaviour. They are detected through matching against lists and resources and through applying various contextual triggers and rules. This group also imposes a number of constraints we need to take into account when perfoming the pseudonymization step, as we show in Section 3.5.

The third group, *sensitive markup*, contains seven categories that are potentially sensitive, but we are not at the moment sure of that. For that reason we are manually (and partly automatically) marking up these categories in the SweLL data, but do not replace them with pseudonyms at this stage. We will review all texts marked with *sensitive* tags at the final stage before the corpus release. Note here that the last four categories, all containing sensitive personal information – *ethnical and sexual information, political and religious views* — might not necessarily be revealing of the person behind the essays. However, if the person is identified on the basis of other data clues, these types of sensitive information can be used to discriminate the person, which contradicts the Ethical Review Boards' requirements. For that reason we need to find a way to conceal this information as well. The top three categories in the *sensitive* group — *education, profession and family members* — are detected and labeled with the help of our pseudonymizer service.This leaves 15 categories to be considered during detection.

3.2 SweLL learner essays

In many ways we use our common sense and linguistic competence to define rules that may capture the categories we focus on. To make sure that the rules are doing the job they are intended to do, we tested our rules on a set of manually pseudonymized essays. The evaluation was performed twice. First time, we used 85 essays availabe at that time. While improving the algorithm following the first evaluation, more essays became available, and we could use 200 essays for the second evaluation. All essays have a manually associated metadata on essay genre, topic and level of a study course where they were collected.

The 85 essays that we used for the first evaluation represent the three levels of linguistic competence that adult language learners in our corpus possess: beginner (which we call A levels), intermediate (B levels) and advanced (C levels). The 200 essays for the second evaluation come mostly from the lower level of proficiency (189 essays) with only 11 essays from the advanced C-level, see Tables 2 and 3. Statistics over the tokens and pseudonymization tags per level of linguistic development, and per genre (in the first evaluation) suggest a tendency for need of pseudonymization to a greater extent at beginner levels. It seems natural since personally oriented topics tend to dominate beginner levels (e.g. *Describe the best/worst day of you life, Present yourself*) which depends on the level of linguistic competence learners at that level master. At more advanced levels descriptive topics decrease giving way to argumentative topics where (detectable) personal details do not show up to the same extent. However, data from the second batch of C-essays contain relatively many manually assigned pseudonymization labels, especially in the essays of narrative genre. Distribution over manually assigned pseudo-tags in the data we have used (see Table 3) suggests, thus, that narrative and argumentative essays at all levels are more likely to contain personal information. To give a more detailed picture over the type of essays, we list examples of topics used for the essays in our data in Table 4.

Evaluation 1	**A level**	**B level**	**C level**	**Total**	**Avg**
Argumentative	110 (19)	–	0 (1)	110 (20)	5.5
Evaluative*	–	–	8 (15)	8 (15)	0.5
Expository	–	0 (10)	–	0 (10)	0
Instructive	17 (7)	–	–	17 (7)	2.4
Narrative	154 (33)	–	–	154 (33)	4.7
Total	281 (59)	0 (10)	8 (15)	289 (85)	3.4

Evaluation 2	**A level**	**B level**	**C level**	**Total**	**Avg**
Argumentative	276 (78)	–	8 (2)	284 (80)	3.5
Descriptive*	107 (43)	–	3 (1)	110 (44)	2.5
Explanatory*	2 (1)	–	–	2 (1)	2.0
Expository	–	–	2 (1)	2 (1)	2.0
Informal mail*	36 (13)	–	–	36 (13)	2.8
Instructive	25 (19)	–	–	25 (19)	1.3
Narrative	182 (35)	–	80 (7)	262 (42)	6.2
Total	628 (189)	–	83 (11)	721 (200)	3.4

Table 3: Manually assigned tags per genre (number of essays in brackets).

Level & genre	**Examples of topics**
A: Argumentative	About your accomodation and the quality of living Argument why you should get back money for the course you cannot attend
A: Narrative	Describe a place where you live, Describe a place you like
A: Instructive	Write a letter to your friend who has just moved to a new place and feels very uncomfortable
B: Investigative	Discuss majority vs minority languages problem (e.g. Finnish vs. Swedish in Finland)
C: Evaluative	Book review: Th. Kallifatides; Film review: Mother of mine
C: Argumentative	Discuss work moral
C: Descriptive	First day at school, Mail to a cousin about your new place of life

Table 4: Examples of topics per level and genre.

