Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

The SweLL Language Learner Corpus: From Design to Annotation

Elena Volodina[1], Lena Granstedt[2], Arild Matsson[1], Beáta Megyesi[3], Ildikó Pilán[1], Julia Prentice[1], Dan Rosén[1], Lisa Rudebeck[4], Carl-Johan Schenström[1], Gunlög Sundberg[4] and Mats Wirén[4]

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

Abstract

The article presents a new language learner corpus for Swedish, SweLL, and the methodology from collection and pesudonymisation to protect personal information of learners to annotation adapted to second language learning. The main aim is to deliver a well-annotated corpus of essays written by second language learners of Swedish and make it available for research through a browsable environment. To that end, a new annotation tool and a new project management tool have been implemented, – both with the main purpose to ensure reliability and quality of the final corpus. In the article we discuss reasoning behind metadata selection, principles of gold corpus compilation and argue for separation of normalization from correction annotation.

1 Introduction

Standard automatic corpus annotation follows a number of steps, including tokenization, PoS-tagging, lemmatization and syntactic parsing. The project on learner language requires us to handle texts exhibiting a great amount of deviation from standard (target) language. While texts with normative target language can be relatively accurately annotated with existing automatic methods (e.g. Megyesi et al., 2016), annotating learner language with the same tools is error-prone due to various (and often overlapping) orthographic, morphological, syntactic and other types of errors, see Figure 1.

It can be argued that texts written by more or less advanced learners can still be relatively reliably annotated automatically (Rosen, 2017; Abel et al., 2014), however, the beginner learner language causes a rather non-negligible drop in reliability of automatic annotation (e.g. Geertzen et al., 2013; Stymne et al., 2017). We argue, therefore, that, before applying standard annotation pipeline, there is a need to add an extra manual step which we call interchangeably "normalization" or "target hypothesis", the latter being the term first introduced by Lüdeling et al. (2005).

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

```
original:  till exampal när jag bor i lite rumet jag käner inte
bra samma lika nu

correct:  Till exempel när jag bodde i ett litet rum mådde jag inte
lika bra som nu

gloss:  For example when I lived in a small room I didn't feel
as good as now
```

Figure 1: Example of deviations in learner language.

The manually performed normalization takes care of the anomalies of learner language where different types of deviations and errors are re-written to represent a standard variant, so that the automatic annotation in the next step is accurate. Normalized texts are further used for training tools that can perform (parts of) normalization and error detection automatically. This sets a high requirement on the quality of manual work, both as far as normalization, correction annotation and linguistic annotation are concerned, since the manual annotation has of late become the only bit of linguistic knowledge hidden in the modern NLP approaches (Fort, 2016, p.xiii), otherwise dominated by discreet machine and deep learning methods that are rather opaque and uninterpretable in linguistic terms (Church, 2017; Doshi-Velez and Kim, 2017).

Thus, we start from a standpoint that high reliability of annotation is a prerequisite for the data to be useful for various future research projects. The (manually) annotated learner data is extensively used for research, among others within Natural Language Processing (NLP), Learner Corpus Research (LCR) and Second Language Acquisition (SLA), and thus the annotation should be both reliable and reproducible. In addition, it should also be comparable between different corpora and other studies (including those not involving electronic corpora), so that conclusions drawn from the data are also reliable and generalizable. To ensure the quality of annotation, there are multiple aspects that need to be considered, among others the size and ambiguity of a tagset, the tools that should prevent one from making (certain types of) mistakes, regular inter-annotator checks, etc., see Fort (2016) and Hovy and Lavid (2010) for an overview.

The above means that in case we expect the current corpus to be of use in SLA, LCR and NLP research (and be cited as a resource), we need to produce a *gold standard* corpus in the best sense of this word, which in turn means that

- corpus design and representativity have to be thought through (Sections 3, 4),
- metadata should be (as good as) exhaustive for the needs of the target research groups (Sections 3),
- manual annotation should be reliable in relation to its complexity (Sections 5, 6),
- and - not least - the resource should be made available (Sections 3.1, 5.2).

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

2 Second language infrastructure project SweLL

SweLL stands for **Swe**dish **L**earner **L**anguage, and is an acronym for an ongoing infrastructure project[1] financed by Riksbankens Jubileumsfond (RJ) 2017-2020. The purpose of SweLL is to lay the main building blocks of an infrastructure in support of Second Language Acquisition (SLA) research on Swedish. This entails preparing and releasing necessary resources and tools for continuous collection, digitization, normalization, and annotation of texts written by learners of L2[2] Swedish. By the end of the project a linguistically annotated corpus consisting of approx. 600 learner texts spanning various levels of linguistic proficiency should be available for researchers alongside with tools for automatic processing, searching and downloading these texts (Volodina et al., 2016).

Texts are collected from adult (16+) second language learners of Swedish either going through *formal education* in Swedish as a Second Language, e.g. at institutions offering Swedish For Immigrants (SFI)[3], preparatory university or upper-secondary courses; or testing their knowledge through *official tests*, e.g. Tisus[4] or CEFR-based ones[5]. Texts are collected together with certain demographic information about the writers, metadata about the tasks they are performing, and the writers' performance on the task.

Texts are undergoing several steps of manipulation, namely, transcription, pseudonymization, normalization, correction annotation, and automatic linguistic annotation before they are made available in a search environment. The final infrastructure consists of:

1. a data collection portal, through file import and via online platform,
2. an annotated corpus of L2 production,
3. methods and tools for L2 analysis,
4. specific search tools for L2-material facilitating filtering for e.g. texts written by male writers, writers of a certain mother tongue, or writers at a certain proficiency level.

The material and tools will be made accessible through Språkbanken[6], partly through the learning platform Lärka (Volodina et al., 2014), and partly through Korp (Ahlberg et al., 2013) and Strix - two tools under development at Språkbanken - for browsing texts and visualization of statistics and analytics.

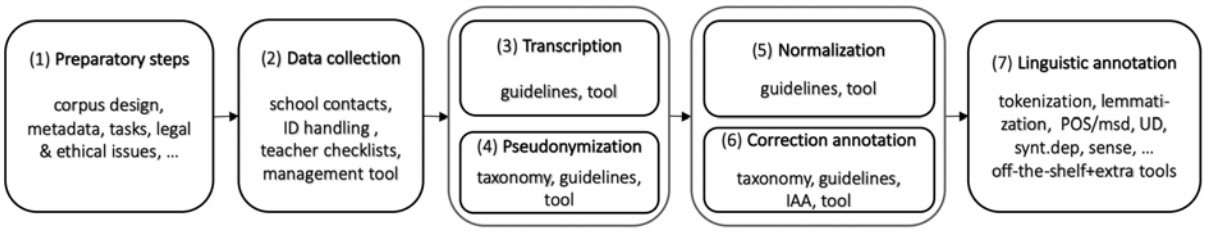

Figure 2: SweLL corpus-building steps.

The SweLL workflow, in terms of corpus-building, involves seven steps, illustrated in

[1] `https://spraakbanken.gu.se/eng/swell_infra`
[2] The abbreviation *L2* covers both second and foreign language(s)
[3] SFI offers courses in Swedish from beginners' to intermediate levels
[4] Tisus stands for Test In Swedish for University Studies
[5] CEFR stands for Common European Framwork of Reference for Language Learning
[6] `https://spraakbanken.gu.se/`

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

Figure 2:
1. Preparatory steps (Section 3), i.e. pre-collection decisions, including the focus learner group, legal considerations, corpus design, metadata of various types;
2. Data collection (Section 4);
3. Transcription and pseudonymization (Section 5);
4. Normalization and correction annotation (Section 6);
5. Linguistic annotation (Section 7).

The tasks of pseudonymization, normalization and correction annotation are carried out using the tool SVALA[7] (Rosén et al., 2018; Wirén et al., 2019), while the project management, upload and export of essays are performed through SweLL portal. Section 8 introduces shortly SVALA and SweLL portal, the two tools developed in the SweLL porject (see Section 8 for more details).

2.1 Securing annotation quality

Building on the arguments in Section 1 about the importance of high quality and reliability of annotation, we perform our manual annotation, i.e. transcription, pseudonymization, normalization and correction annotation, based on the suggested annotation flow in Fort (2016) and Hovy and Lavid (2010). An adapted flow of that work is presented in a chart form in Figure 3. Steps in blue (gray in a printed version) are specific for Fort (2016), the rest of the flow is recommended by both sources.

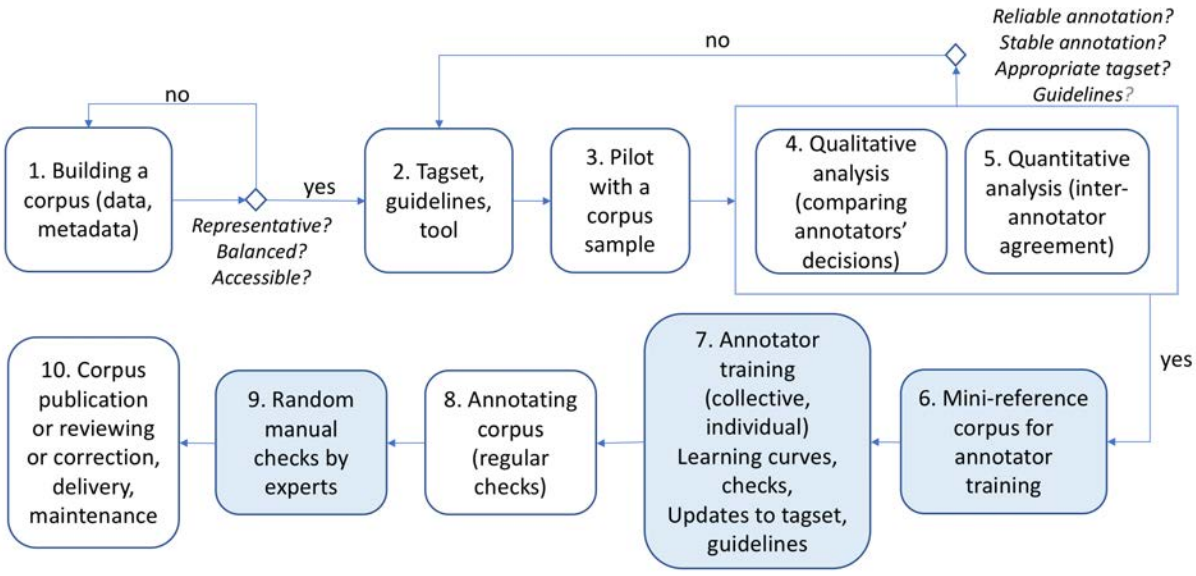

Figure 3: Adopted annotation flow for quality control in SweLL manual annotation, based on Fort (2016) and Hovy and Lavid (2010). Steps in blue (gray in a printed version) are not present in the model described in Hovy and Lavid (2010).

[7] "Svala" in Swedish means "swallow" (noun, in the sense of the bird), but is here an acronym which is a concatenation of SVA (*SV*enska som *A*ndraspråk; in English: *Swedish as a Second Language*) and LA (*L*inking and *A*nnotation).

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

Quality and reliability of annotation depend on the theory-neutral approach, cross-disciplinary insights, taxonomy that is clear and non-overlapping, clear guidelines and instructions, user-friendly annotation tools, and quality checks during the process. In SweLL, steps 2-5 (Figure 3) are repeatedly run in the form of pilots with the project group, to ensure the best possible quality of the final annotation. If we, project researchers, cannot agree on certain annotation decisions, the obvious conclusion is that the guidelines need to be improved and the tag sets need revision. An iterative approach in the form of pilots has proven to be very powerful, and helped us gain a lot of insights into the annotation work.

