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

swell@svenska.gu.se

Abstract

This paper reports on the status of learner corpus anonymization for the ongoing research infrastructure project SweLL. The main project aim is to deliver and make available for research a well-annotated corpus of essays written by second language (L2) learners of Swedish. As the practice shows, annotation of learner texts is a sensitive process demanding a lot of compromises between ethical and legal demands on the one hand, and research and technical demands, on the other. Below, is a concise description of the current status of pseudonymization of language learner data to ensure anonymity of the learners, with numerous examples of the above-mentioned compromises.

1 Introduction

SweLL—Swedish Learner Language—is a project aimed at setting up an electronic infrastructure for collecting, annotating, searching and analyzing Swedish learner language (Volodina et al., 2016a). During the first year of the project, a number of the project aims related to the questions of data accessibility for the research community have been addressed, such as

1. legal and ethical aspects of essay collection,
2. principles of learner language anonymization and pseudonymization, and
3. tools and platforms for ensuring the previous steps.

Annotation in general is where linguistics – as well as pedagogy and other disciplines – nowadays hide in Natural Language Processing (NLP) (Fort, 2016). (Annotated) L2 data is extensively used for research, for instance within NLP, Second Language Acquisition (SLA) and Learner Corpus Research (LCR), and thus the annotation should be reliable, reproducible, and comparable between different corpora, so that conclusions drawn from the data are also reliable. But above all the data needs to be open outside the original project where it has been collected, a challenge that is not so easy to address with the new European Union (EU) General Data Protection Regulation (GDPR)[2]. The demands that we face require careful analysis of what makes the data sensitive and we need to take all possible precautions to reduce the risks of illegal or unethical use of the data before it can be made accessible.

To ensure that the data collected in the project can be used openly in research, we have worked extensively on legal issues, data handling flow, anonymization principles and tools in support of anonymization. Below, we describe the first steps and insights taken in SweLL.

1.1 SweLL infrastructure

The purpose of the SweLL project is to set up an infrastructure for continuous collection, digitization, normalization, and annotation of texts written by learners of Swedish as a second language. The aim is to make available (as open access) a linguistically annotated corpus consisting of a collection of approx. 600 learner texts and tools for automatic processing of these texts by allowing search and download for registered users (Volodina et al., 2016a).

The texts in the collection are produced by learners of Swedish as a second language from the age of 16 on voluntary basis given their consent. The texts are collected in schools where education is given in Swedish as a Second Language such as Swedish for Immigrants (SFI) or Swedish as

https://spraakbanken.gu.se/eng/swell_infra

[2] https://gdpr-info.eu

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. *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018)*. Linköping Electronic Conference Proceedings 152: 47–56.

second language, or where learners are tested for their proficiency in Swedish, such as CEFR (Common European Framework of Reference (Council of Europe, 2001)) or TISUS (Test In Swedish for University Studies (Volodina et al., 2016b)). Our aim is, by the end of the project, to have collected and annotated at least 600 texts and exercise answers written in response to tasks given by the teachers to students in schools, along with additional metadata information about the learners and the writing task.

We envisage a multi-purpose environment that combines data collection, algorithms for automatic processing of data, visualization analytic tools and L2 task generation. SweLL creates an infrastructure consisting of:
1. a data collection portal, through file import and via online exercises,
2. an annotated corpus of written L2 production,
3. methods and tools for L2 analysis, and
4. specific search tools for L2-material facilitating filtering for e.g. writers of a certain mother tongue, or writers at a certain proficiency level.

The material and tools will be made accessible through the learning platform Lärka (Volodina et al., 2014) created and maintained by Språkbanken at Gothenburg University. Lärka has up to now been a login-free online tool used for teaching Swedish grammar to university students and for deploying prototype exercises for learners on Swedish vocabulary. Lärka is extended to include a portal for collecting and processing L2 corpora, and linked to Korp (Ahlberg et al., 2013) and Strix - two tools under development at Språkbanken - for browsing texts and visualization of statistics and analytics.

In the long term, the data in terms of the collected essays and information about the learner, along with its reliability—and above all its accessibility—are the most important issues in the SweLL electronic infrastructure. To assure long-term usage and open access to the SweLL data collection, we were keen to adhere to current law and regulations in the SweLL data management flow.

