On two SweLL learner corpora – SweLL-pilot and SweLL-gold

Elena Volodina

University of Gothenburg, Sweden

Abstract

SweLL – Swedish Learner Language – is a unifying term for the infrastructure module for research on Swedish as a Second Language (L2), deployed and maintained as a part of bigger infrastructure of Språkbanken Text at the University of Gothenburg, Sweden. The SweLL infrastructure module consists of a number of learner data collections, and tools for annotation and management of learner data. As a result, many of its components contain the prefix SweLL in their names, which has created some confusion, especially with regards to the two corpora. In this article we shortly introduce the various SweLL-components with a special focus on the differences between the two SweLL corpora.

Keywords

SweLL, learner corpus research infrastructure, Swedish as a second language, correction annotation aka error annotation, normalization, CEFR labels

1. Introduction

Learner corpora are collections of essays written by learners of some language, where essays are used for empirical evidence in research on the development of learner language or in related fields. Some examples are ASK for L2 Norwegian [1], FALKO for L2 German [2, 3], MERLIN for L2 Czech, German and Italian [4], COBLE2 for L2 Portuguese [5], CzeSL for L2 Czech [6], LAVA for L2 Latvian [7], Icelandic L2 Error Corpus [8] and multiple learner corpora for L2 English [e.g. 9, 10, 11].

For L2 Swedish there exist several collections, such as CrossCheck with essays from different levels of schools/courses [12], ASU with L2 essays and transcribed L2 speech [13], and Uppsala Corpus of Student Writings with an extensive collection of essays from Swedish national exams [14]. These corpora are very valuable, reflecting different aspects of L2 Swedish, but are not easy to gain access to.

The SweLL initiative first emerged in 2012, when several smaller collections were offered to Språkbanken Text for processing and maintenance, namely TISUS-texts from 2006 and SW1203 from 2012. In parallel, compilation of another smaller corpus, SpIn, was in progress at Språkbanken Text itself. Being too small to be released as three individual corpora, the three collections were unified and released under the name of a SweLL-corpus in 2016 [15]. This first SweLL compilation was used as a starting point for a grant application with the same name, SweLL - research
infrastructure for Swedish as a second language,\textsuperscript{1} granted by Riksbankens Jubileumsfond for years 2017–2020. During the SweLL project time, a new corpus of L2 Swedish was developed within a highly cross-disciplinary group of collaborators from the University of Gothenburg, Stockholm university, Uppsala university and Umeå university representing fields of Second Language Acquisition, Learner Corpus Research and Natural Language Processing.

In hindsight, it was unfortunate, that the name of the first corpus and the project name were the same, especially since the corpus produced in the SweLL project was, naturally, also called SweLL corpus. One more term using SweLL prefix is the SweLL portal \textsuperscript{16} - an annotation management tool developed within the SweLL project.

In 2021, during preparation of the SweLL release v.1, a decision was made to rename the first of the SweLL-corpora to SweLL-pilot – which largely reflects its nature, being a proof-of-concept and source of wisdom; and to call the project-generated corpus SweLL-gold – which reflects its status as a corpus with a higher standard with extensive correction annotation of high quality.

To summarize the chronological order of the use of the term (or, rather, modifier) SweLL:

- 2016: SweLL corpus \textsuperscript{15} → since 2021: SweLL-pilot corpus
- 2017: SweLL project\textsuperscript{1} (finished in 2021)
- 2018: SweLL portal \textsuperscript{16}
- 2019: SweLL corpus \textsuperscript{17} → since 2021: SweLL-gold corpus (NOTE! incl. SweLL(-gold) target and SweLL(-gold) original subparts)
- 2021: SweLL infrastructure (module)

Below, we shortly introduce the standards for metadata, annotation and file formats in the Learner Corpora Research field (section 2), and zoom into the two SweLL-corpora, SweLL-pilot and SweLL-gold to summarize their similarities and differences (section 3).

2. Ideal infrastructure for learner language

Despite the relatively long history of Learner Corpus Research (LCR), of at least three decades, there is still no agreement about what to consider an ideal standard \textsuperscript{18}, which reflects the dynamic nature of the fields involved. This refers to sets of metadata; which annotation to include and in which standard; data formats for release; and search tools.