Several pilots have been run in the project, targeting error taxonomies and the annotation tool (pre-pilot and pilot 1), the flow of essay collection including the SweLL portal functionalities (pilot 2), mini-reference corpus annotation and quality of the guidelines (pilots 3 and 4). In the sections below we present the corpus preparation steps in more detail, following the outline in Figure 2 and describing insights from the four pilots.

3 Preparatory steps

The focus of the SweLL infrastructure at the present stage of development is on grown-up learners of L2 Swedish and their written production. This focus can be explained by the need to first set up fundamental building blocks for groups of learners and L2 data types that are abundant and relatively easy to collect - from both legal points of view (i.e. avoiding parental and school involvement), as well as motivated by the technical considerations (e.g. speech is a much more cumbersome type of data to collect and process). In the future, the scope of the infrastructure may grow further, together with the expertise in the legal and technical questions.

The intention with an electronic infrastructure for research on second language acquisition is that the offered data and tools should help answer research questions outside the current project, and thus a certain understanding about what is interesting for the intended target user group is unavoidable.

To name just a few pretty broad needs, SLA research has always been in want of longitudinal L2 data underlying mental representations and developmental processes (Myles, 2005), among others, for studying linguistic features that are characteristic of various stages of L2 proficiency (Housen and Kuiken, 2009), interplay of complexity, accuracy and fluency (Pallotti, 2009), for tracing effects of tasks on language production (Alexopoulou et al., 2017) or cross linguistic inferences between the mother tongue and the target language (e.g. Golden et al., 2017). NLP researchers have been in need of annotated learner corpora to be able to train tools for spelling and grammar error identification and correction (Leacock et al., 2010), for identification of writer's first language (Tetreault et al., 2013), for classification of essays by grades or levels (Östling et al., 2013; Pilán et al., 2016), for development of supportive writing tools (Madnani et al., 2018), learner modeling (Settles et al., 2018), and many other areas of application. LCR has always had a strong connection to language pedagogy and has been aiming at contributing to the development of teaching and teaching materials (Granger, 2009). Among others, LCR has been concerned with development of vocabulary and grammar in learner language (Paquot and Granger, 2012; Forsberg and Bartning, 2010), first language transfer in L2

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

language (Paquot, 2013), effects of feedback on L2 learner production (Hyland and Hyland, 2019), proficiency construct in learner corpora (Carlsen, 2012) and a range of other questions.

Recently, analysis of the bulk of journal publications within the SLA and LCR has revealed that generalizations about second language acquisition are mostly based on convenience samples, namely on learners that researchers have the easiest access to – university students (Paquot and Plonsky, 2017). To balance this bias, calls for replication studies targeting other learner groups started to appear (e.g. Andringa and Godfroid, 2019). Another trend worth mentioning is a shift of focus from the laboratory-staged experiments to more naturalistic instructed situations, i.e. classrooms, and collection data of more varied formats, i.e. not only featuring essays (Meurers et al., 2019; Norris and Ortega, 2009).

The SweLL corpus we present in this article can help answer (at least) a number of the above-mentioned questions for L2 Swedish. This is featured through the type of metadata on learners and tasks, the targeted learner groups and schools, the annotation schemes applied, as well as the data format that opens up for comparisons in a parallel way.

3.1 Legal issues and data management

In order to make a language infrastructure useful, the data shall be publicly available and distributed outside the project, as often also required by the funding agencies that finance research and infrastructure projects. However, sharing electronic language resources such as text corpora requires that cautions are taken to protect the identity of the data subjects - to comply with the EU General Data Protection Regulation (GDPR). In our study, GDPR applies to language learners, who have agreed to provide their texts and information about themselves to the research infrastructure. We need to take precautions from several aspects to minimize the risk of illegal or unethical use of the data before we can make the data available and accessible to the public.

One way to protect the learners' identity and ensure their anonymity would be to eliminate all personal information about the subjects. However, this is not applicable for several reasons. First, according to GDPR, we have the obligation to remove all data related to a particular learner from our register if (s)he so desires. Hence, we need to store personal information about the learners (i.e. mapping keys between names and student IDs) for a long time, longer than the project is financed. Second, we shall provide certain demographic information about the learner for research purposes allowing studies of various kinds. Hence, we need to keep the information about the learners, but ensure that the learner's identity is hidden from users of the language infrastructure.

Personal information might include person names, age, occupation, institution, work place, areas, cities, countries, dates, or sensitive information such as religious or political views, related to an individual. To be able to publicly release the corpus and at the same time protect the learners' identity, we take precautions where learner specific characteristics appear. These can be found not only in the metadata about the learner, but also in the learners' texts, as illustrated in Figure 4.

We take a rather restrictive approach to *demographic metadata* to minimize the risk of the identification of the learner, and only collect information that we believe is needed for

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

SOCIO-DEMOGRAPHIC METADATA

- L1*: Hungarian, Chinese
- Year of birth: 1990
- Gender: female
- Education / highest degree: BA
- Time in L2 country: 5 years
- Other languages: Arabic, English, Japanese, Russian

TASK METADATA:

- Date: April 2019
- CEFR level: B2

TEXT:
I am *29* years old. I lived in *Hungary* and *China* before, and I came to Sweden on *April 24, 2014*. I live in *Sigtuna* and work as a *bus driver*.

Figure 4: Example of some metadata and an essay text for a fake learner, personal information in the learner's text marked in red italics. **L1* stands for *first language*, i.e. mother tongue

long-term research purposes. For example, we do not provide information about learners' country of origin or nationality but we ask for their first language(s). Nor do we ask for age or the year of birth, but rather a 5-year span (e.g. 1990—1994). This is to complicate possible identification of a learner through aggregated personal information. The entire set of metadata is described in Section 3.3.1.

Personal information—often occurring in *learners' texts*—is also pseudonymized by manually identifying, labeling and eventually masking the information that directly or indirectly can reveal the learner's identity. The pseudonymization process is described in Section 5, with further details provided in Megyesi et al. (2018).

On completion of pseudonymization or other data-protective anonymization techniques, in other disciplines a technique known as a *motivated intruder test*[8] or a *re-identification test* has been used to assess the risk levels of a person being identified in a dataset using all sources of information at hand. The risk assessment based on that can, then, regulate the level of access to the dataset to be granted to the public or researchers. While we haven't yet performed any such tests, and are not aware of any learner corpora using this technique, this is an idea that we would like to explore after a major bulk of SweLL essays have been pseudonymized.

3.2 Corpus design

A usable corpus should be based on strict, transparent and systematic design criteria (Granger et al., 2015; Hovy and Lavid, 2010). It is also important to keep good record

[8]https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/guidanceonintrudertesting

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

of data and metadata and to reflect on the future needs of the corpus. The SweLL language learner corpus accordingly should contain specific variables that are of interest for language learner research as well as for pedagogical purposes; and it should be possible to collect a valid and reliable subcorpus for different purposes. Despite these measures, it can be difficult to foresee needs, uses and potential research questions in advance.

One significant variable for a learner corpus is language development, documented cross sectionally or longitudinally. A student's proficiency level can be represented in different ways, such as a language course or a passed test at a certain level assessed by trained teachers or assessors. For the SweLL corpus, written texts at different levels of proficiency are collected from beginners to advanced learners, of which 600 texts, including approx. 100 essays from first language learners of Swedish as a reference (control) subcorpus, are to be manually annotated to form a "gold standard" corpus.

The texts have been collected through the regular educational system for learners of Swedish (from age 16 and on) and in cooperation with teachers and test constructors throughout Sweden. An important criteria is a representative sample of regular courses and official tests of Swedish as a second language. They are collected both cross sectionally and longitudinally.

The other criteria of the corpus design relate to sociolinguistic metadata of the learner. The most important criteria is knowledge of other languages, in the questionnaire expressed as mother tongue(s), but also knowledge and use of other languages, as well as context of learning of Swedish (instructed versus self-study). Many people today define themselves as multilinguals and information about language learning and use can be important for research purposes. Two criteria have been set up to make the corpus design relevant and useful when it comes to language background: the corpus contains (1) around ten common immigrant languages in Sweden of today that the learners describe as their mother tongues and (2) some mother tongues that are compatible to other corpora, like the ASK-corpus (Tenfjord et al., 2006b). Other corpus criteria for the learners give a spread as to length and type of educational background, age (over 16) and gender.

Yet another variable of interest in L2 research is the task variable which, in our case, comprises task formulations, seven different text genres/types and a number of other characteristics that might have a direct or indirect influence over the language learners' production (e.g. Alexopoulou et al., 2017).

Ideally, the final corpus design should reflect the evenly distributed collected texts in terms of learner level (A1, A2, B1, B2, C1), mother tongues (Arabic, Dari, etc), and gender distribution (male vs female), as shown in Table 1. However, this design might prove to be challenging due to a range of socio-cultural characteristics and recruitment pitfalls.

In psychometric test norming (e.g. Lenhard et al., 2018), representative sampling would cover several variables, among others age, educational level, gender, regional distribution. The brackets for age groups, as described by e.g. Lenhard et al. (2018), are rather small (4-6 months) which sets high requirement on collecting sufficient amount of representative data per age bracket. In our case, we strive to spread the sampling across language proficiency levels, mother tongues, educational background and/or gender. And even though we also collect information about age spans and task types, these categories are not primary in the sampling.

With respect to the three-dimensional sampling, several potential problems arise:

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

L1	**A1**		**A2**		**B1**		**B2**		**C1**		**Control**[4]		**Total**
gender[1] –>	M	F	M	F	M	F	M	F	M	F	M	F	
Arabic	5	5	5	5	5	5	5	5	5	5			50
BCS[2]	5	5	5	5	5	5	5	5	5	5			50
Dari	5	5	5	5	5	5	5	5	5	5			50
English	5	5	5	5	5	5	5	5	5	5			50
Farsi	5	5	5	5	5	5	5	5	5	5			50
Greek	5	5	5	5	5	5	5	5	5	5			50
Kurdish[3]	5	5	5	5	5	5	5	5	5	5			50
Somali	5	5	5	5	5	5	5	5	5	5			50
Spanish	5	5	5	5	5	5	5	5	5	5			50
Tigrinya	5	5	5	5	5	5	5	5	5	5			50
L1 Swedish											50	50	100
Total	50	50	50	50	50	50	50	50	50	50	50	50	600

Table 1: Possible design of the SweLL corpus. [1]An alternative option that we are considering is to balance the number of low versus high educated L2 learners of Swedish (instead of gender); [2]BCS = Bosnian, Croatian, Serbian; [3]North & Central Kurdish (Kurmanji & Sorani); [4]Control group consists of L1 learners of Swedish at upper-secondary level; essays are coming from the national tests.

(1) The levels of proficiency are not unified between different educational establishments in Sweden, and several scales exist. One of them, CEFR (Council of Europe, 2001), the Common European Framework of Reference for Language Learning, defines six levels of proficiency (A1-C2). The scale is widely accepted throughout Europe, though still not so much spread in Sweden (Oscarson, 2015; Hildén, 2008; Erickson and Lodeiro, 2012). Instead, the stepwise development in language proficiency (for grown-ups) is defined in terms of course levels given at SFI establishments (A-D levels) and Swedish as a Second Language courses in adult education (SVA dk1-4 and SVA 1-3). For younger learners going through secondary and upper-secondary (gymnasial) education, foreign language courses are defined through a 7-step development scale (1-7). Since we are interested in cross-linguistic comparisons with L2 learner corpora in other languages we are considering to have all essays re-assessed according to CEFR and - optimally - also other systems of language proficiency scales in Sweden.