2 Legal issues and learner corpora

2.1 Data protection and free access

The European Union's new General Data Protection Regulation (Regulation EU 2016/6791), enforced on May 25 2018, regulates the processing of personal data related to individuals by an individual, a company or an organization in the EU. Personal data ”means 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).

GDPR demands that stored data containing personal information undergo either an anonymization or a pseudonymization process. *Anonymization* is the removal of all personal identification so that the person is not or no longer identifiable. Thus, the data must be stripped of any identifiable information, making it impossible to derive insights on a certain individual, even by the party that is responsible for the anonymization. Anonymous data cannot be re-identified. *Pseudonymization* according EU GDPR (Article 4) is ”the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately”. “Additional information” is typically a translation table by which pseudonymized personal data can be mapped back to the original data. But since this ”additional information” should be the *only* means to re-identify a person, a consequence is that it must not be possible to do re-identification with the help of *other* information openly available, for example, on the Internet or in public registers, or by coordinated processing of such information[3]. This is what put such high demands on pseudonymization.

Contemporary trends in modern research has caused an increase in building large infrastructures in support of research. With respect to data collection, an electronic research infrastructure ideally consists of: (Volodina et al., 2016a):

1. freely accessible data in electronic format,
2. a technical platform for exploring the data, including tools and algorithms for data analysis, and visualization,
3. a set of tools and technical solutions for new data collection and preparation, including data processing and annotation, and

[3] in Swedish: ”samkörning”

4. relevant expertise within the area.

On the one hand, the most important aspect for research as promoted by the major granting offices is *freely accessible data* in electronic format. On the other hand, modern legislation makes it more and more difficult to collect data for open use in research, especially in connection to the recently adopted GDPR, which sets to protect *data subjects' integrity*, and which—in combination with Swedish legislation on open access to public data (Riksdagen, 1949, ch.2)—sets certain limitations on the metadata types we are able to collect and types of information that we are able to keep in the original texts, see discussion of that in Volodina et al. (2018).

2.2 Pseudonymization in learner corpora

Out of the above follows the need to take precautions not only when it concerns the metadata, but also when it comes to the contents in the learner-written texts. This step usually takes form of pseudonymization—a general term which covers all possible ways of manipulating such information in the texts that can reveal an author behind them. This information might include, for example, person name, age, locations like home town, address, work place, family related issues, or text items revealing information that can be used for any kind of discrimination, being it political views, religious convictions, or sexual orientation.

To minimize the chance that personal data records and identifiers lead to the identification of subjects, all identifiers in the essays need to be overseen, masked and eventually replaced to ensure anonymity. Thus, pseudonymization includes the identification of personal information that can relate to the subject (e.g. My name is *Ali*), and the classification of that information, masked into certain predefined types (e.g. My name is *first_name*). Each information type can then be replaced in a systematic way to reproduce a "natural" text to increase reading flow (e.g. My name is *Robert* where the original first name is replaced randomly by another first name).

There are several ways to mask the sensitive information in the pseudonymization process, among others through substitution (e.g. Poland →Greece); by making text noisy (e.g. Poland →Europe); or by completely removing a text segment.

Different approaches to pseudonymization (which is also often called anonymization in the NLP literature, see e.g. Medlock (2016)) are used across learner corpus projects[4]. For instance, in CzeSL (Rosen, 2017) all names are substituted with *Adam*, *Eva* or *Sin*, in corresponding morphologically inflected forms, preserving possible spelling errors in suffixes or endings. In many other cases the notation uses codes, e.g. village<priv>. In ASK (Tenfjord et al., 2006), codes in the format *@name, @place, @something*, etc. replace the original tokens (Tenfjord et al., 2006). In CroLTec (Preradović et al., 2015), replacement of names was hard-coded during the error annotation without any special guidelines. Essays containing political views and other sensitive information were discarded from the corpus. Next, we will describe the SweLL approach to protect the anonymity of the learners.

3 Data management and pseudonymization in SweLL

In order to assure that the collection and access of the texts written by the learners (i.e. the subjects) comply with applicable laws and regulations, especially GDPR, the data needs to be handled in a secure way during collection and storage, and the subjects in the corpus must be de-identified. *De-identification* occurs when data has been stripped of common identifiers such as 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. These identifiers might occur in metadata about the learner, and in the learner's text(s).