Metadata for learner corpora is extremely important for pursuing different types of research and for the interoperability between corpora \textsuperscript{19, 20}. For example, age, gender and first languages are important for identification of learning problems for different demographic groups; task metadata – for studying the impact of the task on the type of language produced by learners in the essays. However, there are many other metadata aspects that are easily overlooked by corpus compilers, although similarly important.

Work on metadata standardization in LCR was initiated by Paquot and Granger in 2017 \textsuperscript{21}, was followed up by König et al. in 2022 \textsuperscript{22} and is still ongoing \textsuperscript{23}. Paquot et al. \textsuperscript{23} identify eight groups of metadata – administrative, corpus design, learner, text, task, annotation, annotator and transcriber\textsuperscript{2} – with multiple subcategories divided into obligatory and optional. In both

\textsuperscript{1}https://spraakbanken.gu.se/en/projects/swell
\textsuperscript{2}Work-in-progress document available here: https://tinyurl.com/L2metadataV2
SweLL corpora, most of the obligatory metadata are considered, however, the metadata for characterizing annotators and transcribers is not among those, which is difficult to rectify post-factum.

**Annotation standards** in LCR cover both manual and automatic annotation, stratified further into linguistic annotation, anonymization (vs pseudonymization), normalization, error correction (vs correction annotation), etc. [18]. Included here are also tools for annotation and annotation management. Most previous projects relied on xml schema for annotation where corrections were assigned to the original strings [e.g. 1, 5], which has recently started to be replaced by alternative approaches, such as viewing original and corrected versions of essays as independent aligned versions of a parallel corpus [e.g. 6, 7]. Unlike most predecessors, the SweLL-gold corpus has been pseudonymized (not anonymized) – i.e. personal information in texts has been substituted by alternative strings to preserve the integrity of learner texts and to conform to the requirements of the GDPR [24]; normalized, i.e. rewritten to an alternative independent corrected version, and corrections were correct-annotated for the nature of the difference between the original and normalized strings. All in all, the SweLL project has contributed (1) to increased attention to the need for structured pseudonymization of learner essays [25, 26]; (2) to an emerging new paradigm of learner corpora where the original and normalized versions are treated as parallel corpora [27]; and (3) to shifting the focus from ‘errors’ in learner versions to their ‘corrections’ since these corrections are only some of several possible hypothetical ways to interpret (errors in) learner writing [28].

The automatic linguistic annotation present in both SweLL corpora comes from Sparv annotation pipeline [29] and contains tokenization, lemmatization, word sense disambiguation, morpho-syntactic annotation, syntactic dependencies and a few others.

**Formats** are largely influenced by the way annotation is conceptualized, such as whether to treat the corrected version as an independent text, or to attach a corrected string directly into the original sentence. However, even the search interfaces set limitations on formats, most

---

**Table 1**

Overview of SweLL-pilot and SweLL-gold statistics per subcorpus

<table>
<thead>
<tr>
<th></th>
<th>SpIn</th>
<th>SweLL-pilot</th>
<th>TISUS</th>
<th>total</th>
<th>SweLL-gold</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year of collection</strong></td>
<td></td>
<td>2012</td>
<td>-2012</td>
<td>2006</td>
<td>2017</td>
<td>2006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2016</td>
<td>-2013</td>
<td>-2016</td>
<td>-2020</td>
<td>-2020</td>
</tr>
<tr>
<td><strong>Nr tokens</strong></td>
<td></td>
<td>46 911</td>
<td>52 518</td>
<td>60 632</td>
<td>160 061</td>
<td>147 842</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 145</td>
<td>3 422</td>
<td>10 869</td>
<td>7 807</td>
<td>(8 137)</td>
</tr>
<tr>
<td><strong>Nr essays</strong></td>
<td></td>
<td>256</td>
<td>141</td>
<td>105</td>
<td>502</td>
<td>502</td>
</tr>
<tr>
<td><strong>Nr A1 essays</strong></td>
<td></td>
<td>59</td>
<td>0</td>
<td>0</td>
<td>59</td>
<td>Beginner: 289</td>
</tr>
<tr>
<td><strong>Nr A2 essays</strong></td>
<td></td>
<td>143</td>
<td>0</td>
<td>0</td>
<td>143</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Nr B1 essays</strong></td>
<td></td>
<td>46</td>
<td>40</td>
<td>0</td>
<td>86</td>
<td>Intermediate: 45</td>
</tr>
<tr>
<td><strong>Nr B2 essays</strong></td>
<td></td>
<td>2</td>
<td>71</td>
<td>32</td>
<td>105</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Nr C1 essays</strong></td>
<td></td>
<td>0</td>
<td>23</td>
<td>73</td>
<td>96</td>
<td>Advanced: 168</td>
</tr>
<tr>
<td><strong>Nr C2 essays</strong></td>
<td></td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Nr Unknown</strong></td>
<td></td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>N/A</td>
</tr>
</tbody>
</table>