(2) The target L1 groups have been selected based on the most represented immigrant languages. The decision which mother tongues (L1s) to focus on in the corpus has been approached with care. The initial list of languages for sampling was based on (a) the statistics from the Swedish National Agency for Education over the mother tongues for paricipants in adult education programs from 2012-2016 (Skolverket, 2018) and (b) an overview of languages spoken in Sweden by Parkvall (2009). The selection of languages was then discussed within the research team and even with experts on certain languages

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

outside of the team, in order to ensure that we were including languages most frequently spoken as L1s in Sweden and, also, to avoid approaching learners in a way that could offend them, considering language and identity, i.e. being seen as belonging to a certain language community (or not). The socio-cultural characteristics of some of the ethnical groups, however, could pose challenges and overshadow recruitment. As an example, Dari learners are not frequently represented at more advanced levels of language courses, and the number of Kurdish females is not balanced with the number of male representatives. Overall, recruiting might turn out to be more difficult, than thought before. As a result, the following ten languages have been identified as the most important ones, belonging to 2 language families (Afro-Asiatic and Indo-European) and 7 branches/sub-groups:

- Afro-Asiatic: *Cushitic* (Somali), *Semitic* (Arabic, Tigrynia);
- Indo-European: *Germanic* (English), *Hellenic* (Greek), *Indo-Iranian* (Dari, Farsi, Kurdish: Kurmanji and Sorani), *Romance* (Spanish), *Slavic* Bosnian-Croatian-Serbian (BCS).

That said, we have adopted an opportunistic approach to data collection: initially, we collect everything we can. The harsh selection would be imposed on the 600 essays that we manually annotate for inclusion in the "gold standard" corpus. The availability of data and distribution of variables at that stage define the final corpus design for manual annotation.

A way around the recruitment and sampling problem would be not to limit each sampling group to 5+5 essays per each of the 10 languages as shown in Table 1, but to collect more for each individual characteristic, so that they could be subgrouped into various types of balanced samples per variable, e.g. just females of a certain level, just Arabic learners (not taking gender distribution into account) per proficiency level, or all low-educated versus all high-educated learners represented at the same level, etc.

3.3 Metadata

Two types of metadata are collected, all related to factors that are often seen as being significant for second language learning (e.g. Mitchell et al., 2012). These are 1. personal information about the learner, 2. information about the essay and the writing context, and the learner's performance on the essay if applicable.

3.3.1 Learner metadata

Given the importance of certain socio-demographic variables and personal characteristics for answering relevant research questions on the one hand (Sections 1 and 3), and the limitations imposed by the legal and ethical considerations (Section 3.1) on the other, we define a set of metadata about the learner.

As already mentioned, the project has a restrictive approach to collect personal metadata compared to other learner corpora projects. This specifically relates to data concerning personal information about the student's country of origin and age, see 3.1. The metadata shows *year of birth in a five-year interval* instead of exact year (e.g. 2000–2004, 1995–1999, etc.). No data is collected about country of origin, instead there are questions related to the *language background* of the learner, e.g. mother tongue(s), as well as

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

the learner's knowledge of other languages. Also included is *courses taken in Swedish or Swedish as a second language* and in mother tongue in the Swedish school, with allotted total study time for each language. Data regarding language use is identified i.e. *languages that the learner speaks or uses in particular communicative situations (spoken or written)* with family, friends, at work or in school, or elsewhere. Other personal metadata collected includes *gender* (female, male, decline-to-respond/other) and *education level* with number of study years at educational establishments in Sweden and outside Sweden (e.g. elementary school, upper secondary school, technical/vocational school, university degree, etc).

The learner is asked to fill in personal information about him/herself in a metadata sheet[9] in class to allow teachers' guidance. In order to anonymize the learner, each learner is provided with a SweLL-id that is used instead of the learner´s name on all essays collected for the project.

Based on an overview of current (L2) learner corpora metadata, Granger and Paquot (2017) have been working towards developing a standardization of core metadata for L2 corpora. They are proposing the scope of five main components: administrative, corpus design, annotation, text and learner metadata. The different variables of data within the five components are proposed to be either obligatory or optional. Comparing the kind of learner metadata collected in the SweLL-project to the obligatory variables proposed by Granger and Paquot (2017) shows a very high level of agreement. The opposite can be said of the proposed optional variables: almost none are collected in the SweLL-project. Some of the optional variables – e.g. profession/occupation, interaction with native speakers, scores on the modern language aptitude tests (MLAT) – have been part of early discussions in the project but have later been excluded due to factors like a wish to make it relatively easy and not too time consuming for either teachers or learners to fill in the metadata sheet.

3.3.2 Task metadata

Task structure is known to have a significant effect on the accuracy of L2 writing (or speaking) performance (e.g. Skehan and Foster, 1999; Kuiken and Vedder, 2007), as well as the type of language it elicits from the learner (e.g. Alexopoulou et al., 2017). Thus, it is of importance for a corpus relevant to research on learner language to contain task variables that pertain to the language situation (Granger, 1998). Therefore, in addition to the metadata about the learner, metadata about the task is collected. This includes information about the essay and the writing context as well as information about the learner's grade on the essay (if applicable). The information is based on instructions for the tasks provided by the teachers. The information includes (i) date when the writing task was carried out, (ii) school subject and year of schooling/study level, (iii) if the task has been graded, and if so - which grading scale was used, (iv) type of assignment, (v) topic/theme of assignment, (vi) maximum time permitted for the writing task, (vii) genre/type of text, (viii) instructions given to the learner (if applicable), (viiii)

[9]Personal metadata form in Swedish: https://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/01_SweLL-Personlig_metadata-Swedish.pdf;
in English: https://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/ENGELSKA_SweLL%20-%20Personlig%20metadata_EV.pdf

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

the learner's access to aid/tools to carry out the assignment, (x) the learner's access to additional material (dictionary, textbook, internet, peer response etc.). The teacher is also asked to include any additional material used as hand-outs. Lastly, if the essay has been graded by the teacher, a metadata sheet about the grade of each individual essay is filled in by the teacher. The teachers are asked to fill in the information on the metadata sheets[10]. All information provided is then stored in the database along with the essays and metadata information about the learner.

Compared to the set of variables on task metadata by Granger and Paquot (2017) SweLL task metadata, again, shows a high level of agreement regarding variables proposed as being optional as well as obligatory. There are a few differences and one is concerning what version of the given text it is (proposed obligatory variable), which refers to the versions of submitted and corrected-and-resubmitted tasks, and the time used for the assignment (proposed obligatory variable if timed). The metadata collected in SweLL does not include what the version of the given text is, which is presumably logical since we are targeting mostly exam settings. Nor does it include the time the student has used for writing the text but only the time allowed for the assignment. There is also a difference regarding the content of the writing task. While the writing task as proposed by Granger and Paquot (2017) is defined quite broadly, the collected information about the writing task in the SweLL project is based on the text types as proposed in the steering documents from the Swedish National Board of Education.

4 Data collection

In order to acquire data for the project, teachers, learners and researchers are involved. Schools are contacted, and teachers who teach Swedish as a second language from beginner to advanced levels as well as teachers carrying out assessment tests of Swedish are approached by the researchers. Teachers who are interested in participating are then provided information about the goals of the project, the method of data collection and a description of what their involvement would include. When a teacher has agreed to participate, a researcher registers the school and provides a school-id, makes sure that the schools have a safe storage for the texts and metadata sheets, and sets a date for an initial visit to the school to initiate the data collection. In some cases, the data collection has been initiated by the researcher explaining the project to the teacher who in turn explains it to her/his students, and in other cases by the researcher together with the teacher explaining the project to the group of students. The students who are interested in participating then receive information about the project, sign a consent form to donate their texts, and fill in the personal metadata sheet. Here, a note of caution needs to be added: even though students are provided, where possible, with consent and metadata forms in their native languages (L1s), they are asked to use the Swedish version for filling in the information. Translations are, thus, used only as support for their better understanding. The texts written by the participating students are from then on collected by the teachers and kept in the safe storage until they are picked up by the researcher or the project assistant.

[10]Task metadata form is available in Swedish only: https://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/SweLL-task-form.pdf

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

	# documents	# tokens	avg. # tokens
Collected	546		
Digitally born	140	N/A*	N/A
Handwritten	105	N/A	N/A
Transcribed	301	N/A	N/A
Annotated			
Pseudonymised	80	19703	246
Normalized	2	341	171
Correction-annotated	0	0	0

Table 2: Corpus size as of September 2019.
* Until essays are pseudonymised, they are stored in a protected Kiosk environment in possession of SweLL assistants and are unavailable on our servers for statistic inspection.

A pilot study involving the researchers as well as the teachers interested in participating in the project was conducted in February 2018. The aim of the pilot was to validate the flow of essay collection and to evaluate the portal (described in Section 8.1) for registration of metadata forms and the uploading of essays. A check list for the different steps was provided to the teachers. The outcome of the pilot concerned issues like the time frame for collection, the clarity of how to fill in the metadata sheets on both the students' and the teachers' parts. The teachers' comments helped develop the metadata sheets and the school-based part of the essay collection. Some of the teachers opted out from the continued collaboration due to the lack of time as a result of the insights gained during the pilot.

Once the metadata forms are collected, project assistants or researchers enter all information into the portal – a project management system which links to a database, keeps track of all the data, helps assign tasks to project assistants, and provides statistics over the corpus collection and preparation status (see Section 8.1).

As of September 2019, 546 L2 texts have been collected. The largest portion of the collected texts come from SFI schools. Many texts also stem from university courses in Swedish as a Second Language, while only a rather small portion comes from upper-secondary (Swe. *gymnasium*) courses in L2 Swedish. A part of the learners are recurrent ones, which allows for collecting some amount of longitudinal data. Around 150 essays, not counted towards the above-mentioned L2 essays, come from the Uppsala Corpus of Student Writings (Megyesi et al., 2016) and have been added to the SweLL collection to function as a Control, or Reference, corpus (see Table 1) representing essays written as part of national tests by L1 Swedish upper-secondary learners.

Out of the 546 collected essays, 80 have been pseudonymised, and two have further been normalized (Table 2). No essay has yet been correction-annotated. The pseudonymised essays sum up to nearly 20,000 tokens, and the average essay has a length of 246 tokens. Assuming a similar distribution among remaining essays, the complete manually annotated corpus of 600 essays will contain around 150,000 tokens.

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

5 Transcription and pseudonymization of L2 data

The two steps - transcription and pseudonymization - require special care to protect data from accidental use by unauthorized users. Hand-written texts contain all personal information including hand-writing itself, which can give away a person behind the text. Original non-pseudonymized texts may contain information that needs to be masked before reseachers outside the project can be given access to them. That is why the two steps are performed in a special encrypted environment, kiosk (Section 8.2), before the pseudonymized essays are saved to a database.

5.1 Transcription

The transcribed version of a text should ideally be a reproduction of the string of conventional symbols (letters, numbers, punctuation marks, etc.) which make up the handwritten text. The major difficulty for transcribers is to resist making unintentional corrections, and the most important part of securing the quality of the transcriptions seems to be finding and training meticulous transcribers used to second language texts. Random checks of transcribed files are always performed by another transcriber or a researcher. In SweLL three transcribers have worked on the texts, all of them used to transcribing handwritten texts at different language levels.

Uncertainty about how to best represent the handwritten text arises primarily due to unclear handwriting or due to changes in the text made by the writer. The fundamental principle for resolving such uncertain cases is *the principle of positive assumption*: Whenever one of the alternatives involves better intelligibility or closer adherence to standard norms, that is the alternative which should be chosen. While transcriptions are primarily carried out individually, transcribers are encouraged to discuss uncertainties with another transcriber or a researcher. This *cross-consultation routine* secures the quality and consistency of the transcriptions. Thorough description of the process is provided in the SweLL transcription guidelines[11].