It is obvious that we have more varied data from the beginner levels (A), even though by the token count, data are relatively balanced between the levels in the first evaluation as shown in Table 2. We would need to repeat the analysis once we have more manually annotated data for better understanding whether presence of personal data points has a certain correlation with levels, genres and topics.

3.3 Detection of personal information

For effective detection, as well as for selection of appropriate pseudonyms, relevant resources are necessary. To that end, the raw resources have been collected from various openly available official statistical agencies or open services, among others using *GeoNames*[3] and *Swedish Central Statistics Agency*[4] (see more in Appendices A and B). In raw format, these lists could not be used and certain curation, restructuring, cleaning and cross-matching was necessary to adapt the sources to our needs. The adapted resources are available at an open repository for this project[5].

Only detected segments can proceed into labeling and pseudonymization steps, which makes detection the most crucial step. We split categories of personal information into three groups (as shown in Table 1):

Hard replacement group covers information that is either numerical in nature (e.g. bank account) or falls into structurally forseeable patterns (e.g. emails). Regular expressions are used as the primary technique for detecting categories in this group. For avoiding ambiguities, certain contextual clues/triggers are used, e.g. for making sure we are dealing with *Age*, we double-check the sentence for strings (triggers) like *turn, birthday, old, years*, etc.; to make sure a 4-digit combination represents *Year*, we look for indications of month, appropriate prepositions, etc. In other cases, matching against a list is necessary, for example for detection of date-strings (*4th of November*), a list of months is used. An additional source of information comes from automatically assigned parts of speech, e.g. *numeral* would have stronger association with *Year* or *Age* categories than other parts of speech.

Placeholder group represents categories that are trickier to detect and demand a more intelligent approach to pseudonymization than the group above. To start with, this group does not represent any structurally predictable patterns: a sensitive segment can consist of one word (e.g. *Adam*), a group of words (e.g. *Volvo Trucks*), linguistically inflected forms (e.g. *Stadsbiblioteket*), and may contain misspellings (e.g. *Stokhulm*). It also exhibits potential to homonymy, e.g. *Hans* (1st male name) versus *hans* (pronoun, Eng. "his"). Note, in connection, that handwritten essays are often difficult to interpret

[3] https://www.geonames.org/

[4] https://www.scb.se/hitta-statistik/statistik-efter-amne/befolkning/ amnesovergripande-statistik/namnstatistik/

[5] https://github.com/SamirYousuf/Pseudonymization with a CC BY-NC-SA license

with regard to capitalization; and in case of digitally-born essays, we have observed multiple cases of negligence to capitalization conventions (e.g. *stockholm*).

Sensitive group is separated from the others since we are at the moment not convinced whether this information is revealing enough of a person. While *family members, education*, and *profession* can be detected in an unstructured text through matching to lists, we are not certain how to formalize ways of capturing *political, religious, sexual or ethnical information* that characterizes the author of the essays. *Ethnical information* at the moment is formalized into the mentions of languages and language-based nationalities which we can match against a list of world languages. However, not all such mentions bear any ethnicity information, and it is easy to overgenerate on the task.

The other three categories — *sexual information*, *political* and *religious views* — are especially challenging when it comes to detection. For example, which part of the sentence below should be pseudonymized to render it neutral when it comes to political views (we mocked the style and errors of the original Swedish essay)? *One day we saw a big demstration there were many people wanted not Turkey minister Ardogan and we were very happy because this was the first day we see a free demstration.*

Is it the word *happy* that expresses the attitude to a political event and is revealing of the author's political views? Is it the *free demonstration* that makes it charged with political judgement? Or is it the fact itself — *demonstration in Turkey against Erdogan* — that renders this segment revealing of a person's political views? Since this type of interpretation is difficult to automatize without seeing enough examples first, we leave this category for manual markup without pseudonymization, and will return to it later to analyze examples and draw conclusions from them.

3.4 Labeling

Labeling in our experiment is conflated in one step with detection if the detected string fits into one category only. However, since a number of sensitive data points could fit into several classes (e.g. *Elena* being both a female first name and a town in Bulgaria), there is a need to disambiguate polysemous strings. To that end, we exploit several approaches (see also Figure 1):

Contextual triggers, which are category-dependent. For example, to distinguish between a telephone number and a personal identity number, we check the nearby context of a current sentence for related words. Among others, *call, phone* would associate strongly with the "telephone"-hypothesis.

Frequency approach. For example, in case of *Elena* we check which of the two potential categories — *first name* or *city* — is more likely according to frequency. In this particular case, *Elena* is more frequent as a *female first name*, and the category is assigned on that ground.