The SweLL project adopted a rather restrictive approach to metadata describing important aspects about each produced text and learner in a way that learners are de-identified while still providing important information for research purposes about the learner's gender, age, total time in Sweden, education level and languages spoken in various communicative situations. The full set of metadata will be described in Section 3.2.

De-identification through metadata might not be solely satisfactory, since the texts written by a learner may also contain personal information

[4] Information about anonymization approaches in other projects comes from personal communication with involved researchers

connected to the learner, see for example Figure 1 where metadata in combination with the text may give away the physical person behind them. This means that we need to manipulate the text written by the learner with the purpose of hindering the possibility of going back to the original text, e.g. mention of profession *web developer* in Figure 1 to guarantee that the learner is de-identified.

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.”

Figure 1. Example of (selected) metadata and an essay text for a fake learner.

Since we need to keep the information about the learner throughout the project in order to be able to delete his/her record in the database if the learner so requests, we pseudonymize (rather than anonymize) both the text and the information about the learner. How we handle the identification of personal information and pseudonymization in learners' texts is described in detail in Section 3.3.

3.1 Data management in SweLL

The processing of SweLL data—from collection through storing to search and retrieval—is based on the ethical frontier *Building digital trust: The role of data ethics in the digital age*, developed by Accenture labs (Accenture, 2016), which describes best practices for data sharing. The model for the SweLL project data handling process is based upon this seven-step model (as described by *Data ethics and digital trust*). The model includes i) acquisition, ii) storing, iii) aggregation iv) analysis v) usage, vi) sharing and vii) disposal. Here, we give a brief outline to this process.

During *data collection* the teachers inform learners of the project and its aims. To ensure that the learners understand what they agree to we provide information not only in Swedish but also in several other languages common as mother tongues (L1) among learners of Swedish, including Arabic, Bosnian-Croatian-Serbian, Dari, English, Farsi, Greek, Kurmanji, Sorani, Somali, Spanish and Tigrinya[5]. In the consent, we inform the learners about the project, and describe the management of personal information throughout the project, including the statement that participation is entirely voluntary and the subject can opt out of continued involvement whenever he/she wants without the need to provide any explanation. Further, we state that we will not disclose the person's name and we will remove personal information from the texts to guarantee anonymity. Since the agreement covers a period of a learner's involvement in the project (e.g. a year) which is stated in the agreement, we do not need to ask for a new agreement every time we collect a text from a learner.

Once the learner agreed to donate his or her text(s) to the project, the teachers are responsible for the collection of the essays and additional personal- and task-specific metadata about the learner, the assignment, and the learner's grade. For each learner, we collect 1) the agreement form signed by the learner and 2) personal information about the learner. From each teacher, we collect 1) information about the assignment, and 2) the learner's grade of a particular essay when applicable. We collect agreements and metadata forms from the learners under teachers' guidance (in some cases in the presence of researchers or project assistants).

Data and data-related documents are handled and stored, making them both secure and easily accessible within the project for further processing. Teachers keep agreements, metadata sheets, task sheets and hand-written essays in safes at their schools until the documents are collected by researchers/project assistants. In the case of electronic essays, they are copied to a USB-memory stick and kept in a safe. Once the data and all related documents are transported by the project assistants/researchers from schools, all agreements

[5] However, as a word of warning—to ensure that project assistants can interpret the filled forms correctly, subjects usually fill in the Swedish form, and use translations only as support.

are collected and stored (on paper) in a safe.

To hide the identity of the learner for each essay, we assign a SweLL-ID to each learner. The SweLL-ID is inserted into the personal metadata sheet's special field. Project assistants register the personal metadata on the SweLL portal creating a "learner"-record. The list with the mappings between the learner's name and SweLL-ID is defined as *the key*. The key is kept in a safe, together with agreements, metadata sheets and the hand-written essays. The key is necessary to be kept making it possible to delete learner specific data if a participating subject (individual) so requests.

Information about the assignment provided by the teachers is uploaded to a portal by creating a task-ID, which is then linked to relevant essays. Where there are handouts, they are scanned and saved to the "task" profile. The forms containing information about the assignment are delivered either on a USB-memory stick or on paper.