5.2 Pseudonymization of personal information

When the original texts of the learner are transferred to a computer-readable format, the texts are pseudonymized to ensure the anonymity of the learner. The purpose of pseudonymization is to de-identify all information that can reveal the identity of the person who wrote the text. This information might include, for example, person name, profession, age, dates, account or license numbers, locations like home town, address, work place, transportation, 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. We developed a taxonomy of personal information adapted to L2 learners' texts, described in Megyesi et al. (2018).

During the pseudonymization step, all personal information that can relate to the learner is manually identified and classified according to the pre-defined taxonomy (e.g personal name, city, country). As tokens representing personal information are classified by type, they can be automatically replaced with pseudonyms, randomly from a small set

[11] https://spraakbanken.github.io/swell-project/Transcription_guidelines

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

of fake terms organized by type. For instance, *Göteborg* might be replaced by *B-city*. As the pseudonymisation type (i.e. label) is stored along with each pseudonymized token, the corpus user can choose whether they want to use the provided pseudonyms or utilize the pseudonymisation type directly.

Numeric indices, ascending from 1, are automatically added to each classified token. Whenever a token occurs more than once, the number can be manually changed to match that of the first occurrence, whereupon the same pseudonym is automatically applied. In addition, the items are annotated with morphological information so that the pseudonym can have the same morphological form as the original (e.g. case and definiteness). The entire process is described in detail in Megyesi et al. (2018).

6 Normalization and correction annotation

6.1 Normalization

The purpose of normalization is to correct words or segments of the learner source text that deviate from the norm of the target language, ideally by making minimal changes of the form while retaining the content of the learner text. Since the process of reconstructing the learner's intended message requires interpretation, many changes are possible for the same deviation, and there is an inescapable element of subjectivity in deciding on a correction. Hence, the corrections have the status of hypotheses, referred to as *target hypotheses*.

For example, consider the learner sentence **jag trivs mycket bor med dem* (EN (English): "I enjoy live with them"), see Figure 5. Applying the main principle of normalization that *any change to a grammatically correct version should be as small 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* (EN: live, verb) →*bra* (EN: well, adverb). 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. SLA researchers involved in the SweLL project were unanimous about changing this sentence to *Jag trivs mycket med att bo med dem*.

The problem of normalization has been approached in different ways. In some projects, normalization is merged in one step together with error annotation, e.g. in ASK (Tenfjord et al., 2006b) where provided error corrections can be used to derive a normalized version of the text, the target hypothesis. For example, to extract a correction of *stillingen* (EN: position) in the example below, one needs to access the attribute *corr* (correction):

```
<sic type="F" desc="AGR" corr="stilling">stillingen</sic>
```
[12]

This might be called flattened normalization. Markup of word order errors in that case becomes more challenging since larger text segments have to be included into both the xml-element and the corr-attribute.

[12]The example is taken from a presentation by Paul Meurer at CLARIN workshop on Interoperability of L2 resources and tools, ASK project: https://sweclarin.se/sites/sweclarin.se/files/ASK-Goteborg.pdf

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

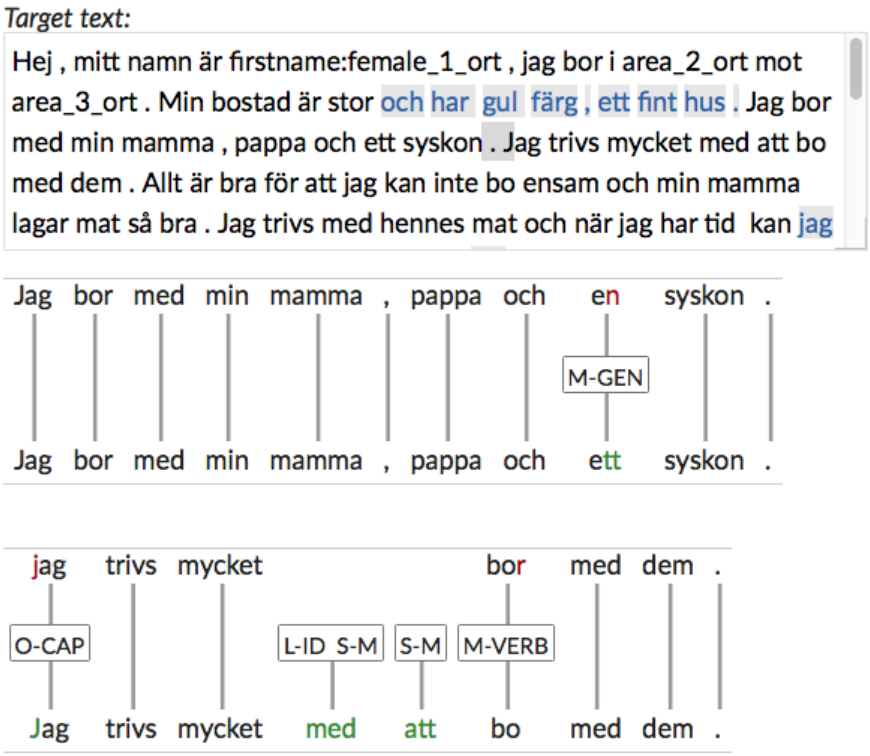

Figure 5: Original and normalized versions of a learner text, with correction tags added on the edges. Gloss of the original layer (with some imitation of the errors): *I live with my mother, father and an sibling . i enjoy live with them.* The question is, should the second occasion of *live* be changed to *living* or *(my) life*? Note that in the normalization mode it is still possible to edit pseudonymization in case some need may arize.

In some recent projects, normalization has been performed in a parallel fashion, where alongside the original version a full-fledged target hypothesis is generated (Boyd et al., 2014; Rosen et al., 2014). Figure 5 shows how this is visualized in the SweLL normalization/correction annotation and pseudonymization tool SVALA (Rosén et al., 2018; Wirén et al., 2019). The row on the top shows the original text, the row at the bottom represents the target hypothesis, with error labels on the edges linking original and target segments. Parallel view handles word order errors in a more intuitive, and possibly also more elegant, way.

Several experiments with normalization and correction annotation in the SweLL project have shown that normalization as a separate step is preferable for several reasons:

- It helps the annotator to build a better understanding of a learner's linguistic competence (for example, 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

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

much less time compared to correction annotation/labeling 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.

- Correction 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 correction labels at the same time as normalizing a text segment).

- Inter-annotator agreement with respect to correction codes can be objectively measured only given that the annotators are working on the same normalized version.

Following the two pilots (3 & 4, as mentioned in Section 2.1) a decision was made to separate normalization from correction annotation, and to outsource this step to SLA-trained staff. Only one normalization version is created at this stage of SweLL infrastructure development. However, we are aware of the potential need to produce several target hypotheses reflecting several interpretations of the same learner-produced text passage, and we will assess which way that could be incorporated into the SweLL infrastructure in the future.

6.2 Correction annotation

Correction annotation means labelling the changes that have been made in the normalization step (Section 6.1), using a hierarchical taxonomy of correction types (discussed in Section 6.3). This process is commonly referred to as "error annotation" in the literature on learner corpora research (e.g. Granger, 2013; Lüdeling et al., 2005), but there are subtle but crucial differences between the two terms which make us prefer the former one. First, "correction annotation" is meant to signal that what we label is not an alleged objective behaviour of the learner but rather inherently subjective behaviour on our own part, namely, postulating a hypothesis about how the learner text could be changed. The fact that it is a hypothesis means that another annotator or researcher might prefer a different hypothesis on equally probable grounds. 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 5, a wrong form of a verb (*bor* →*bo*) plus idiomaticity problem in using the verb *att trivas* (*trivs* →*trivs med att*). In other words, we cannot claim that we are error-labelling the learner language; rather, we are labelling the corrections that we have introduced.

But there may also be real differences between correction annotation and error annotation. Thus, when a correction involves a replacement, we do not annotate errors that appear in the material that has been replaced, but only the replacement as such. For example, if "He always make all the shopping" is changed to "He always does all the shopping", we annotate the lexical replacement ("make" replaced by "do") but not the agreement error of "make" (> "makes") since this token is not present in the corrected text. It is, however, important to point out here, that the original learner text will still be accessible for the corpus researcher, however, it will not appear in a search for third-person-singular verb-subject agreement.

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

In SVALA, correction annotation is carried out by selecting one or several (not necessarily contiguous) tokens in the source and target texts, or the edges that connect them. A pop-up menu with correction labels is displayed to the left. The annotator selects one or more labels, which are displayed on the corresponding edges (see Figure 6). The rationale for labelling the edges is that we view correction labels as relations between the learner and corrected versions of the text.

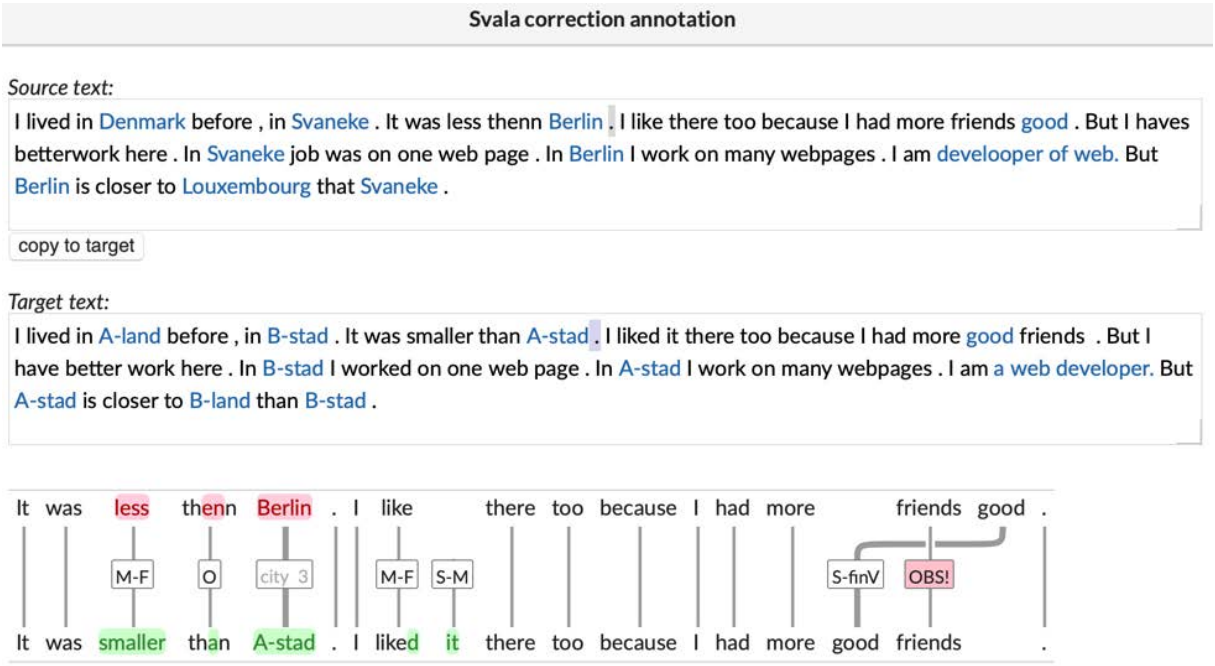

Figure 6: SVALA correction annotation.

As indicated above, we have good reasons not to use the term error annotation, since it simply does not accurately describe our annotation approach. The fact that the question of terminology regarding error annotation was raised in the first place, however, has to do with the cross-diciplinary character of the project. While the term error annotation is a widely adopted one both within Natural Language Processing (NLP) and Learner Corpus Research (LCR), the notion of error, as well as the use of the term, has been controversial within many other branches of modern day SLA, since the field turned its back on error analysis (EA) (e.g. Corder, 1967) in the 1980s (cf. Tenfjord et al., 2006a; Prentice and Granger, In prep.). This issue was raised by the SLA researchers in the SweLL project and several other terms have been discussed, including *norm deviations, interlanguage phenomenon* (Dıaz-Negrillo et al., 2010), *non-norm adequate form* (Dobric, 2015), and *unexpected uses* (Gaillat et al., 2014). However, neither of the terms entirely served the purpose since the problem is not solved by renaming alone. The same problem was discussed in relation to comparable projects for other languages, such as the creation of a developmental corpus for Slovene (Holdt et al., 2017), where the lessons learned included moving the perspective from talking about learner errors to teacher corrections.