Disambiguation. In certain cases we rely on parts of speech to disambiguate dubious cases. In other cases we leave disambugation to an assistant, who has to pick one of the suggested categories, or rewrite adding another one. In the future, we plan to experiment with crowdsourcing correction/disambiguation of automatic pseudonymization by learners/authors who write essays.

3.5 Pseudonymization

Out of the three groups of personal data points (see Table 1) only the first two groups are pseudonymized: the *hard replacement* group and the *placeholder* group.

The *hard replacement* categories are replaced once without any further possibilities to fiddle with the formats. Each category in this group maps to a single format of replacement, for example, *urls* are always pseudonymized with *url.com*. Two categories in this group differ a bit, namely, *age* and *year*. In both cases we replace the original numbers with the one that fall within a span of +/-2 from the original value. For example, if the age in the original is *21*, the range for randomly instantiating a pseudonym will be *19–23*. Age spelled as a string is first converted to a number, and then replaced/pseudonymized with a digit. Subcases of digital representation of dates, months, weekdays and years — spelled out as strings — are especially provocative since they entail various misspellings and we may fail on the detection step.

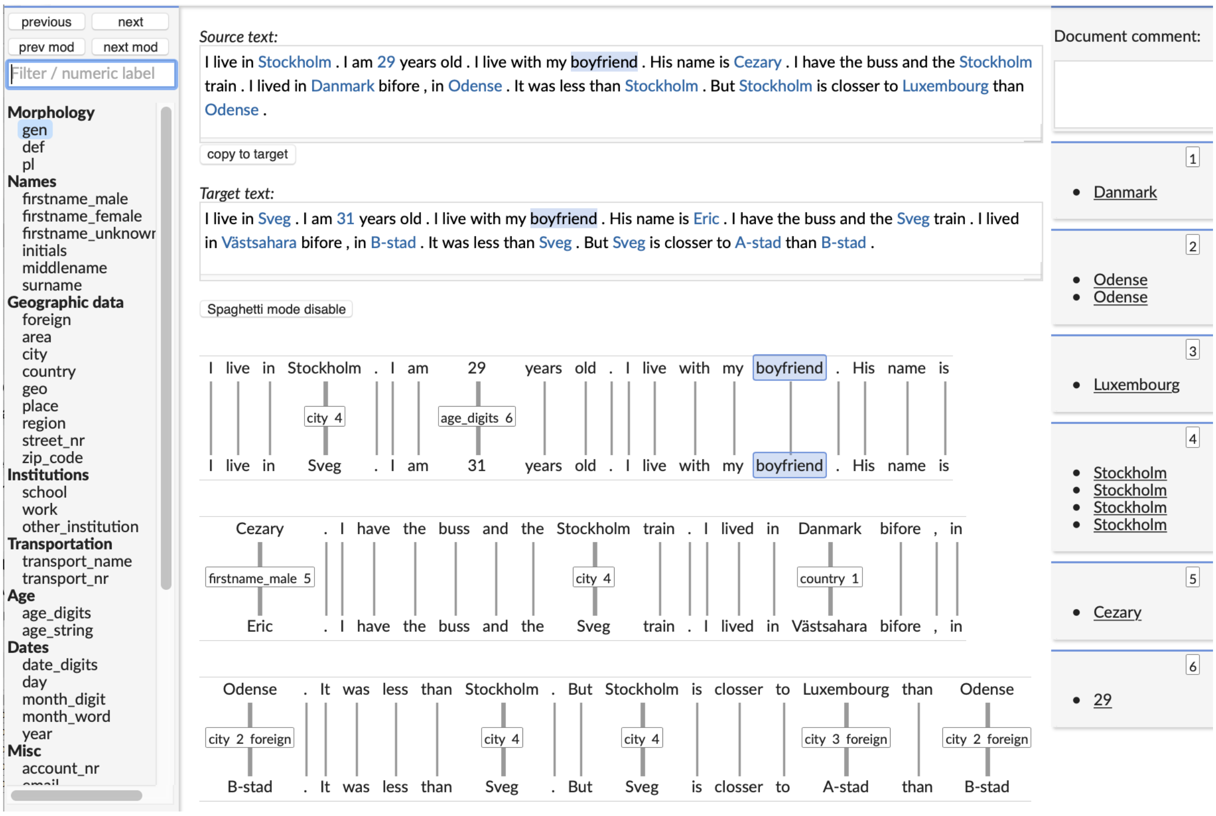
Categories in the *placeholder* group can be recurrently instantiated to new pseudonyms after the initial pseudonymization. There are some rules we need to follow to ensure consistency, namely:

1. Names are pseudonymized by gender correspondence. If gender is uncertain, we use gender-neutral names, e.g. *Kim*, to select from. To make sure we use relatively common names, we randomly select

pseudonyms from the top 50 most frequent names per list (female, male, neutral, surnames).
2. City mentions are checked with country mentions to keep the context to the same geographical area. Thus, if the original says *I lived in Danmark in Odense*, the pseudonymizer will check whether the country-mention *Danmark* and the country of the mentioned city *Odense* are the same, and the selected pseudonyms will also be selected from the same country. To avoid unusual cities, we randomly pick only the 5 most frequent cities per country. Another variant we are experimenting with is to use fake cities for Swedish cities, and a list of *A-city*, *B-city*, etc. placeholders for all other cities. The second (fake) alternative is used to avoid grammatical and semantic infelicities that could be observed when using real geographic names in inappropriate contexts, e.g. *I live in Barcelona where I can ski all year round.*
3. Street mentions are coordinated with the currently pseudonymized city, although this information is not all-covering in our resources. In that case, a fake street name from a special list is picked.
4. If the detected segment has a morphological form that we keep track of, e.g. genitive case, plural or definite form, we assign a morphological label (*gen, pl, def*) to render the form in the pseudonym.

3.6 Visualization

The code – written in python – has been integrated into the SweLL annotation tool *SVALA* (Wirén et al., 2019), for initial tests and user-friendly inspection. The pseudonymizer service can be called from the tool on any uploaded Swedish text (some categories can work for other languages as well, for example if writing conventions for personal data points coincide with the Swedish ones, e.g. many personal and geo names are spelled similarly in Swedish and English). The pseudonymizer gets a raw text, tokenizes it, segments into sentences and applies all the rules described above. Additionally, an automatic annotation Sparv pipeline for Swedish is run (Borin et al., 2016).



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 .

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 .

Figure 2: Pseudonymization mode in the annotation tool SVALA.

The output is delivered in two formats – xml and the so-called SVALA format. In the case of the SVALA format, three objects in json format are generated: *original* text, *target* (in this case pseudonymized) text, and *edges* that describe links between the original and the target versions with labels assigned to links.

Figure 2 shows a pseudonymization mode in *SVALA*. In the middle zone (from top down) you can see *Source text zone* – with certain tokens in a different colour (marked in blue). These are the ones that have been detected by the pseudonymizer service.

Target text zone – where the marked tokens (in blue) are the pseudonyms that have been automatically selected to match the detected category.

A graph (informally called *spaghetti*) – representing parallel versions of the same essay, original and pseudonymized, with links connecting them token by token. Certain links contain labels that consist of a pseudonymization tag and a numerical reference-id to keep track of the mappings, so that the same name in the original version will be replaced with the same pseudonym in the target version if used more than once (see data points listed on the right, with their reference IDs).

On the left, there is a list of all pseudo-tags and some other clickable options for manual correction/pseudonymization. On the right, clickable reference IDs together with all detected sensitive data points (strings) are listed in groups.

Each time a pseudonym tag is inserted, the change is pushed into the *edges* object, and is represented graphically in the user interface (i.e. *spaghetti* area) for better visualization. The *target text* area is also updated with an automatically assigned pseudonym. Browsing through the text will highlight the currently hovered token in all the three fields – *Source text, Target text, Graph (spaghetti)*, see, for example, the word *boyfriend* in Figure 2.

4 Results

Table 5 shows the number of pseudonymization labels per genre as assigned (1) by human annotators, (2) by the rule-based approach and (3) by a combined rule-based approach using automatic part-of-speech tagging (POS) for disambiguation, referred to as *POS+rules* approach. All numbers except for the asterisk-marked(*) genres come from the second evaluation, whereas Evaluative* and Investigative* genres were not represented in the evaluation dataset for the second round. For the sake of discussion we chose to report the numbers from the first round of evaluation for these two genres.

During the first evaluation, we observed that the automatic pseudonymization performed more reliably in narratives, argumentative and instructional texts, but failed in investigative and evaluative genres. On manual inspection and comparison of the two modes of pseudonymization, we noted that in investigative and evaluative genres personal information is practically never used. We observed the following:

Evaluative texts are based on book and film reviews with plenty of names (person, places) without any reference to the author of the essay. Thus, the human annotator did not consider labeling the passages, whereas the automatic service followed the same principles despite the topic change.

Investigative texts are usually based on a newspaper article and the writer discusses different aspects of the topic touched in the article. This entails references to names, places and various facts that do not reveal the writer's identity, but most often point to some famous political or cultural figures. Like with evaluative texts, texts of this genre could potentially be exempt from pseudonymization. Care should be taken, though, when mentions of first person personal pronouns take place (e.g. *In Vietnam, we...*).