The essays written by the learners are processed by researchers and research assistants. The essays originally written on computer as non-anonymized are saved on USB-memory and kept in a safe. On upload of an essay, the essay is linked to the specific SweLL-ID (with the learner's personal metadata). The handwritten essays are transcribed by project assistants using encrypted portal functionalities (SweLL-kiosk mode). All the essays written by the same person are connected systematically through the SweLL-ID and metadata information without revealing the identity of the learner.

Within the project, we operate under GDPR for the essay collection and pseudonymization. Neither the participating learners', nor the teachers' identity are to be revealed to the public. However, the list of participating teachers and schools, and the list of participating subjects with their SweLL-IDs are kept throughout the project in a safe to secure contact information in the long-term during the entire project period. Once the data is de-identified and the texts are pseudonymized, the data is made available to the public with a restricted license, which requires login and password for access to the portal.

3.2 Pseudonymization in metadata in SweLL

When designing the set of metadata, we tried to strive for necessary and detailed information for research purposes without jeopardizing the identification of the learners. Metadata concerning personal information about the learners is required for the project purpose to develop methods and exercises to particular groups of learners with various first and second languages, language skills and grades.

Personal metadata includes information about the learner's gender (<female>, <male>, <decline to respond/other>); instead of exact year of birth or age, the date of birth is given in 5-year interval spans (e.g. 1950–1954); instead of arrival date to Sweden, we ask for total time in Sweden in years and months; no information is provided on the educational establishment where the essays have been collected, but we ask for education level outside and in Sweden in years (<elementary school>, <introductory programme>, <gymnasium/upper secondary school>, <technical/vocational school>with degree, <university/other inst. of higher education>with degree and <other>. To further complicate possible identification of a learner through aggregated personal information, the metadata does not provide a country of origin or nationality of the learner but we restrict to information about the mother tongue (L1) only. Lastly, we ask information about how the learner learned Swedish (self-taught or took Swedish courses given as number of years and months).

In order to ensure high quality and usefulness of the corpus in research and development, information about the writing task is also essential, represented as additional task-oriented metadata without any personal information about the learner. We also ask teachers to provide the grade or result of the exercise for each particular essay written by each learner; For identification the learner's SweLL-ID is assigned instead of the name of the learner.

3.3 Pseudonymization of texts in SweLL

For the SweLL data set of the texts written by the learners, we manually identified text segments that reveal personal information in a subset of the corpus data. The following named entity types with sub-types were identified :

- Personal name: including <first_name>, <middle_name>, and <surname>. *Descriptor*: GENDER: <male>, <female>, <unknown>; CASE: <genitive>; INITIALS: in case of initials <ini>.

- Institution: referring to schools, working places, sport team, etc. *Descriptor*: <school>, <work>, <other_institution>.

- Geographic data: country, city, Swedish city, region, geographical areas (e.g. forest, lake, mountain), areas (e.g. city areas, municipalities), street, number (e.g. of building), zip code,

- Transportation: <transport>(e.g. subway, train, bus), <transport_line>(e.g. line no. 3, or green line)

- Age: the person's age given as a random number from a 5-year interval (age:FROM-TO) (e.g. 20 given as age: 18-22)

- Dates: elements directly related to an individual: <day>, <month_digit>expressed as digit (e.g. 5), <month_word>expressed as word (e.g. May), year <FROM–TO>given as a five-year span (e.g. 2018 as 2016–2020).

- Phone numbers: <phone_nr>

- Email addresses: <email>

- Personal web pages: <url>

- Social security numbers: <personid_nr>

- Account numbers: <account_nr>

- Certificate/licence numbers (e.g. vehicle): <license_nr>

- Profession: the person's profession <prof>, or the person's education <edu>

- Sensitive information that might reveal physical and mental disabilities, political views, unique family relations such as a large number of siblings, etc. <sensitive>

- Extra: any other items that are not covered by the previous categories. Distinction is made between objects that need to be replaced because of sensitivity *oblig*, and objects that might be sensitive but can be replaced later *nonoblig*

The list is not exhaustive, and we expect to refine the identified types above as we manually add more texts to the corpus.

Since we want to be able replace the information in the same morphological form as the original written by the learner, morphological features are also added to text strings containing personal information. These include Case: genitive <gen>, Form: definiteness <def>, and Number: <plural>. However, noteworthy that we do not keep track of spelling errors during pseudonymization as these are difficult to replicate in a pseudonymized version.