The ideal taxonomy for learner language annotation, should, from our (and other SLA and LCR researchers', e.g. Thewissen (2013)) point of view, create a balanced picture of

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

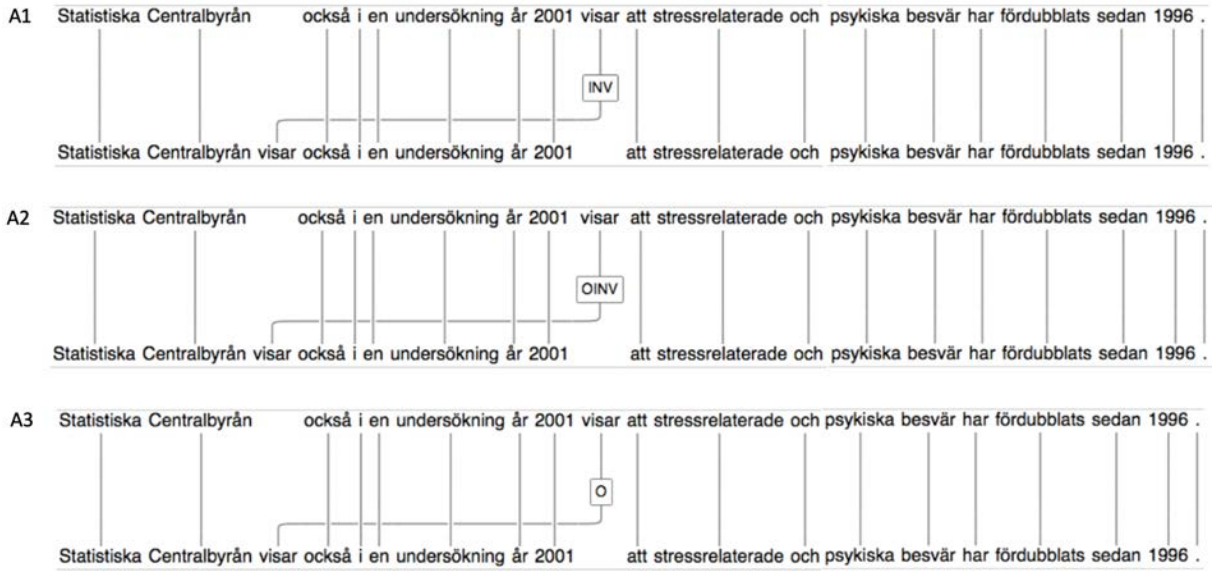

Figure 7: Three versions of correction annotation of the same segment, by three annotators (A1, A2, A3). Gloss of the original layer: *Central Statistical Agency [...] also in a report from 2001 [shows (finite verb)] that stress-related and psychological troubles have doubled since 1996.* The three correction tags for word order *INV*, *OINV* and *O* have been used to describe the same change introduced into the original text.

learners' can- and can't do-s in relation to a relevant linguistic norm in a given context (Prentice and Granger, In prep.). The limitations in time and resources for this project do, however, only allow for the implication of correction annotation. As discussed in the previous section, the term correction annotation has been adopted mainly for other reasons and is hence not exclusively related to the different views on the notion of error within the disciplines involved in the SweLL-project.

6.3 Correction taxonomy

The design of a taxonomy for the annotation is a complex matter which requires many detailed decisions on different levels. These decisions have to be pragmatic to a certain extent, since the conductibility of the annotation work depends on it. It is, however, hard to imagine, that the researchers designing the taxonomy would not to some degree apply their theoretical views on language to it, even if some corpus projects have an explicitly non-theoretical approach to taxonomy design (cf. Tenfjord et al., 2006a). So far few learner corpus projects have managed to reuse each other's error taxonomies, even though attemps to build on previous work have been made. At an early stage, the SweLL-project tried 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 existing error

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

	STEP 1		Taxonomy 1.0	STEP 2	Taxonomy 2.0
Taxonomy (L2 corpus project)	**MERLIN** (64 codes), applied in pre-pilot	**ASK** (23 codes), applied in pre-pilot	**SweLL 1.0** (ASK+, 29 codes), applied in pilot 1		**SweLL 2.0** (29 Codes),
ERROR CATEGORY and codes	GRAMMAR, N=21	MORPHOLOGY, N=3 SYNTAX, N=7	MORPHOLOGY, N=7 (ASK+4) SYNTAX, N=8 (ASK+1)	PILOTS 1 AND 3	Morphological codes, N=8 Syntactical codes, N=8
	ORTHOGRAPHY, N=8	PUNCTUATION, N=4	PUNCTUATION, N=2 (ASK-2)		Punctioation codes, N=4
	INTELLIGIBILITY, N= 8		INTELLIGIBILITY, N= 1		Intelligibility, N=1
	VOCABULARY; N=10	LEXICON, N=7 ORTHOGRAFY, N=1	LEXICON, N=10 (ASK+2)		Lexical codes, N=4 Orthografic codes, N=3
	COHERENCE, N=4				
	SOCIOLINGGUISTICS, N= 10				
	PRAGMATICS, N= 3				
		UNIDENTIFIED, N=1	UNIDENTIFIED, N=1		Unidentified, N=1
					Concistency, N=1

Figure 8: From Merlin and ASK to SweLL taxonomy.

annotation taxonomies, namely of ASK (Tenfjord et al., 2006b) and MERLIN (Boyd et al., 2014). While comparability is, of course, an important issue, a one-size-fits-all taxonomy for annotating learner language is probably neither realistic nor necessarily desirable, considering the various aims and circumstances for different projects. That said, it becomes even more important for different projects to be as transparent as possible when it comes to the principles and decisions that their taxonomies are built on.

The SweLL tagset has been developed in several steps, with the ASK (Tenfjord et al., 2006b) taxonomy (23 tags) and the MERLIN (Boyd et al., 2014) taxonomy (64 tags) as a starting point. In a first pre-pilot annotation these two taxonomies were applied on two Swedish learner essays, randomly selected from a beginner and a rather advanced levels. It turned out that annotating with the highly intricate MERLIN taxonomy took

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

twice as much time as with the ASK taxonomy, also resulting in a lot of inter-annotator disagreements. As a result of this pilot experiment, the ASK taxonomy was adopted with several modifications and was tested in a second pilot study within the SweLL-research group.

Once again, the practical implementation of the taxonomy raised important questions with reference to tag names and their coverage. See for example Figure 7, where three annotators - A1, A2 and A3 (in this case - SweLL researchers with linguistic training, one native speaker of Swedish, two others - near-native users of Swedish) - agreed on both the segment in need of correction (top row) and on the target hypothesis (second row), but not on the correction label (O, INV, OINV describing various types of word order errors[13]). As a result of the second pilot, both the tag names and the number of tags were reviewed and changed again in order to minimize ambiguity. As a consequence, the taxonomy has become increasingly independent from the ASK-error taxonomy, even if it still bears a clear resemblance to it when it comes to the general linguistic categories annotated, cf. Tenfjord et al. (2006b); see also Volodina et al. (2018). The development process for the SweLL taxonomy (2.0) is described in Figure 8.

Testing the taxonomy in several annotation pilots has been a time consuming and, at times, frustrating process, since the need to discuss the details and inconsistencies in the taxonomy and the revision of the taxonomy were seemingly endless. In hindsight, pilots and the following (sometimes circular) discussions were absolutely necessary to achieve gradual improvement process that is driven by the experience of practical annotation in relation to detailed discussions about the pragmatic, methodological and theoretical issues involved in the annotation of learner language. One example of a reoccurring issue was the initial distinction in the SweLL taxonomy between grammatical words and content words. While one could argue that there is a difference between e.g. the omission of a grammatical word, like a preposition, and omission of a content word, both when it comes to the form and syntactic form of a structure, there are too many cases where the annotator would have to make an ad-hoc decision on whether a certain word is a content word or not, which would affect inter-annotator agreement (and reliability of the annotation) in a negative way. Since the automatic POS-tagging provides information about the types of words that are annotated in correction annotation, after testing, re-evaluating and discussing this issue (like several others before) the decision to abandon the distinction between grammatical and content words was made, and thereby the ambiguity within the taxonomy decreased.

The SweLL taxonomy allows for a somewhat more fine-grained annotation with regards to certain categories, compared to the ASK taxonomy. These cases have to do with language specificities for Swedish and Swedish learner language (i.e. common error types), e.g. the use of definiteness in complex noun phrases (covered by the correction code M-DEF) or subject omission (S-Msubj), see Figure 9. The final SweLL-taxonomy consists of 29 correction codes that have been tested and reviewed again at an annotation meeting in January 2019. The final adjustments to the taxonomy have been made due to internal consistency between the correction categories and within and between the linguistic levels annotated. The different steps in the development of the Swell correction taxonomy are illustrated in Figure 9.

[13]ASK 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

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

LEXICON (4)	MORPH (8)	ORTH (3)	PUNC (4)	SYNT (8)	Other (3)
L-W wrong word	M-ADJ/ADV	O spelling error	P-W wrong punct.	S-adv adverb. placement	C Consistency (follow-up corrections)
L-DER diviant (existing) affix		O-CAP capitalization	P-R redundant punct.	S-COMP Wrong choice mwu vs. compound	X intelligibility
L-FL foreign word	M-CASE	O-COMP Oversplitting/compounding	P-M missing punct.	S-finV finite verb placement	Uni Unidentified error
L-REF wrong reference	M-DEF deviation definiteness		SENT- segmentation	S-Msubj subject missing	
	M-F correct grammatical category, wrong paradigm			S-M word missing	
	M-GEND			S-R redundant word/phrase	
	M-NUM			S-WO word orde (other)	
	M-other			S-CON problematic syntactical construction	
	M-VERB				

Figure 9: SweLL correction taxonomy, version from April 2019.

To support normalization and correction annotation in a parallel fashion, and to guarantee transparency and reliability during annotation, a tool SVALA has been developed (Rosén et al., 2018; Wirén et al., 2019), see Section 8.3 for details.

6.4 Inter-annotator agreement

To ensure that the proposed set of correction tags is clearly defined and straightforward to apply, two pilot annotation experiments (1 and 3, see Section 2.1) have been carried out for testing the reliability of annotation with the taxonomy at hand. In both cases a number of learner texts have been selected for annotation by two or more annotators. We then measured inter-annotator agreement (IAA) on these texts, i.e. the extent to which annotators agreed on the correction tags assigned to the words or segments they

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

considered incorrect.

Since correct words did not need to be annotated with a correction tag, we opted for Krippendorff's α (Artstein and Poesio, 2008) as an IAA measure, which is computed based on observed disagreements. Moreover, as annotators could provide one or more tags for an error, we compared the assigned tags based on Jaccard distance (JD, Jaccard, 1908), which allowed us to capture partial agreements on a set of assigned tags. JD is computed by taking the union of the set of tags assigned by annotators, subtracting from it the intersection of these tags and dividing the resulting value by their union again. For example, if one annotator assigned tags X and Y to an error, while the other annotator indicated Y and Z, JD would be (3-1)/3 = 0.66. We employed the NLTK Python module for calculating both the distance and the agreement measure (Bird and Loper, 2004). Besides errors consisting of single tokens, there were cases in which multiple tokens constituted one error. When comparing annotations for these, we computed: i) the agreement on the exact span of tokens marked by annotators (EXACT); ii) and the agreement rate with correction tags projected down to individual token level (SINGLET).

