Annotation of learner corpora: first SweLL insights

Elena Volodina[1], Lena Granstedt[2], Beáta Megyesi[3], Julia Prentice[1], Dan Rosén[1], Carl-Johan Schenström[1], Gunlög Sundberg[4], Mats Wirén[4]

[1]University of Gothenburg (Sweden), [2]Umeå University (Sweden)
[3]Uppsala University (Sweden), [4]Stockholm University (Sweden)
elena.volodina@gu.se

1. Introduction

SweLL - Swedish Learner Language - is a project aimed at setting up an electronic infrastructure for collecting, annotating, browsing and analyzing Swedish learner language (Volodina et al., 2016). During the first year of the project, a number of the project aims have been addressed, such as

1. legal and ethical aspects of essay collection
2. principles of learner language annotation
3. tools and platforms for securing the previous steps

As the practice shows, annotation of learner texts is a very 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 the SweLL project with numerous evidence of the above-mentioned compromises[1].

2. Legal issues and their consequences

Spreading an electronic resource through an infrastructure entails responsibility to the data subjects, in our case language learners, who have agreed to provide their texts and personal information. The requirement of collecting and storing informed consents, obligation to remove a learner and their data from the registers if they desire so as well as national and international laws and ethical regulations regarding personal integrity and discrimination create certain difficulties in making the data open for all types of uses. To argue for the data to be accessible to users outside individual projects, handling of data should be 'bulletproof' at each stage and there are several stages to consider, namely, data acquisition, data storage, data aggregation, data analysis, data usage, data sharing and data disposal (Accenture, 2016). Most of the steps deal with organizational and management decisions/precautions or preparatory steps before uploading data to the infrastructure. In the text below, we concentrate on the stages relevant to infrastructure usage where learner specific characteristics in the texts and metadata present risks at the data usage and data sharing stages.

To start with, within European countries, there is a requirement to ensure personal non-identifiability when adding essay information with personal metadata. According to the EU General Data Protection Regulation (GDPR), Article 4[2], "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..." (Commission, 2016, art.4). Consider Figure 1, where adding up information from the two sources – a learner text and socio-demographic metadata – can give away a learner. Even though the name as such is not revealed to the data users, indirect clues can be used to identify a person.

SOCIO-DEMOGRAPHIC METADATA

- L1: Luxembourgian, Chinese
- Year of birth: 1986
- Gender: male
- Education / highest degree: MA
- Time in L2 country: 3 years
- Other languages: Russian, Korean, German, French

TASK METADATA:

- Date: April 2018
- CEFR level: B1

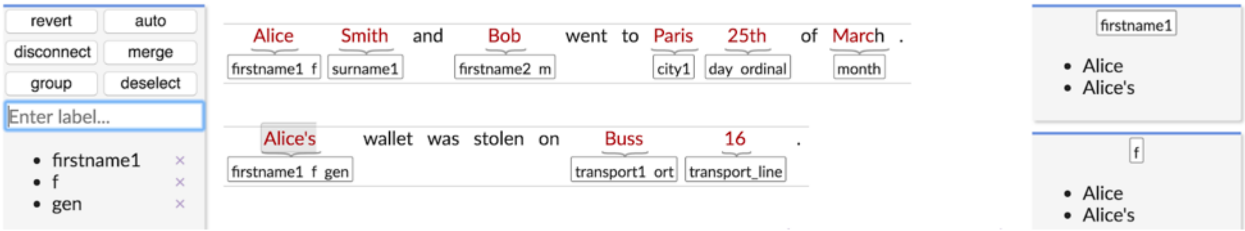
TEXT:
I lived in Denmark before, in Svaneke. It was less thenn Berlin. I like there too because I had more friends. But I have better work here. In Svaneke job was on one webpage. In Berlin I work on many webpages. I am web develooper. But Berlin is closer to Louxembourg that Svaneke.

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

In view of this, unlike a number of learner corpora projects, the SweLL project adopted a rather restrictive approach to metadata. For instance, it does not provide a student's country of origin or nationality (restricting information to the mother tongue (L1) only), nor the year of birth, but rather a 5-year span (e.g. 1970–1974), to complicate possible identification of a learner through aggregated personal information. For the same reason, no information is provided on the educational establishment where the essays

[1] Parts of Sections 2 and 3 have originally been written by the abstract co-authors for the article by Stemle et al. (2019) and are re-used with the permission of the LCR volume editors

[2] https://gdpr-info.eu/art-4-gdpr/



Alice Smith and Bob went to Paris 25th of March .

Alice's wallet was stolen on Buss 16 .

Figure 2: Anonymization compact view in SweLL anonymization tool (Rosén et al., 2018)

have been collected. This comes as a natural consequence of, on the one hand, the national Swedish legislation on open access to public data (Riksdagen, 1949, ch.2), and on the other, the stricter current European legislation on personal data integrity (Commission, 2016).