To keep the information about named entities with the same reference, each unique type (e.g. name or city) gets its own running number, starting with 1. If the particular word is repeated in the text, the same running number is assigned to it.

In the SweLL project, data—where possible—is pseudonymized in two steps: first we mark-up the text string containing personal data token by token on the basis of the named entity types by using a placeholder to keep track of which tokens in the text have been changed; then we replace the marked text string (i.e. placeholder) either by rendering, or by replacement with another token of the same named entity type. In some cases, when the annotator does not know how to categorize a certain text string, the original text is kept but marked by the placeholder, see Figure 2:

1. ORIGINAL TEXT →@PLACEHOLDER →RENDERING

2. ORIGINAL TEXT →@PLACEHOLDER →REPLACEMENT

3. ORIGINAL TEXT →@PLACEHOLDER →ORIGINAL

Figure 2. Pseudonymization steps, three ways to handle personal information, the SweLL approach.

Thus, pseudonymization consists of two distinct steps: 1. first marking up (i) information that directly or indirectly can reveal the author as well as (ii) sensitive information about the author, using *@placeholders*, and then
2. replacing the *@placeholders* by rendering or replacement.

Figure 3 illustrates an example of the pseudonymization tool where the male first name 'Ali' 'firstname:male_1' is identified (marked in red) and marked up as 'firstname:male_1'. Then, the male name 'Ali' is replaced, randomly selected from a list of male names registered in Sweden, in this case by 'Peter'.

This two-step process potentially opens a possibility to set an essay into different cultural contexts, for example by selecting names and cities from a certain country or part of the world. The first case in Figure 2 (1), that is rendering, can be applied to the information that can be collected from general resource lists, such as 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.

However, we need to refine our approach even further, among other things, when it comes to different numerical types of information with different formatting where general resource lists cannot suffice. Thus, the second way of handling personal information, see Figure 2 (2), is replacement, and applies to the cases where we need to replace information directly during the pseudonymization phase. This covers the following cases:

- middle names and initials are replaced with an ”A” for each token used in those names;
- all numerical information (dates, phone numbers, certificate/license numbers, etc) is replaced according to the pattern used in the original, preserving all delimiters, e.g. dates: 2018/01/01 →@DATE_DIGITS →1111/11/11 or phone numbers: 089-777-654-22 →@TEL_NR →000-000-000-00;
- age, both written in digits and in strings. We replace @age with a random number from the range of plus/minus two years from the number provided in the text, for instance a number between 16 and 20 if the original age is 18. However, the complicating moment here is that learners may write the age in strings and make an error with that, so that it needs to be interpreted first by an assistant, and second the number range needs to be provided for the tool to apply a random number selection, preserving only the @placeholder tag in the end. For example:

[ORIGINAL] My elder sister is thirty and my younger sister is *eity.
→[CORRECTION] My elder sister is thirty and my younger sister is eighteen (or eight ?).
→[@PLACEHOLDER + RANGE] My elder sister is @AGE_STRING(28-32) and my younger sister is @AGE_STRING(16-20)
→[RANDOM REPLACEMENT] My elder sister is @AGE_STRING(28) and my younger sister is @AGE_STRING(20)

The third case of handling personal information according to Figure 2 (3) is, in fact, a sub-case of (1), where rendering is not applied. In that case we are marking up a text segment, but do not take any actions until further notice (or rather decision). This covers cases where it is not clearcut whether the information may be considered risky to keep or not. Consider the following examples:

- professions: *I am a web developer.*
- education: *I am taking courses in Linguistics.*
- political or religious views: *We were happy to participate in a demonstration against Erdogan.*
- number of siblings or family members: *I have five sisters and three brothers.*

The different approaches across various learner corpus projects have their advantages and disadvantages. By manually replacing the learner text with strings like *Adam* or *Eva*, there is little chance that the general flow of text will be changed in an unwanted way, that is, the context, the morphological form and imitation of a learner error will be manually taken care of. The necessary prerequisite, then, is to keep track of the tokens that have been manipulated (i.e. not originally written by the learner) for potential post-pseudonymization purposes. However, the possibility of setting a learner text into a different context or other types of studies is lost. Also, they give rise to strings of the following type: *I have three sisters and four brothers. Their names are Eva, Eva and Eva, and Adam, Adam, Adam and Adam.*

In case of @*placeholders* of various kinds (including XML notation) that are preserved in the final text, the readability of the text is hampered, for instance *Hi, my name is @firstname:female_1, I live in @area_2 towards @area_3*. Besides, the possible errors that have been made by the learner are not reflected in this notation, e.g. @*area_2* was originally misspelled as **Stokhulm* (instead of Stockholm).