For the second evaluation, we have updated the pseudonymizer service and added a version which consults automatically assigned parts of speech for improved disambiguation. We can see that the pseudonymizer service in both its variants is quite close to the human annotation numbers for all genres (Table 5). Table 6 shows the precision, recall, F1 and F2 scores per label for the POS+rules algorithm. We can see that the accuracy of automatic annotation reaches the value of 0.89 (F2 score) which can be seen as a good result. However, we can also see that both F scores, i.e. a combined score of catching true positives and excluding true negatives, is low for certain categories, i.e. *place, surname, date_digits*. Manual inspection of these categories for *false positives*, i.e. segments that are not personal in nature, but have been erroneously assigned a personal label, and *false negatives*, i.e. segments that are personal in nature, but have been missed by the automatic annotation, has revealed consistent problems with:

(1) lack of capitalization of names, surnames, cities and countries, which leads to false negatives during the detection step (almost 50% of false negatives).

(2) misspellings that are difficult to compensate for automatically, also leading to false negatives. A

	Manual	Rules	POS+Rules
Argumentative	283	280	249
Descriptive	110	148	140
Evaluative*	8	276	–
Explanatory	2	4	4
Expository	2	4	4
Informal mail	36	32	22
Instructive	25	26	21
Investigative*	0	476	–
Narrative	262	267	249
Total	720	761	689

Table 5: No. of pseudo tags per genre and annotation mode.

	Precision	Recall	F1 score	F2 score
age_digits	1.00	1.00	1.00	1.00
city	0.91	0.91	0.91	0.91
country	1.00	0.75	0.85	0.78
date_digits	0.40	1.00	0.57	0.77
day	0.00	0.00	0.00	0.00
edu	0.00	0.00	0.00	0.00
email	1.00	1.00	1.00	1.00
firstname	0.94	0.98	0.96	0.97
island	0.00	0.00	0.00	0.00
month_word	1.00	1.00	1.00	1.00
other_nr_seq	0.00	0.00	0.00	0.00
phone_nr	1.00	1.00	1.00	1.00
place	0.33	1.00	0.50	0.71
prof	0.00	0.00	0.00	0.00
school	1.00	1.00	1.00	1.00
surname	0.67	0.32	0.44	0.37
year	0.96	1.00	0.98	0.99
zip_code	1.00	1.00	1.00	1.00
Accuracy			0.90	0.89

Table 6: Performance scores for automatic pseudonymization service incl. some sub-categories.

possibility could be to continuously collect lists of misspellings to add to our resources that we use for matching.

(3) overgeneration of *(sur)name* and *city* tags for capitalized words in the text. Since the name database contains a lot of international names that often coincide with Swedish pronouns, verbs or other common vocabulary, they are labeled for a category on the basis of matching. Here, automatic POS-tagging reduces the majority of ambiguous cases, even though it doesn't solve all issues.

The analysis has also shown that we lack resources for the *place* category, and hence systematically fail to detect *place* mentions, e.g. *I live in Stockhom in Bromma*, where *Bromma* is a part of the city. This information is not available for downloading or scraping. However, since it might be revealing of a person, we need to find a workaround to collect this resource. An option could be to enrich the list every time this label is used by the human annotators, and start from there.

From the point of view of protection of personal integrity, *recall* is a more important measure (i.e. the pseudonymizer does not miss personal information) than *precision* (i.e. the pseudonymizer does not assign personal labels to non-personal information), which is captured by the **F2 score** by giving extra weight to recall. However, from the point of view of readability and research value of the data, overgeneration is a serious drawback, and hence precision is nonetheless important, which is reflected rather by the **F1 score**. There is therefore important to note that *surname* is missed more often (F2 score < F1 score), whereas *date_digits* and *place* are overgenerated (F2 score > F1 score), see Table 6.

Further, we computed Inter-Annotator Agreement (IAA) for the two modes of annotation — manual and POS+rules — using NLTK implementation (Bird and Loper, 2004). The agreement is reported using Fleiss kappa (Davies and Fleiss, 1982) and Krippendorff's alpha (Krippendorff, 2018). In both cases, the value of 0.86 is reached. Fleiss' kappa within the range 0.81-1.00 means almost perfect agreement, which is a very encouraging result. However, given that pseudonymization should protect people from accidental privacy breaches leaving no room for chances, we need to improve the performance further.