---

3In majority of other learner corpora this is called ‘error annotation’
prominently, corpus workbench depending heavily on TEI-XML. Most error-annotated corpora are, therefore, distributed in xml file formats with only a few distributed alternatively also in json format [30]. *SweLL-gold* and *SweLL-pilot* are distributed in three file formats: raw texts, linguistically annotated xml (TEI-XML) and json (in case of *SweLL-gold* containing correction and pseudonymization tags).

**Search tools** are critical for accessing and analyzing learner data, with multiple solutions, often adapted to a corpus in question. For both *SweLL* corpora, it was possible to use Korp [31], where the user can see each subcorpus individually under 'L2 Korp' – SpIn, SW1203, TISUS, SweLL-origial, SweLL-target (Table 1) and perform searches in any combination of those.

### 3. The two SweLL corpora

The *SweLL* (Swedish Learner Language) infrastructure currently contains two *SweLL* corpora (which are shown as five subcorpora in Korp search tool), collected at two different periods of time: *SweLL-pilot* between 2006–2016 [15] and *SweLL-gold* between 2017–2020 [17]. As the name suggests, *SweLL-pilot* was the initial attempt to collect learner essays; whereas *SweLL-gold* is built upon those experiences, accounting for the lessons learnt, correcting the limitations and extending the scope of annotation. Notably, during the *SweLL-gold* period a larger group of researchers and annotators was involved and richer annotation schemes and tools were developed. Table 1 provides an overview of statistics over the two essay collections.

#### 3.1. SweLL-pilot - a corpus of learner essays with CEFR labels

*SweLL-pilot* is a corpus of essays written by adult learners of Swedish during exam settings and collected from students who have signed consents. It was collected during the period of 2006-2016, with the first release of 339 essays in 2016 [15] – transcribed from hand-written essays and anonymized. In 2018, 163 more essays were transcribed, anonymized and added to the *SweLL-pilot* collection. In 2020-2021 the *SweLL-pilot* collection was added to the *SweLL* portal [16] to ensure comparable json format [27] and harmonized metadata attributes with the *SweLL-gold* collection. Nowadays, *SweLL-pilot* contains 502 essays that have been anonymized and labeled with the CEFR levels.

The *SweLL-pilot* collection contains three subcorpora, all of which represent multiple first languages (L1) and age groups:

- **SpIn**[^4] - 256 essays collected from Language Introduction course (mid-term exams) for newly arrived refugees. Some of the students are recurrent.
- **SW1203**[^5] - 141 essays collected from university students in exam setting, most of who wrote three essays each.
- **TISUS**[^6] - 105 essays written as a part of a Test In Swedish for University Studies. All essays are on the same topic "Stress" and within the argumentative genre.

[^4]: SpIn metadata: https://spraakbanken.github.io/swell-release-v1/Metadata-SpIn
[^5]: SW1203 metadata: https://spraakbanken.github.io/swell-release-v1/Metadata-SW1203
[^6]: TISUS metadata: https://spraakbanken.github.io/swell-release-v1/Metadata-TISUS
Table 2  
<table>
<thead>
<tr>
<th>Metadata</th>
<th>Privacy</th>
<th>Normalized</th>
<th>Correct-annotated</th>
<th>CEFR labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>SweLL-pilot</td>
<td>Harmonized</td>
<td>Anonymized</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>SweLL-gold</td>
<td>Harmonized</td>
<td>Pseudonymized</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

Table 2  
Major differences between the annotations present in the two SweLL subcorpora

*SweLL-pilot* is the first and the only CEFR-labeled learner corpus of L2 Swedish. The Inter-annotator agreement on CEFR labeling measured for the SW1203 subcorpus (141 essays) is 0.80% Krippendorff’s alpha [32] which corresponds to high annotation quality.