In case of @*placeholders* that are replaced automatically in the final version or rendered automatically on upload of an essay, on top of the previously described loss of error information, there is a non-negligible chance of

(1) introducing an error that was not originally made by a learner, e.g. *Ukrainians are ...* → *@nationality are ...* → *Swede are ...* where the pseudonymization has failed to preserve the plural form. Another example is *I worked in Charing Cross Hospital* → *I worked in @workplace* → *I worked in Volvo* where the preposition *in* sounds incorrect in a combination with Volvo as a company.

(2) not being able to preserve the forms that a learner has used, e.g. *Alice's wallet was stolen* → *@female_name wallet was stolen* → *Jane wallet was stolen* where the genitive form has not been automatically added and hence an error is introduced into the pseudonymized version. Even though the possessive form seems easy to be fixed, certain languages have rich inflectional morphology - which is impossible to reproduce unless a full morpho-syntactic tag (MSD) is added to the pseudonymized segment, something that makes the manual pseudonymization work by far more complex, error-prone and time-consuming, whereas projecting automatically assigned morpho-syntactic descriptors (MSDs) from automatically annotated original version might be non-straightforward and need further testing for reliability.

There is a trade-off between the benefits of adding the information on errors, MSDs, on lexical and syntactic restrictions (e.g. combinability with prepositions) and common knowledge (e.g. to avoid sequences like *I lived in Berlin, the capital of Venezuela*) and the increased time investment and error rate of doing that.

3.4 Pseudonymization tool in SweLL

During the pseudonymization phase, the research assistants work with essays on a special encrypted hard drive, *SweLL-kiosk*, designed for the purposes of transcription and pseudonymization. The environment does not allow any access to the internet except to a single url-address (i) for reporting technical issues and annotation considerations for discussion with other project members, (ii) for transferring the original essay to a secure data storage outside of anybody's—even project members'—reach and (iii) for transporting pseudonymized essays to an online database, from where any other authorized users can start working on normalization and annotation.

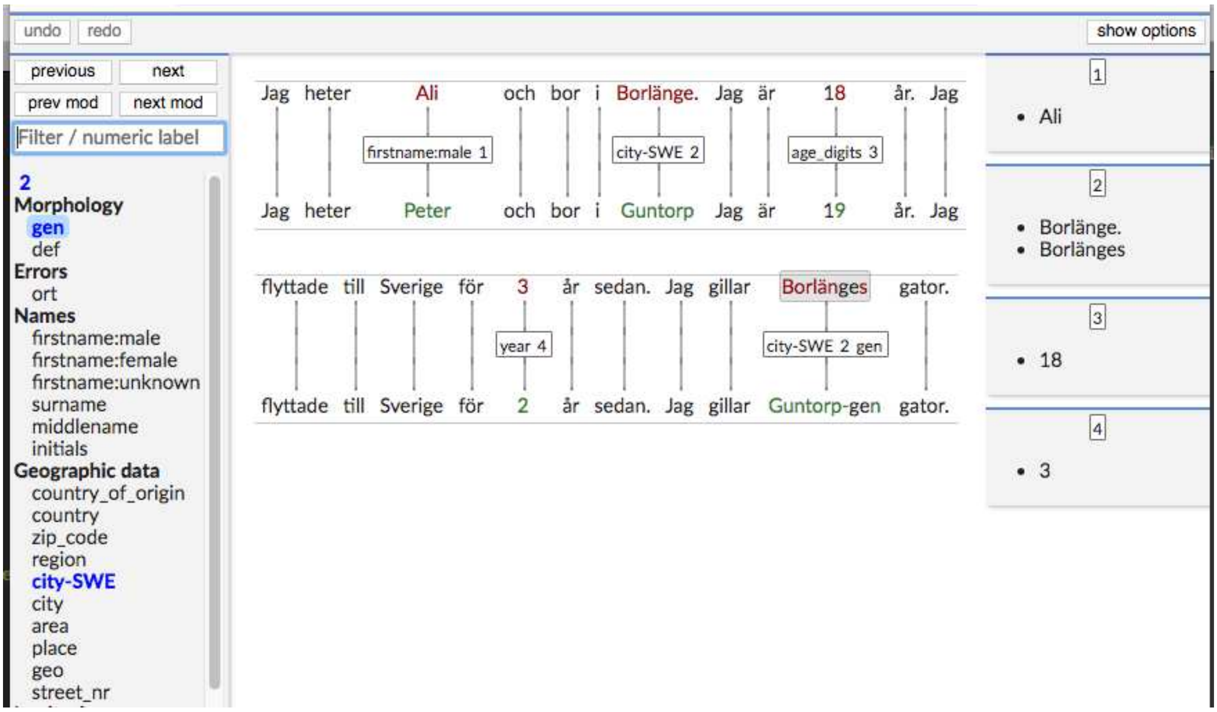
SweLL-kiosks contain a specially designed database and annotation management functionalities, that give an overview over the tasks at hand and completed tasks. On upload of new essays, they are tokenized, and in future we plan to test using full linguistic annotation to explore named entity recognition (NER) for support of anonymization, as well as to evaluate the relevance and benefits of projecting MSDs to the pseudonymized segments. During the work on pseudonymization, continuous versioning is enabled.