Ethical Review Boards set further requirements on the so called sensitive data, i.e., data that can reveal a (potentially identifiable) person's sexual orientation, religion, political views or ethnicity, which may lead to discrimination. Unless it can be ensured that the person behind the (meta)data will not be revealed, Ethical Review Boards are entitled to require an application which should list all potential scenarios for data usage, moreover restricting data usage to internal use only, within the project. This in itself is counterproductive since a research infrastructure is aimed at providing electronically available data to researchers outside the project for any potential research questions that cannot be foreseen in advance.

To make learner data less "sensitive" (according to the Ethical Review Boards' definition) as well as to minimize personal identifiability from a text, learner essays need to be anonymized, so that information in the actual text that may give away the author, is either substituted/pseudonymized (e.g. Poland →Greece); made noisy (e.g. Poland →Europe); or completely removed, see text in Figure 1 where a lot of personal information is provided. Whereas suggestions for anonymization of "structured" or "listed" types of personal information (e.g., personal names, city names, telephone numbers, etc.) can be supported through use of automatic methods as adopted from the medical domain (El Emam and Arbuckle, 2013), "unstructured" types of potentially sensitive information (e.g. *We were happy to participate in a demonstration against Erdogan*) will still need to be marked up manually.

In the SweLL project, data is anonymized in two steps – first manually marking up (1) information that directly or indirectly can reveal the author as well as (2) sensitive information about the author, and then rendering the 'placeholders', e.g. 'firstname1' in Figure 2, according to an associated algorithm. Thus, for 'firstname1 f' a female name will be randomly selected from a list of names registered in Sweden. 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. However, the question of the influence of anonymization on readability, reader attitudes and assessment is still an open one, as well as how it is best to render personal or potentially sensitive information.

To secure a safe environment for anonymization, a special solution has been developed in the SweLL project, called SweLL-kiosk. A SweLL-kiosk is an encrypted environment that protects unauthorized users to get access to the non-anonymized versions of the essays. Kiosks are equipped with a project management system, a database for storing all versions of the files, and a simplified version of SVALA, SweLL annotation tool, containing anonymization functionalities. Essays that have been anonymized, are exported from the kiosk database to Språkbanken's databases.

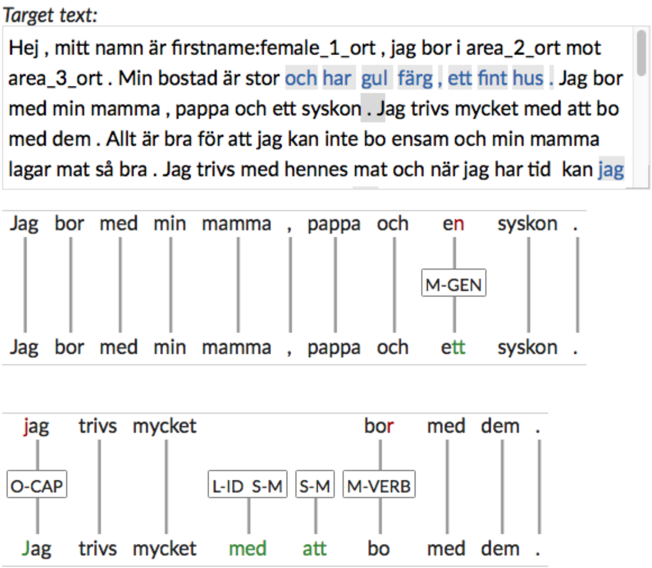
3. Normalization and error annotation

Annotation of a standard corpus follows a number of steps including tokenization, morphosyntactic tagging, lemmatization and parsing, all of them assuming a standard language. However, a learner corpus includes texts exhibiting deviations from the standard version of the target language for which the tools have been designed. While standard language can be relatively accurately annotated with existing automatic methods, annotating learner language with the same tools is more error-prone due to various (and often overlapping) types of errors, as in e.g. **I has was* (morphology and agreement) or **We wrote down it* (word order).

Automatic tools aimed at standard language can sometimes be applied with more or less satisfactory results even to learner language. Where available, spelling or grammar checking tools providing suggestions can be used to approximate a corrected version of the text. Alternatively (and more commonly), an additional manual step is added, namely *normalization* which means rewriting the original learner text to a grammatically correct target hypothesis (Lüdeling et al., 2005), before applying a standard annotation pipeline. Most projects, further, combine normalization with *error-annotation*, i.e. labelling the type of change that has been applied to the original text. In SweLL, the two steps - normalization and error-annotation - are separated as conceptually independent ones.