During the first pilot experiment, we measured IAA on two texts annotated by three annotators each. Then, after a slight revision of the correction taxonomy, a second, larger-scale annotation experiment was carried out. This consisted of the annotation of 24 texts by two or three annotators each. The IAA results computed as described in the previous paragraph are presented in Table 3. The numbers in parenthesis indicate results for broader (macro) correction categories, that is, the first five column headers from Figure 9 (LEXICON, MORPH, ORTH, PUNC, SYNT) as well as the C, X, and UNI tags.

Pilot Exp. 1		**Pilot Exp. 2**	
EXACT	SINGLET	EXACT	SINGLET
0.677	0.602	0.627 (0.689)	0.591 (0.656)

Table 3: IAA results in terms of Krippendorff's α.

To put these results into perspective, $\alpha = 0.667$ is the minimum threshold proposed for drawing conclusions and $\alpha = 0.80$ indicates a high annotation quality (Krippendorff, 2004; Artstein and Poesio, 2008). It is worth noting that the inherent complexity of this task, reflected also by the number of correction tags, undoubtedly influences the achievable rate of agreement.

To gain more insights into the annotation process and the nature of disagreements, in the second experiment, we calculated some statistics for tag use and inspected tag pairs commonly disagreed on. Figure 10 shows the summed pairwise correction tag assignments by all annotators over all texts.

The numbers in the confusion matrix in Figure 10 indicate the number of times all annotator pairs assigned two correction tags – axis X and axis Y respectively – to the same error over all texts. The last element of each row represents matching correction tags, i.e. the number of times annotator pairs agreed on the type of correction, while

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

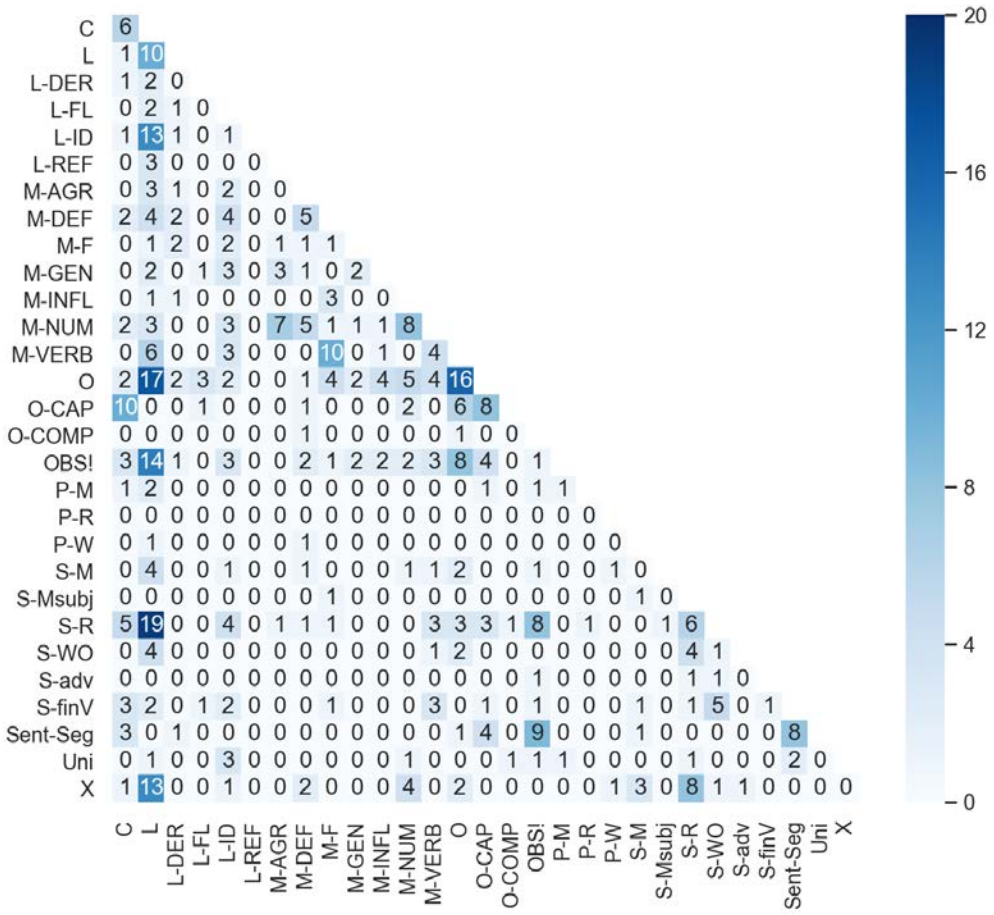

Figure 10: Confusion matrix over correction tags.

all other cases illustrate disagreement. We found that the labels L[14], O, S-R, M-VERB and M-DEF were the five most frequently used tags. Furthermore, the analyses revealed that the most common tag disagreements occurred between the incorrect use of existing Swedish words (L) and the following three types of correction tags: redundant word or phrase (S-R), orthographic / spelling errors (O) and the non-idiomatic use of expressions (L-ID). Other common disagreements included consistency "follow-up" errors stemming from other corrections (C) vs. capitalization errors (O-CAP); gender-related errors (M-GEN) vs. morphological errors related to verbs (M-VERB) and definiteness errors (M-DEF) vs. incorrect use of number agreement (M-NUM).

All in all, early analysis of IAA has influenced revisions of the code taxonomy and guidelines, and has been an important step towards improvements in the manual anno-

[14]This correction tag has been replaced by L-W (with a slightly different coverage) after pilot 3, as shown in Figure 9

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

tation process[15].

7 Linguistic annotation

While the linguistic annotation of the SweLL texts has not yet been performed, our intention is to process normalized, target versions of the learner texts linguistically by using state-of-the-art natural language processing tools developed for Swedish. There are several options of those available off-the-shelf, and we will explore the possibility to set up a "model" (recommended) pipeline, but at the same time allowing the infrastructure user choose a custom-assembled toolkit.

To provide linguistic annotation of the learner texts, we consider two major alternatives: adapting the SweGram annotation tool (Näsman et al., 2017; Megyesi et al., 2019) based on efselab (Östling, 2016) - to the SweLL infrastructure, as well as offer Sparv pipeline (Borin et al., 2016). The automatic linguistic annotation in both pipelines consists of several steps: preprocessing, i.e. the text is tokenized and the sentences are segmented first, before normalization and pseudonymization are applied. Each token in the normalized, pre-processed text is lemmatized and annotated with part-of-speech and morphological information using the SUC tagset (Gustafson-Capková and Hartmann, 2006), as well as the universal part-of-speech tagset[16] (Nivre et al., 2016) (currently available only the SweGram version). On the basis of the lemmatized PoS-tagged tokens, each sentence is parsed using dependency parsing (Nivre et al., 2016).

When adding new learner texts (i.e. the ones that are not manually corrected) to the infrastructure, the same annotation pipeline can be used allowing a uniform annotation to the entire dataset applying the same principles. Since the new learner texts are expected to contain errors, SweGram provides automatic spelling normalization to the texts, using a modified version of HistNorm (Pettersson et al., 2013) which can be adapted to learners' writings.

8 SweLL tools

In the current project we have opted for development of our own tools rather than reusing the ones from the previous projects. The reasons for that were several.

First, the previous projects that built on existing tools, reported problems with lossy conversions and non-matching formats (e.g. Boyd, 2017). Second, our intention was to build an intuitive tool that SLA and LCR researchers would find easy to work with, which we didn't find off-the-shelf. Third, we aimed at building a parallel corpus where the original text and the target ones would be linked. None of the existing tools could offer all of the above. Finally, we needed a tool for managing the project, the annotation process, and statistics, as well as for supporting encrypted environment for initial "sensitive" steps of transcription and pseudonymization. The setup of a project management environment together with an intuitive tool in one was seen as a prerequisite for reliable manual annotation.

[15]Note that the IAA numbers come from the pilots carried out BEFORE we separated normalisation from correction annotation.

[16]http://universaldependencies.org/

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

8.1 Project management

For project management, a SweLL portal environment was set up. It opens a user-friendly interface to a database where we can store, access and manage information about learners, tasks, and have access to the essays. It also allows to assign tasks to assistants for annotation, calculate inter-annotator agreement, have an access to metadata statistics, and have a general overview of the corpus work. The tool is in its beta-version and is under active development.

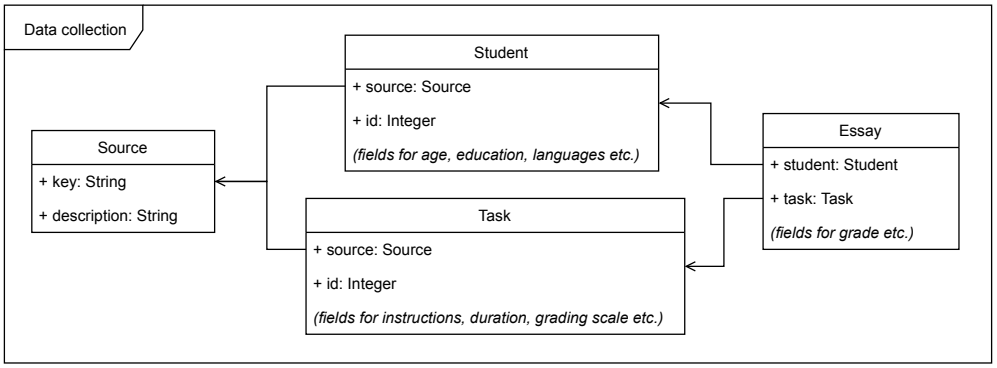

Figure 11: Overview of the database model for metadata collection.

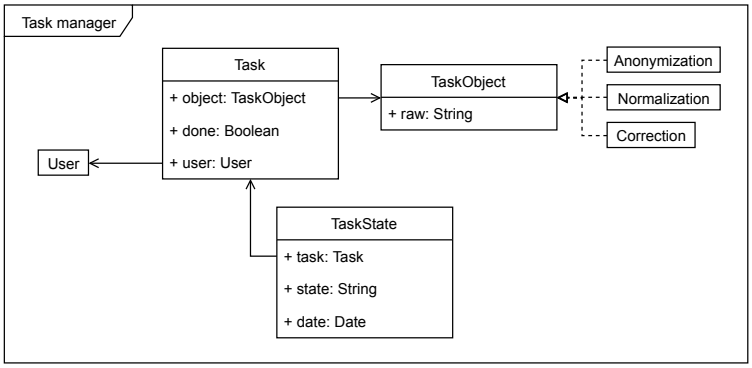

Figure 12: Overview of the database model for task management within the corpus construction workflow.

Figure 11 and Figure 12 provide a simplified overview of the database model underlying the SweLL portal.

The metadata repository contains records for *students*, *tasks* and *essays*, organized by source (for example, the Swedish course at a certain school). An essay is modeled as the response of one student to one task.

The anonymization, normalization or correction annotation of a given essay is modeled as a *task object* record. The actual student-written texts live here, and not in metadata *essay* records. The assignment of a *task object* to an assistant (*user*) results in a *task* record (which models a task for assistants, as opposed to a task for students).

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

8.2 Kiosk

To secure a safe environment during data collection and pseudonymization, a special solution has been developed in the project, called SweLL-kiosk which is an encrypted hard-drive designed specifically for the purpose of pseudonymization, an environment that protects unauthorized users to get access to the non-pseudonymized versions of the essays. Kiosks are equipped with a project management system (based on the SweLL portal, Section 8.1), a database for storing all versions of the files, and an installation of a minimized version of SVALA annotation tool (Section 8.3).