All personal information is marked up and masked according to the types described in Section 3.2, using the *SVALA tool* for pseudonymization (Rosén et al., 2018). SVALA links original text to the pseudonymized text building a parallel version with links going from one version to another, token by token. *@placeholder* tags are assigned to the links, as shown in Figure 3. The menu on the left shows a list of *@placeholder* tags, the menu on the right keeps track of unique *@placeholders*.

Data is stored in a JSON format, where information is kept about the source text, the target text, which segments have been manipulated, and the edges between the source and target segments. The edges are displayed as shown in Figure 4, describing the token *Borlänges* and its *@placeholder* label *city-SWE_2*.

To understand the de-identified and masked version of the essay, we keep track of references to the same persons and places, as we described in Section 3.2: if a unique name or place occurs more than once in the text, these are enumerated with the same number, and replaced by a unique pseudonym, as shown in the case of *Borlänge* in Figure 3 which is replaced by *Guntorp* in both places in the text.

The collected data is aimed for research scenarios of many kinds so we mask the absolutely necessary personal information only but without taking any risk of the possibility to identify the person behind the essay. This is not straightforward, and needs manual supervision. Even though we have named entity recognizers that can automatically detect names, places, or numeric expressions (phone numbers, street addresses) with high precision, learner data contains many spelling mistakes, and less well-formed sentences which make these tools less reliable. To guarantee anonymity, we carry out the identification and masking of personal information manually, sentence by sentence,



Jag heter Ali och bor i Borlänge. Jag är 18 år. Jag

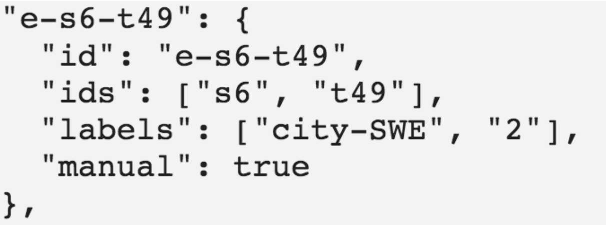
Jag heter Peter och bor i Guntorp Jag är 19 år. Jag

flyttade till Sverige för 3 år sedan. Jag gillar Borlänges gator.

flyttade till Sverige för 2 år sedan. Jag gillar Guntorp-gen gator.

Figure 3: Example of pseudonymization in the SVALA tool.
Gloss-original: *My name is Ali and I live in Borlänge. I am 18 years old. I moved to Sweden 3 years ago. I like Borlänge's streets.* Gloss-pseudonymized: *My name is Peter and I live in Guntorp. I am 19 years old. I moved to Sweden 2 years ago. I like Guntorp's streets.*

essay by essay.



```
"e-s6-t49": {
  "id": "e-s6-t49",
  "ids": ["s6", "t49"],
  "labels": ["city-SWE", "2"],
  "manual": true
},
```

Figure 4: SVALA data format for edges.