3.1 Normalization

Normalization entails interpretation of intentions of the author, which on many occasions is difficult to make. Consider the following example: **jag trivs mycket bor med dem* (Eng. I enjoy live with them) (see Figure 3). Applying the main principle of normalization that *any change to a grammatically correct version should be as minimal as possible*, i.e. THE PRINCIPLE OF MINIMAL CHANGE, the seemingly best way would be to change the original sequence to *Jag trivs mycket bra med dem*, that is, *bor* →*bra*. However, this change does not reflect objectively the knowledge of the learner, namely usage of the verb *att bo* versus the adjective *bra*, with *bra* being used correctly by the learner in the other parts of the text. The referenced minimal change does not seem to reflect the semantics that the learner is trying to convey, either. The Second Language Acquisition (SLA) researchers involved in the SweLL project were unanimous about changing this sentence to *Jag trivs mycket med att bo med dem*.



Target text:

Hej , mitt namn är firstname:female_1_ort , jag bor i area_2_ort mot area_3_ort . Min bostad är stor och har gul färg , ett fint hus . Jag bor med min mamma , pappa och ett syskon . Jag trivs mycket med att bo med dem . Allt är bra för att jag kan inte bo ensam och min mamma lagar mat så bra . Jag trivs med hennes mat och när jag har tid kan jag

Jag bor med min mamma , pappa och en syskon .

Jag bor med min mamma , pappa och ett syskon .

jag trivs mycket bor med dem .

Jag trivs mycket med att bo med dem .

Figure 3: Original and normalized versions of a learner text, with error tags added on the edges. Gloss of the original layer (with some imitation of the errors): *I live with my mother, father and e syster . i enjoy live with them.* The question is, should *live* be changed to *living* or *(my) life*?

Error annotation that is applied to the corrected version is in fact NOT about labeling errors that a learner has made. It rather reflects the difference between the original and normalized versions, and depends upon which normalization variant is accepted. It makes the normalization step extremely important. In the example with the two correction versions of the sentence **jag trivs mycket bor med dem*, error labels could describe either a spelling correction (*bor* →*bra*) or, as we see in Figure 3, a wrong form of a verb (*bor* →*bo*) plus idiomaticity problem in using the verb *att trivas* (*trivs* →*trivs med att*). As such, we cannot claim that we are error-labeling the learner language. We are labeling the type of correction we have introduced.

Several experiments with normalization and error-annotation within the SweLL project have proven that normalization as a separate step is a conceptually right way to go for several reasons:

- It helps to build a better understanding of a learner's linguistic competence (e.g. that (s)he is able to spell the adjective *bra* correctly) so that the changes in the normalized version would take that into account.
- It can be outsourced to SLA researchers for doing it, since (1) normalization takes much less time compared to error-annotation and thus can be done quickly, and (2) SLA researcher reasoning rests on a basis of competence in the SLA field and experience with second language learners, whereas project assistants, who are often L1 students within linguistics, do not have this type of insights into learner language.
- Error annotation depends on the change applied to the original text, and thus should rather start from comparison of the two versions (in contrast to adding error labels at the same time as normalizing a text segment).
- Inter-annotator agreement with respect to error codes can be objectively measured only given that the annotators are working on the same normalized version.

3.2 Error annotation

We start this section with an anonymous quotation: "Taxonomies are like underwear; everyone needs them, but no one wants someone else's." With respect to error annotation projects, this is both true and false. Even though so far very few learner corpus projects have managed to reuse each other's error taxonomies, several projects have tried to build on previous work. Let us demonstrate the problems of re-using someone else's taxonomy with an example from the SweLL project.

Since the SweLL project is in an early stage, there is a direct incentive to learn from the experience of other projects to ensure a certain degree of comparability. In this respect, the SweLL project has looked into some error annotation taxonomies, namely of ASK (Tenfjord et al., 2006) and MERLIN (Boyd et al., 2014).

The initial SweLL tagset was a result of testing the ASK taxonomy (23 tags) and the MERLIN taxonomy (64 tags) on a set of Swedish essays. It turned out that annotating with the highly intricate MERLIN taxonomy took twice as much time as with the ASK taxonomy, leaving a lot of inter-annotator disagreements. As a result of this experiment, the ASK taxonomy has been adopted with several modifications and was tested in a pilot study with the involved researchers. Once again, practical usage of the taxonomy led the SweLL researchers to important insights with reference to tag names and their coverage. See for example Figure 4, where three annotators agreed on both the segment in need of correction (top row) and on the target hypothesis (second row), but not on the error label (O, INV, OINV describing various types of word order errors). Consequently, both the tag names and the number of tags have been reviewed to avoid ambiguity – leaving very little of the original ASK taxonomy as a result.

The strongest argument for reviewing the ASK taxonomy was the possible drop in annotation quality unless the tagset is reduced or changed, an idea also supported in previous annotation projects (Fort, 2016).

To support normalization and error-annotation in a parallel fashion, a tool SVALA has been developed (Rosén et al., 2018) which is now undergoing an extensive testing in its beta version.