To facilitate the manual labor of pseudonymization, we developed the SVALA pseudonymization tool to assist the annotators. The tool is used for the purpose of *pseudonymization* before the essays are uploaded to the web-based portal. The tool SVALA attaches the original text to the pseudonymized text token by token, building a parallel version of the two. The annotator marks up the personal information token by token, and classifies it so that information is kept about the source text, the target text, and the manipulated segments along with the edges between the source and target segments (Figure 13, e.g. "B-stad"/"city" for "Berlin").

To guarantee anonymity, we carry out the identification and masking of personal information manually, sentence by sentence. Once the text is pseudonymized, the deidentified essay is exported to the online SweLL portal for further processing, to normalize and annotate the text accordingly.

The functionalities in the kiosk version of SVALA are practically the same as described in the section below, apart from the fact that it does not allow to perform any other annotation steps.

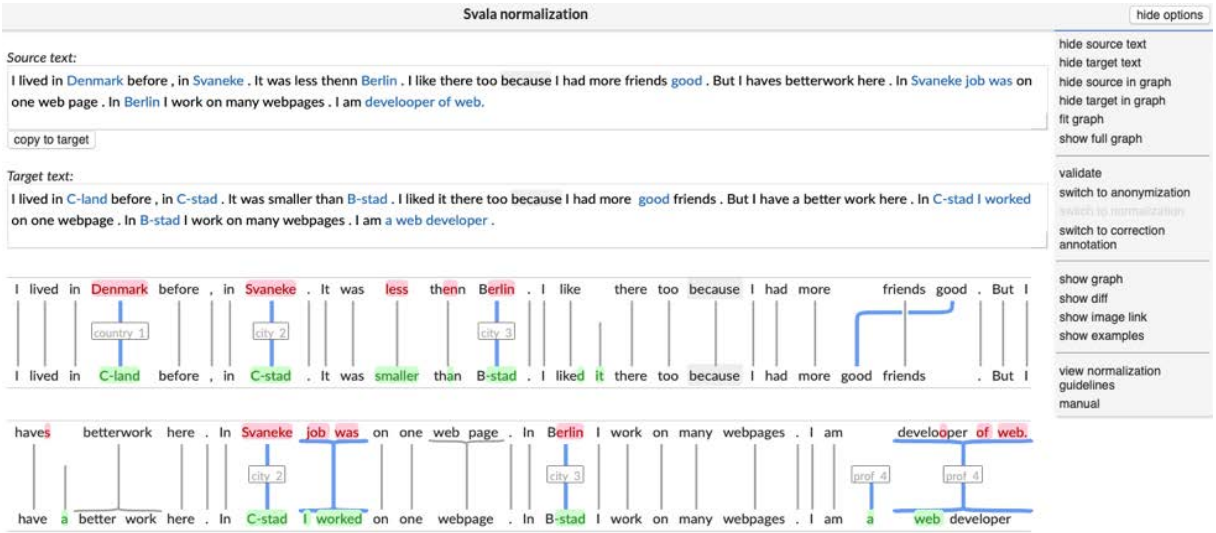

Figure 13: Example of a "parallel" annotation. Normalization mode with pseudonymization tags and rendered pseudonyms.

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

8.3 The SVALA annotation tool

In the past, projects often re-used tools from a number of different other projects for the steps outlined in Figure 2 (e.g. Boyd, 2017). As a result, various input and output formats between the tools needed to be converted, which increased the complexity of the task.

In the present project, we are developing a tool that handles all of the manual annotation steps (transcription, pseudonymization, normalization, correction annotation) in one environment maintaining a stable interpretable format between the steps. A distinguishing feature of the tool is that users work in a usual environment (e.g. plain text) while the tool visualizes all edits that are performed via a graph that links an original learner text with an edited one, token by token. This way word order changes are easily handled. Within the current project, SLA researchers find it convenient and intuitive to work with the tool. Based on that, we assume, that the tool can be of interest to LCR and SLA community outside the project as well.

Pseudonymization is a preprocessing step described earlier, which is performed in Kiosk environment. However, corrections to pseudonymization are also possible during the normalization step.

During *normalization*, a user is working directly in the field "target text" editing a copy of an original text (Figure 14). All the while, the tokens in the original texts are automatically aligned with the tokens in the target texts, a graph is incrementally updated and the result is visualized in a parallel view.

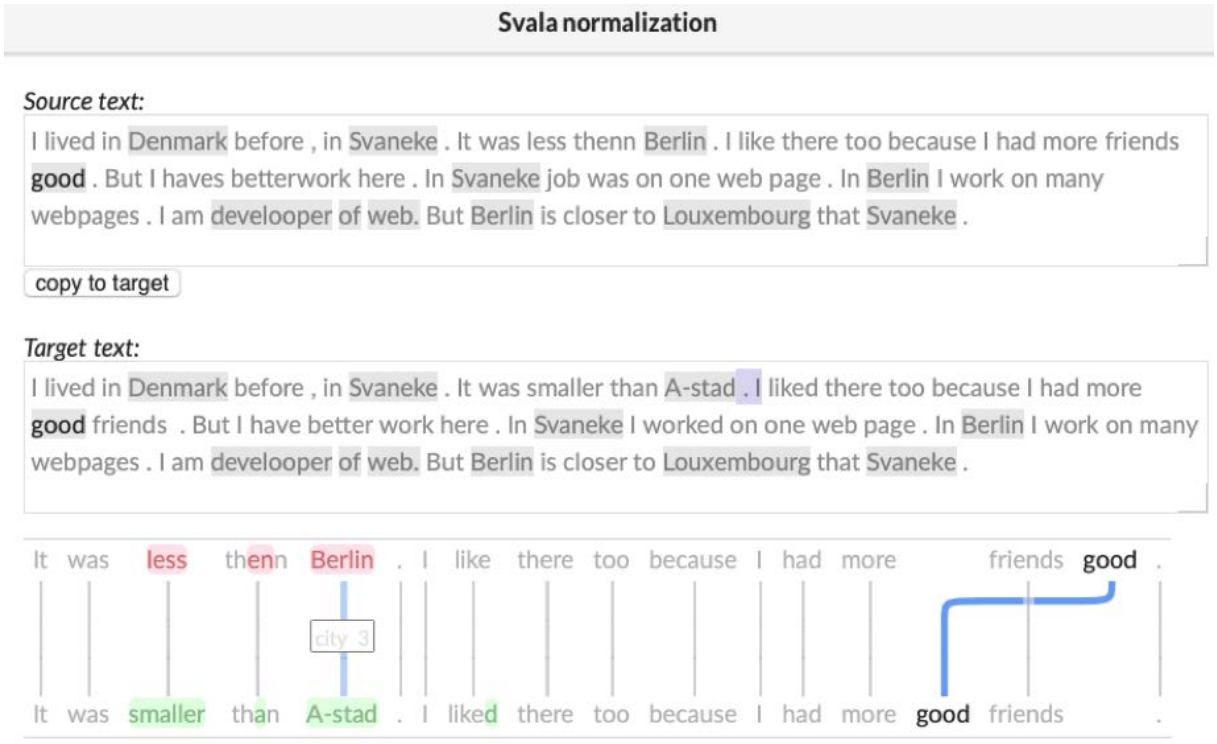

Figure 14: SVALA normalization mode.

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

In *correction annotation* mode, correction labels are assigned to the links in the graph between the tokens that display difference between the original and the edited text (e.g. "thenn" and "than", see Figure 6) and characterize the nature of the difference.

Alignments (i.e. links) in the graph are built automatically, and can be manually corrected. It may especially be necessary when it comes to the word order changes (see "friends good" vs "good friends" in Figure 14).

The choice has been made in the project to support JSON format representation of the data as an alternative to a more universally accepted XML, since JSON ensures a relative light-weightedness of the tool and supports structuring the data in an easily accessible way. Three data objects are created (see Figure 15):

(1) to handle the original text,
(2) to handle the target text and
(3) to describe edges (links) with attached correction labels

For further technical details, see Rosén et al. (2018); Wirén et al. (2019).

Conversion from SVALA format to XML TEI[17] (Text Encoding Initiative) format or any other format is a trivial step, and in case there is interest to SVALA outside the project, we can consider adding it to a format conversion tool, such as Pepper (Zipser and Romary, 2010).

Adapting SVALA format to existing search environments may present challenges. We foresee wasting some of the information encoded in the present format when flattening it to a less expressive XML format. Those challenges and potential other consequences will be evaluated and addressed in the not so distant future.

9 Insights and pitfalls

Among the burning questions in emerging learner corpora infrastructures (e.g. MacWhinney, 2017), the two questions below remain to be the most important at our current stage of development:

(a) how to collect *enough* L2 learner data that is both balanced, representative and accessible, as well as

(b) how to obtain *reliable, consistent* and *useful* annotations of data.

While (a) has not been addressed up to now in our infrastructure work, we plan to explore collection of essays through an online platform where pseudonymization of newly submitted essays will be done "on-the-fly" using manually pseudonymized essays as training data. Our hope is to achieve a more effective data collection process this way, and to prevent teachers from opting out because administrating consent forms and socio-demographic forms to their students consumes valuable teaching time and adds to their already heavy work load.

Adding manual annotation (b) is a further complication on the way. According to Hovy and Lavid (2010) (reliability of) annotation in corpora has two types of major consequences, namely *theoretical* ones for shaping, extension and re-definition of theories, and *practical* ones for use in classrooms, teaching and assessing practices. Unreliably

[17]https://tei-c.org

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667


```
{
  "source": [
    {"id": "s0", "text": "It "},
    {"id": "s1", "text": "was "},
    {"id": "s2", "text": "less "},
    {"id": "s3", "text": "thenn "},
    {"id": "s4", "text": "Berlin "},
    {"id": "s5", "text": ". "}
  ],
  "target": [
    {"id": "t16", "text": "It "},
    {"id": "t17", "text": "was "},
    {"id": "t28", "text": "smaller "},
    {"id": "t31", "text": "than "},
    {"id": "t36", "text": "B-stad "},
    {"id": "t21", "text": ". "}
  ],
  "edges": {
    "e-s4-t36": {
      "id": "e-s4-t36",
      "ids": ["s4", "t36"],
      "labels": ["1", "city"],
      "manual": true
    },
    "e-s0-t16": {
      "id": "e-s0-t16",
      "ids": ["s0", "t16"],
      "labels": [],
      "manual": false
    },
    "e-s1-t17": {
      "id": "e-s1-t17",
      "ids": ["s1", "t17"],
      "labels": [],
      "manual": false
    },
    "e-s2-t28": {
      "id": "e-s2-t28",
      "ids": ["s2", "t28"],
      "labels": ["L-W"],
      "manual": false
    },
    "e-s3-t31": {
      "id": "e-s3-t31",
      "ids": ["s3", "t31"],
      "labels": ["O"],
      "manual": false
    },
    "e-s5-t21": {
      "id": "e-s5-t21",
      "ids": ["s5", "t21"],
      "labels": [],
      "manual": false
    }
  }
}
```


Figure 15: SVALA format (*source, target* and *edges* objects in JSON format).

Northern European Journal of Language Technology, 2019, Vol. 6, Article 4, pp 67–104
DOI 10.3384/nejlt.2000-1533.19667

annotated data can lead to biased – if not erroneous – theoretical conclusions and generalizations, as well as influence teaching and assessing practices in unwanted ways. To circumvent this challenge, the adopted flow of piloting all steps (Figure 3) on project members (before involving assistants) have proven to be a very wise decision. Use of the custom-developed tools that are both user-friendly and rich in visualization, is a further contributing factor to producing reliable annotations.