In addition, the learners might write personal information in several essays which altogether might reveal the identity of the learner. To prevent such cases, we manually check all the essays written by a specific learner.

Once the text is pseudonymized, the de-identified essay is moved from the encrypted environment to the online SweLL portal for further processing, to normalize, correct and annotate the text accordingly.

4 Conclusions and future outlook

We presented on-going work on building a research infrastructure for Swedish as a second language with the focus on pseudonymization of learner essays. We described the legal issues influencing the way data needs to be handled and manipulated to ensure anonymity of data subjects, i.e. learners providing us with essays. This influences the way the data is collected, stored and pseudonymized. We gave an overview of the taxonomy for pseudonymization and presented the approaches and tools used for that.

The corpus is under development, as are the tools, and we envisage a number of experiments in order to

- add rendering functionality to our SVALA pseudonymization tool, and prepare resources that can be used for that,
- evaluate the necessary constraints—linguistic and extralinguistic—to ensure logical rendering, so that we do not get strings of the type *I lived in Berlin, the capital of Venezuela*,

- evaluate NER for support of manual pseudonymization, and

- evaluate projecting MSDs for keeping track of grammatical and orthographical choices made by learners.

We expect the corpus and the tools to be released as open source by the end of 2020.

To date there are no systematic studies that focus on the questions of the influence of pseudonymization of learner corpora on readability, text fluency, reader attitudes, assessment and annotation quality, or how it is best to render personal or potentially sensitive information. Nor does there seem to be tools that exploit automatic methods, e.g. Named Entity Recognition, for fully or semi-automatic learner text pseudonymization. All of which opens a whole new field for research.

Acknowledgements

We are very grateful to the two anonymous reviewers for useful comments and suggestions for improvements on the draft of our paper. This work has been supported by an infrastructure grant from the Swedish Foundation for Humanities and Social Sciences (Riksbankens Jubileumsfond: SweLL - research infrastructure for Swedish as a second language, project IN16-0464:1).

References

Accenture. 2016. *Building digital trust: The role of data ethics in the digital age*. https://www.accenture.com/t20160613T024441__w__/us-en/_acnmedia/PDF-22/Accenture-Data-Ethics-POV-WEB.pdf.

Malin Ahlberg, Lars Borin, Markus Forsberg, Martin Hammarstedt, Leif-Jöran Olsson, Olof Olsson, Johan Roxendal, and Jonatan Uppström. 2013. Korp and Karp - a bestiary of language resources: the research infrastructure of Språkbanken. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 429–433.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

Ben Medlock. 2016. An Introduction to NLP-based Textual Anonymisation. In *Proceedings of Language Resources and Evaliation*, pages 1051–1056.

Nives Mikelić Preradović, Monika Berać, and Damir Boras. 2015. Learner Corpus of Croatian as a Second and Foreign Language. In *Multidisciplinary Approaches to Multilingualism*. Peter Lang.

Riksdagen. 1949. *Tryckfrihetsförordningen (1949:105)*. http://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/tryckfrihetsforordning-1949105_sfs-1949-105.

Alexandr Rosen. 2017. Introducing a corpus of non-native Czech with automatic annotation. *Language, Corpora and Cognition*, pages 163–180.

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*.

Kari Tenfjord, Paul Meurer, and Knut Hofland. 2006. The ASK corpus: A language learner corpus of Norwegian as a second language. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1821–1824.

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. In *Proceedings of SLTC 2018, Stockholm, Sweden*.

Elena Volodina, Beáta Megyesi, Mats Wirén, Lena Granstedt, Julia Prentice, Monica Reichenberg, and Gunlög Sundberg. 2016a. A Friend in Need? Research agenda for electronic Second Language infrastructure. In *Proceedings of SLTC 2016, Umeå, Sweden*.

Elena Volodina, Ildikó Pilán, Lars Borin, and Therese Lindström Tiedemann. 2014. A flexible language learning platform based on language resources and web services. In *LREC*, pages 3973–3978.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016b. Swell on the rise: Swedish learner language corpus for European reference level studies. *Proceedings of LREC 2016*.