To gain more insights into the use of various pseudo-categories and to identify the ones causing most disagreement, we calculated statistics in the form of a confusion matrix. The result has shown that the most used tags are *city, country, firstname* and *year*; most frequently confused ones are *city–country* and *place–city/country*. Most other tags are used relatively rarely.

5 Discussion and concluding remarks

There are two main risks with the detection of entities that contain personal information — false positives, i.e. flagging for presense of a risk that is not really there, overgeneration, and false negatives, i.e. failure to detect a real risk. *False positives*, or *overgeneration* could be a real problem when it comes to learner essays for certain topics, e.g. book reviews or argumentative essays based on political articles. By the nature of those topics, there will be names, places and dates used in the text which are by no means revealing of the author of the essay. Obtrusively pseudonymizing names of political figures or book characters may introduce (unnecessary) readability, or simply introduce errors if a pseudonym should e.g. agree with the rest of the context semantically, syntactically or morphologically. Potentially, there is a possibility to "turn off" pseudonymization of essays that have any of such topics. Provided we have a list of topics which we are safe to apply, it would be possible.

False negatives, on the other hand, can reveal weaknesses in the rules, e.g. lack of coverage of the underlying resources, failure of the rules to capture some specific cases, or inability of the algorithm to identify sensitive personal data due to misspellings. If no manual control is applied, false negatives can set the author of the essay to serious risk of being identified. In the future, challenges with spelling (and other) errors need to be addressed alongside with questions around whether errors and grammatical forms used by the learners need to be projected to pseudonyms.

Pseudonymized version of a text represents a manipulated — and presumably more protective — version of an original text. An important question in connection to this is what is more important: that a pseudonymized text reads like an original, or that it is obvious that the text is pseudonymized? Language teachers and assessors within the SweLL project have been unanimous about keeping clear identifiers of all text segments that have been manipulated in order not to mistake pseudonyms for learner's production. For this reason, we have not considered the need to conceal pseudonyms. On the contrary, we keep all pseudonymized segments clearly highlighted.

This attitude seems, however, to be domain dependent. In medical domain, for example, publications reveal an intention to pseudonymize medical records in such a way so that no one using the data would suspect which sections are pseudonymized. For example, Dalianis (2019) describes an evaluation procedure of pseudonymization where human evaluators are given the task to identify which of the medical records are pseudonymized, and which are not. If pseudonymized texts are taken for being orignal ones, this is taken as a sign of a high quality of pseudonymization.

However, the current paradigm has been shifting from hiding the fact of pseudonymization towards actually pointing out which text segments have undergone careful examination to hide revealing personal information. In light of this, possible consequences of linguistic infelicities that pseudonymization might introduce into the texts seem to be minor compared to the risk of getting access to an original. For the sake of readability, however, pseudonymized texts need to retain the same level of (natural) flow, as well as grammatical and semantic coherence. The latter aspects present non-inconsiderable challenges. A number of unstructured data points that might be attributed to a person through context are still left outside the scope of this work (e.g. *When I was four I climbed up a mango tree in my uncle's garden and fell off. I have scars and limp since then.*), as are misspelled names that cannot be matched to the resources we are using. Even though the GDPR (EU Commission, 2016, Recital 15) admits that where personal data is stored in an unstructured way it *might* not be covered by the GDPR, the ethical restrictions still apply. In connection to which we double-check all essays manually to identify such cases. In the future, we will test automatization of that part of the work as well.

To summarize, use of *pseudonymization* holds two strong benefits in the research context: compliance with GDPR and permission to use data beyond the original purposes of collection. However, pseudonymized data is still personal data, and needs to be protected in further ways, such as encryptions, authorizations, etc. In the future, we would like to test machine learning for this problem, and to test crowdsourcing for correction of automatic pseudonymization with the ultimate goal to start collecting learner essays online with a secure on-the-fly pseudonymization.

Acknowledgements

This work has been funded by a grant from the Swedish Riksbankens Jubileumsfond (SweLL - research infrastructure for Swedish as a second language, dnr IN16-0464:1) and by *Nationella Språkbanken* – jointly funded by its 10 partner institutions and the Swedish Research Council (dnr 2017-00626).

References

Lars Ahrenberg and Beata Megyesi. 2019. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–341.

Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, pages 69–72. Association for Computational Linguistics.

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, pages 17–18.

Hercules Dalianis. 2019. Pseudonymisation of swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation, September 30, 2019, Turku, Finland*, number 166, pages 16–23. Linköping University Electronic Press.

George Danezis, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Métayer, Rodica Tirtea, and Stefan Schiffner. 2014. Privacy and Data Protection by Design – from policy to engineering. https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design (Accessed 2019-11-17).