To ensure the acceptance of data annotation by SLA and LCR researchers (and hopefully teachers) while remaining of interest to NLP researchers, a non-negligible gap in "subject cultures" had to be clozed, one example being the use of the term "error annotation" instead of (the curently adopted term) "correction annotation". The trip between the first discussion on "error annotation" to adoption of "correction annotation" took us two years. A number of methodologically important insights have been gained due to collaboration between the two groups of research cultures, notably separation of normalization from correction annotation, and decision to delegate normalization step to SLA-trained researchers or teachers. Access to system developers and NLP expertise made it possible to guide the development of annotation tools NOT by technology limitations, but by SLA researcher needs. Design of the annotation tool and visualization of each step has been formed in discussion with the representatives of the two cultures, and our hope is that it would be easy to work with both the tool and the data format for all future users.

10 Final remarks

We have presented on-going work on building a new L2 learner corpus of Swedish as a central part of a research infrastructure for Swedish as a second language. Our emphasis is on the importance of ensuring quality of each step of manual corpus annotation. We described the collection, pseudonymization, normalization, and annotation with special focus on a preliminary taxonomy of codes describing corrections introduced to fix learner-produced errors, and the tools used for data preparation and processing. We explained some legal issues and the data management flow involved to ensure anonymity of the participating subjects.

One of the questions left for near future, is setting up a search interface with functionalities that would address SLA research interests. The search interface to the corpus will be based on both Korp (Ahlberg et al., 2013) and Strix. The former is a corpus management tool that presents searches through a concordancer, whereas the latter is document-oriented, allowing larger portions of the corpus to be viewed. It will be possible to construct search expressions using words as well as regular expressions. In the latter case, we plan to allow information from several sources to be independently combined: words in the learner and/or normalized texts, categories in the correction annotation, elements of the linguistic annotation, and metadata associated with the texts (language background, age, etc.). Specifically, the linguistic annotation will be produced by a standard analysis pipeline which includes part-of-speech tagging and dependency annotation. The primary target of this will be the normalized text, but a possibility is to project the annotation of this to the learner text by using the alignments between the two levels.

To summarize, the SweLL infrastructure has been extensively developing towards

opening a possibility for continuous collection and annotation of learner essays. So far four 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 2019-2020. 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. The corpus is planned to be released in 2020 as freely available open-access.

Acknowledgements

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

Abel, Andrea, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2014. KoKo: an L1 Learner Corpus for German. In *Language Resources and Evaluation Conference (LREC)*, pages 2414–2421.

Ahlberg, Malin, 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.

Alexopoulou, Theodora, Marije Michel, Akira Murakami, and Detmar Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning* 67(S1):180–208.

Andringa, Sible and Aline Godfroid. 2019. SLA for all? Reproducing SLA research in non-academic samples. In *OSF Project, published January 25, 2019 (https://osf.io/mp47b/)*.

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4):555–596.

Bird, Steven 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.

Borin, Lars, 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.

Boyd, Adriane. 2017. MERLIN: Lessons Learned. In *Presentation at CLARIN workshop on Interoperability of Second Language Resources and Tools. University of Gothenburg, Sweden. December 2017*.

Boyd, Adriane, 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 *Language Resources and Evaluation Conference (LREC)*, pages 1281–1288.

Carlsen, Cecilie. 2012. Proficiency level—a fuzzy variable in computer learner corpora. *Applied Linguistics* 33(2):161–183.

Church, Kenneth Ward. 2017. Emerging trends: I did it, I did it, I did it, but... *Natural Language Engineering* 23(3):473–480.

Corder, Stephen Pit. 1967. The significance of learner's errors. *IRAL-International Review of Applied Linguistics in Language Teaching* 5(1-4):161–170.

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

Dıaz-Negrillo, Ana, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards interlanguage pos annotation for effective learner corpora in sla and flt. In *Language Forum*, vol. 36, pages 139–154.

Dobric, Nikola. 2015. Quality measurements of error annotation-ensuring validity through reliability. *The European English Messenger* 24:36–42.

Doshi-Velez, Finale and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* .

Erickson, Gudrun and Julieta Lodeiro. 2012. (In Swedish) Bedömning av språklig kompetens–En studie av samstämmigheten mellan Internationella språkstudien 2011 och svenska styrdokument. ISSN 1652-2508. *Skolverkets aktuella analyser* .

Forsberg, Fanny and Inge Bartning. 2010. Can linguistic features discriminate between the communicative CEFR-levels?: A pilot study of written L2 French .

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

Gaillat, T., P. Sébillot, and N. Ballier. 2014. Automated classification of unexpected uses of this and that in a learner corpus of English. *Recent Advances in Corpus Linguistics* pages 309–324.

Geertzen, Jeroen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*.

Golden, Anne, Scott Jarvis, and Kari Tenfjord. 2017. *Crosslinguistic influence and distinctive patterns of language learning: findings and insights from a learner corpus*. Multilingual Matters.

Granger, Sylviane. 1998. The computer learner corpus: a versatile new source of data for SLA research. *Granger, S. (Ed.). Learner English on Computer.* pages 3–18.

Granger, Sylviane. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching. *Corpora and language teaching* 33:13–32.

Granger, Sylviane. 2013. Error-tagged learner corpora and CALL: A promising synergy. *CALICO journal* 20(3):465–480.

Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier. 2015. *The Cambridge handbook of learner corpus research*. Cambridge University Press.

Granger, Sylviane and Magali Paquot. 2017. Towards standardization of metadata for L2 corpora. In *Keynote talk at CLARIN workshop on Interoperability of Second Language Resources and Tools. University of Gothenburg, Sweden. December 2017*.

Gustafson-Capková, Sofia and Britt Hartmann. 2006. *Manual of the Stockholm-Umeå Corpus, version 2.0*. Stockholm University, Stockholm, Sweden (https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf, last accessed October 2019).

Hildén, R. 2008. Analys av svenska kursplaner i relation till den Europeiska Referensramen. *Skolverket.* Internal report. ISBN: 978-91-87115-72-1.

Holdt, Špela Arhar, Iztok Kosem, and Polona Gantar. 2017. Corpus-based resources for L1 teaching: The case of Slovene. In *Handbook on digital learning for K-12 schools*, pages 91–113. Springer.

Housen, Alex and Folkert Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics* 30(4):461–473.

Hovy, Eduard and Julia Lavid. 2010. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation* 22(1):13–36.

Hyland, Ken and Fiona Hyland. 2019. *Feedback in second language writing: Contexts and issues*. Cambridge university press.

Jaccard, Paul. 1908. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles* 44:223–270.

Krippendorff, Klaus. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research* 30(3):411–433.

Kuiken, Folkert and Ineke Vedder. 2007. Task complexity and measures of linguistic performance in l2 writing. *IRAL-International Review of Applied Linguistics in Language Teaching* 45(3):261–284.

Leacock, Claudia, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies* 3(1):1–134.

Lenhard, Alexandra, Wolfgang Lenhard, Sebastian Suggate, and Robin Segerer. 2018. A continuous solution to the norming problem. *Assessment* 25(1):112–125.

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

MacWhinney, Brian. 2017. A shared platform for studying second language acquisition. *Language Learning* 67(S1):254–275.

Madnani, Nitin, Jill Burstein, Norbert Elliot, Beata Beigman Klebanov, Diane Napolitano, Slava Andreyev, and Maxwell Schwartz. 2018. Writing Mentor: Self-Regulated Writing Feedback for Struggling Writers. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 113–117.

Megyesi, Beáta, 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.

Megyesi, Beáta, Jesper Näsman, and Anne Palmér. 2016. The Uppsala corpus of student writings: Corpus creation, annotation, and analysis. In *Language Resources and Evaluation Conference (LREC)*.

Megyesi, Beáta, Anne Palmér, and Jesper Näsman. 2019. (In Swedish) SWEGRAM Användarmanual. (https://cl.lingfil.uu.se/ bea/publ/swegram-manual-2019.pdf – Last accessed October 2019).

Meurers, Detmar, Kordula De Kuthy, Florian Nuxoll, Björn Rudzewitz, and Ramon Ziai. 2019. Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics* 39.

Mitchell, Rosamond, Florence Myles, and Emma Marsden. 2012. *Second language learning theories*, vol. 3 ed. London: Routledge.

Myles, Florence. 2005. Interlanguage corpora and second language acquisition research. *Second Language Research* 21(4):373–391.

Näsman, Jesper, Beáta Megyesi, and Anne Palmér. 2017. SWEGRAM: A Web-Based Tool for Automatic Annotation and Analysis of Swedish Texts. In *21st Nordic Conference on Computational Linguistics, Nodalida 2017*, pages 132–141.

Nivre, Joakim, Marie-Cathrine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of Language Resources and Evaluation Conference (LREC)*.

Norris, John M and Lourdes Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30(4):555–578.

Oscarson, Mats. 2015. (In Swedish) Bedömning på systemnivå - En komparativ studie av stegsystemet i språk i den svenska skolan och språknivåer i Europarådets Common European Framework of Reference. *EDUCARE* 2:128–153.

Pallotti, Gabriele. 2009. CAF: Defining, refining and differentiating constructs. *Applied Linguistics* 30(4):590–601.

Paquot, Magali. 2013. Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics* 18(3):391–417.

Paquot, Magali and Sylviane Granger. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32:130–149.

Paquot, Magali and Luke Plonsky. 2017. Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3(1):61–94.

Parkvall, Mikael. 2009. *(In Swedish) Sveriges språk - vem talar vad och var?*. Institutionen för lingvistik, Stockholms universitet.

Pettersson, Eva, Beáta Megyesi, and Joakim Nivre. 2013. Normalisation of Historical Text using Context-Sensitive Weighted Levenshtein Distance and Compound Splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics, NODALIDA*.

Pilán, Ildikó, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111.

Prentice, Julia and Sylviane Granger. In prep. Error - (still) a controversial notion in SLA research. In *In preparation*.

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

Rosen, Alexandr, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2014. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation* 48(1):65–92.

Rosén, Dan, 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*.

Settles, Burr, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65.

Skehan, Peter and Pauline Foster. 1999. The influence of task structure and processing conditions on narrative retellings. *Language learning* 49(1):93–120.

Skolverket. 2018. Kommunal vuxenutbildning i svenska för invandrare - Elever och kursdeltagare - Riksnivå. (https://www.skolverket.se/skolutveckling/statistik/ Last accessed December 2018.).

Stymne, Sara, Eva Pettersson, Beáta Megyesi, and Anne Palmér. 2017. Annotating errors in student texts: First experiences and experiments. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 47–60.

Tenfjord, Kari, Hilde Johansen, and Jon Erik Hagen. 2006a. The" hows" and the" whys" of coding categories in a learner corpus (or" how and why an error-tagged learner corpus is not'ipso facto'one big comparative fallacy"). *Rivista di psicolinguistica applicata* 6(3):1000–1016.

Tenfjord, Kari, Paul Meurer, and Knut Hofland. 2006b. 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.

Tetreault, Joel, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 48–57.

Thewissen, Jennifer. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal* 97(S1):77–101.

Volodina, Elena, 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* .

Volodina, Elena, Beáta 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 Swedish Language Technology Conference (SLTC) 2016, Umeå, Sweden*.

Volodina, Elena, Ildikó Pilán, Lars Borin, and Therese Lindström Tiedemann. 2014. A flexible language learning platform based on language resources and web services. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3973–3978.

Wirén, Mats, 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* .

Zipser, Florian and Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, Language Resources and Evaluation Conference (LREC) 2010*.

Östling, Robert. 2016. Efficient Sequence Labeling: efselab. https://github.com/robertostling/efselab – Last accessed February 2018, Stockholm University, Stockholm, Sweden.

Östling, Robert, Andre Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. 2013. Automated essay scoring for Swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47.