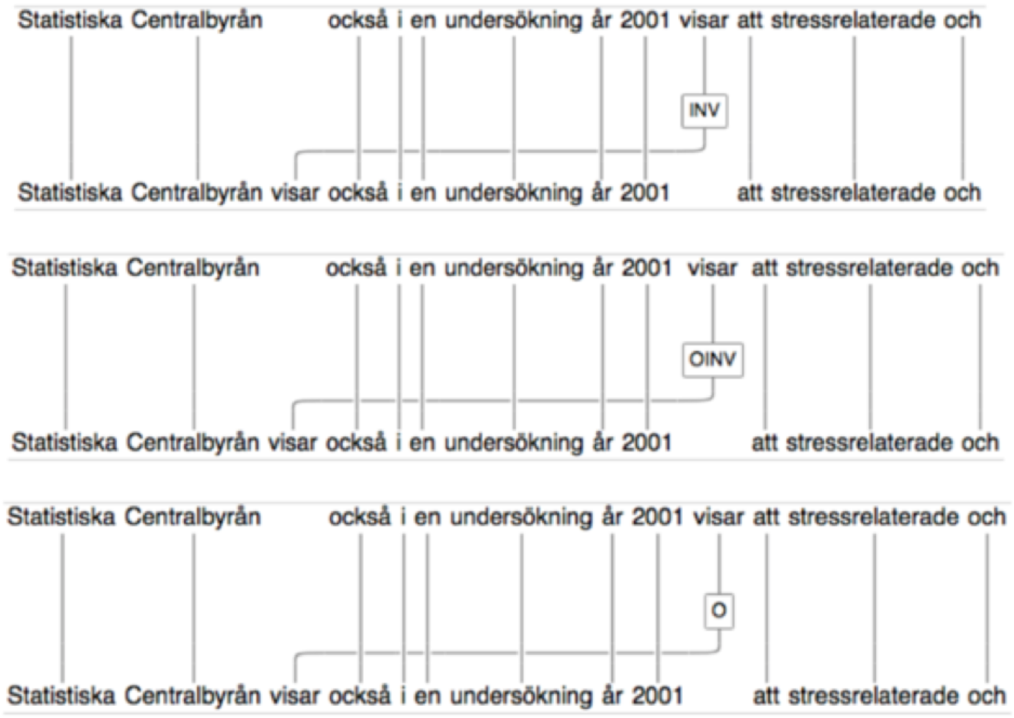

Statistiska Centralbyrån också i en undersökning år 2001 visar att stressrelaterade och
INV
Statistiska Centralbyrån visar också i en undersökning år 2001 att stressrelaterade och
Statistiska Centralbyrån också i en undersökning år 2001 visar att stressrelaterade och
OINV
Statistiska Centralbyrån visar också i en undersökning år 2001 att stressrelaterade och
Statistiska Centralbyrån också i en undersökning år 2001 visar att stressrelaterade och
O
Statistiska Centralbyrån visar också i en undersökning år 2001 att stressrelaterade och

Figure 4: Inspecting error annotation done by three annotators, SweLL error annotation pilot
Gloss: Central Statistical Agency [...] also in a report from 2001 [shows] that stress-related and…
Error code explanations: *INV* Non-application of subject/verb inversion, *OINV* Application of subject/verb inversion in inappropriate contexts, *O* word (or phrase) order error

4. Future prospects

To summarize, the SweLL infrastructure has been extensively developing towards opening a possibility for continuous collection and annotation of learner essays. So far three pilot studies have been carried within the project group, with the aim to produce high quality guidelines, non-ambiguous tag sets and top performing tools. The work is still ongoing. A full scale annotation of essays is planned for 2019.

Next, SweLL will look into the necessary functionalities for visualizing, browsing and statistically analyzing learner corpora - to make learner texts as accessible for SLA research as possible.

References

Accenture. 2016. *Building digital trust: The role of data ethics in the digital age.*

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014. The merlin corpus: Learner language and the cefr. In *LREC*, pages 1281–1288.

European Commission. 2016. *General data protection regulation.* Official Journal of the European Union, 59, 1-88.

Khaled El Emam and Luk Arbuckle. 2013. *Anonymizing health data: case studies and methods to get you started.* ” O’Reilly Media, Inc.”.

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

Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. *Proceedings of corpus linguistics 2005*, 1:14–17.

Riksdagen. 1949. *Tryckfrihetsförordningen (1949:105).*

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.*

Egon W. Stemle, Adriane Boyd, Maarten Janssen, Therese Lindström Tiedemann, Nives Mikelić Preradović, Alexandr Rosen, Dan Rosén, and Elena Volodina. 2019. Working together towards an ideal infrastructure for language learner corpora. *Learner Corpus Research 2017, post-conference volume.*

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