Mark Davies and Joseph L Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051.

Luis Hernández Encinas, Agustín Martín Muñoz, Víctor Gayoso Martínez, Jesús Negrillo Espigares, José Ignacio Sánchez García, Claude Castelluccia, and Athena Bourka. 2015. Online privacy tools for the general public. Towards a methodology for the evaluation of PETs for internet mobile users. https://www.enisa.europa.eu/publications/privacy-tools-for-the-general-public (Accessed 2019-11-17).

ENISA. 2017. Privacy Enhancing Technologies: Evolution and State of the Art. A Community Approach to PETs Maturity Assessment. https://www.enisa.europa.eu/publications/pets-evolution-and-state-of-the-art (Accessed 2019-11-17).

ENISA. 2018. A tool on Privacy Enhancing Technologies (PETs) knowledge management and maturity assessment. https://www.enisa.europa.eu/publications/pets-maturity-tool (Accessed 2019-11-17).

EU EU Commission. 2016. *General data protection regulation.* Official Journal of the European Union, 59, 1-88. https://gdpr-info.eu/ (Accessed 2019-11-19).

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology.* Sage publications.

M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, M. Rodríguez, Antonio Lo Martin, M. Villegas, and M Krallinger. 2019. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019).*

Ben Medlock. 2006. An introduction to nlp-based textual anonymisation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC).*

Beáta Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén, and Elena Volodina. 2018. Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish. In *Proceedings of the 7th NLP4CALL, Swedish Language Technology Conference, SLTC 2018*, pages 47–56.

Medical Research Council MRC. 2019. GDPR Guidance note 5: Identifiability, anonymisation and pseudonymisation. (Accessed 2019-11-22).

Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):1–9.

Frances Eileen Rock. 2001. Policy and practice in the anonymisation of linguistic data. *International Journal of Corpus Linguistics*, 6(1):1–26.

Dan Rosén, Mats Wirén, and Elena Volodina. 2018. Error Coding of Second-Language Learner Texts Based on Mostly Automatic Alignment of Parallel Corpora. In *CLARIN Annual conference 2018.*

Latanya Sweeney. 2000. Simple Demographics Often Identify People Uniquely. *Health (San Francisco)*, 671:1–34.

Sumithra Velupillai. 2014. Temporal expressions in swedish medical text–a pilot study. In *Proceedings of BioNLP 2014*, pages 88–92.

Elena Volodina, Lena Granstedt, Sofia Johansson, Beáta Megyesi, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2018. Annotation of learner corpora: first SweLL insights. *Proceedings of Swedish Language Technology Cconference (SLTC) 2018.*

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2020. The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology, Special Issue.*

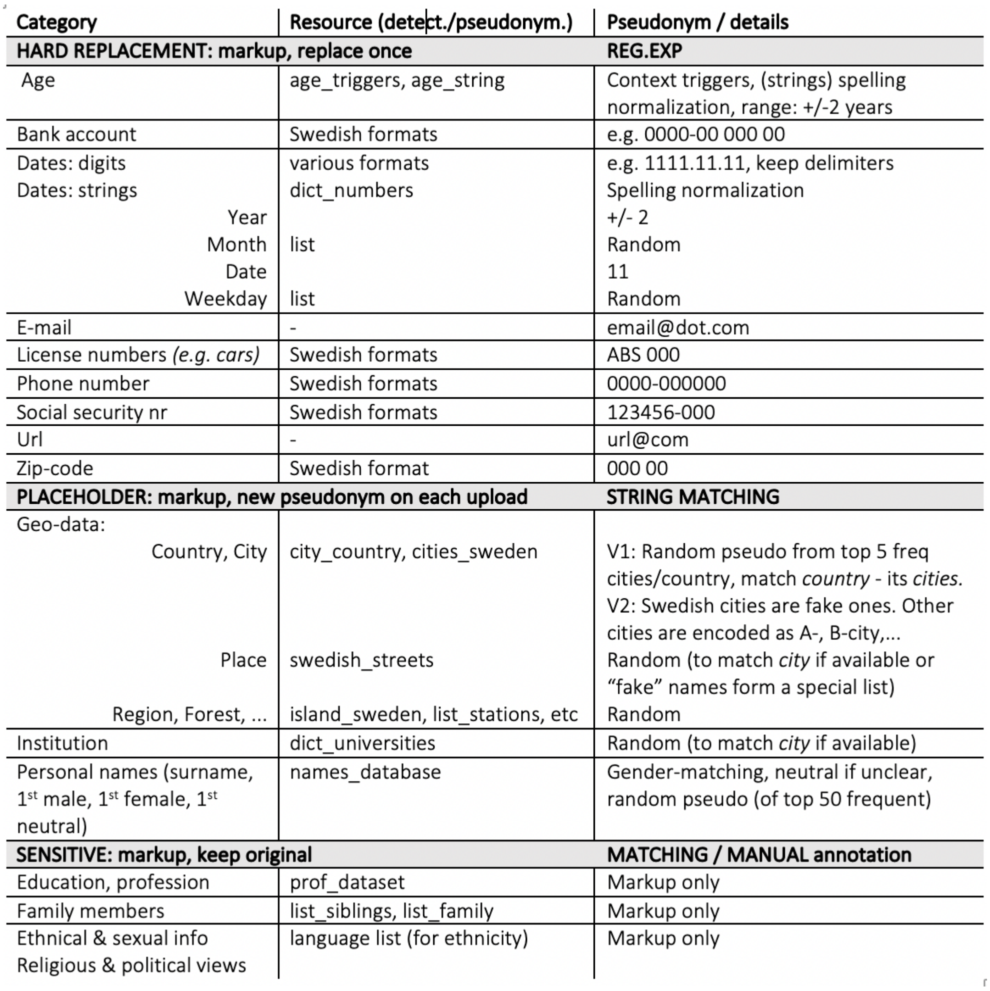
Mats Wirén, Arild Matsson, Dan Rosén, and Elena Volodina. 2019. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. *Post-conference proceedings of CLARIN 2018.*

APPENDICES

A Resources and gazeteers[6]

1. various rich geographical information was obtained from *geonames*[7]
2. lists over personal names in Sweden with their statistics of use have been provided by the *Swedish Central Statistics Agency* [8]
3. a list over registered companies in Sweden have been provided by *Bolagsverket* [9]
4. names of streets, places and islands in Sweden come from the online service *Svenskaplatser*[10]
5. list of languages and their ISO-639-3 codes were downloaded from *geonames*[11]
6. lists covering professions and education come from *Yrkesguiden*[12]
7. lists with Swedish universities come from *Wikipedia*[13]

B Applied techniques per category



Category	Resource (detect./pseudonym.)	Pseudonym / details
HARD REPLACEMENT: markup, replace once		**REG.EXP**
Age	age_triggers, age_string	Context triggers, (strings) spelling normalization, range: +/-2 years
Bank account	Swedish formats	e.g. 0000-00 000 00
Dates: digits	various formats	e.g. 1111.11.11, keep delimiters
Dates: strings	dict_numbers	Spelling normalization
Year		+/- 2
Month	list	Random
Date		11
Weekday	list	Random
E-mail	-	email@dot.com
License numbers *(e.g. cars)*	Swedish formats	ABS 000
Phone number	Swedish formats	0000-000000
Social security nr	Swedish formats	123456-000
Url	-	url@com
Zip-code	Swedish format	000 00
PLACEHOLDER: markup, new pseudonym on each upload		**STRING MATCHING**
Geo-data:		
Country, City	city_country, cities_sweden	V1: Random pseudo from top 5 freq cities/country, match *country* - its *cities.* V2: Swedish cities are fake ones. Other cities are encoded as A-, B-city,...
Place	swedish_streets	Random (to match *city* if available or "fake" names form a special list)
Region, Forest, ...	island_sweden, list_stations, etc	Random
Institution	dict_universities	Random (to match *city* if available)
Personal names (surname, 1st male, 1st female, 1st neutral)	names_database	Gender-matching, neutral if unclear, random pseudo (of top 50 frequent)
SENSITIVE: markup, keep original		**MATCHING / MANUAL annotation**
Education, profession	prof_dataset	Markup only
Family members	list_siblings, list_family	Markup only
Ethnical & sexual info Religious & political views	language list (for ethnicity)	Markup only

[6] We are grateful to anonymous reviewers suggesting to have a look at the *European Open Data* for the Swedish language: https://www.europeandataportal.eu/sv and at *Heideltime*, also available for the Swedish language (Velupillai, 2014)

[7] http://www.geonames.org

[8] https://www.scb.se/hitta-statistik/statistik-efter-amne/befolkning/ amnesovergripande-statistik/namnstatistik/

[9] https://bolagsverket.se/ff/foretagsformer/namn

[10] https://www.svenskaplatser.se/

[11] http://www.geonames.org

[12] https://www.gymnasium.se/yrkesguiden/alla-yrken-10957

[13] https://sv.wikipedia.org/wiki/Lista_%C3%B6ver_universitet_och_h%C3%B6gskolor_i_Sverige