### 3.2. SweLL-gold - a corpus of learner essays with error annotation

The *SweLL-gold* corpus⁷ [17] was developed within the *SweLL project¹* [33], the purpose of which was to set up an infrastructure for collection, digitization, normalization, and annotation of L2 Swedish adult learner written production, as well as to make available a linguistically annotated parallel corpus, where it would be possible to search for various types of linguistic structures, without the researcher having to guess what such a structure might look like in original essays, since there is a parallel normalized version available.

The essays were collected from several schools around Sweden where teachers assisted with consent forms, personal and task metadata forms, and essays. The type of school was used as an indication of the approximate level of learners, e.g. *upper-secondary* and *university preparatory* courses being representative of 'Advanced' levels (C); *SVA* (Swedish as a Second Language) courses for adults representing 'Intermediate' levels (B); and *SFI* courses (Swedish For Immigrants) representing 'Beginner' levels (A). The original essays were transcribed and normalized (i.e. rewritten in standard Swedish that conforms to grammatical norms); and all corrections were labeled as to their nature, i.e. correct-annotated (in other corpora called error-annotated). The result was an aligned parallel corpus of original and normalized essays with correction labels attached to the aligned segments of the essay.

The *SweLL infrastructure* components, such as *SweLL portal* [16], SVALA annotation tool [27] and multiple guidelines for annotation [34, 35, 36, 37] were developed to ensure high quality of data annotation, which resulted in Inter-Annotator Agreement of 88% by Fleiss’ kappa and 76% by Krippendorff’s alpha [38, 32] as measured on 10% of the essays (i.e. 50 essays).

*SweLL-gold* is the first and the only correction-annotated L2 learner corpus of Swedish.

### 3.3. Differences between SweLL-pilot and SweLL-gold

The general overview of the statistics for the two corpora, provided in Table 1, shows that the size of the corpora is comparable, albeit relatively modest, amounting in total to 1004 essays representing 307 903 tokens. We can also see from Table 2 that the metadata and attribute names have been harmonized between the two corpora. However, there are four critical aspects that differ between the corpora, namely, (1) the way personal data in the text was handled ('Privacy');

---

⁷SweLL-gold metadata: https://spraakbanken.github.io/swell-release-v1/Metadata-SweLL.
absence or presence of a corrected version of the original essay (’Normalized’); (3) absence or presence of the manually assigned labels for corrections (’Correct-annotated’); and (4) absence or presence of CEFR labels (’CEFR labeled’).

This means that the two corpora can practically never be used for the same research questions or applied to the same development problems. For example, if you are interested in Grammatical Error Detection (GED) or Correction (GEC), only SweLL-gold is appropriate. If you want to develop an automatic essay grading (AEG) system or identify a scope of vocabulary/grammar used at particular level, only SweLL-pilot can be used.

In an ideal world, each of the corpora would be complemented for the missing annotation. However, in the real world it demands additional funding to make it happen. On the bright side, SweLL-pilot is currently being pseudonymized in accordance with the standards of the SweLL-gold corpus [35, 39] so that it can be used for work on automatic pseudonymization of research data within the ’Grandma Karl’ project [26]. That version of SweLL-pilot will be released in the future. Normalization and correction annotation of SweLL-pilot, as well as CEFR-labeling of SweLL-gold are, however, left for future.

3.4. Access to the data

The two SweLL corpora contain private information - both in the form of metadata and as private mentions in texts, and are therefore under the GDPR [24] protection. This sets limitations to the openness of data, namely, that only individuals living and working in Europe can have access to the data; with a further restriction that the area of application should be connected to education (teaching, learning, research or development).

Due to that, access to the SweLL corpora is administered through an application form. The approved user gets access to the data in three file formats: raw text, linguistically annotated xml and json; as well as through a corpus search system Korp [31].

4. SweLL impact: a game changer in Swedish L2?

It is a fact that languages and research domains, that can boast rich data collections, have more empirical and data-intensive research done on them [40, 41]. This makes us believe that now, with the SweLL data available for research and development, the field of Swedish as a second language and related research fields will get a boost. Since the release of SweLL-pilot in 2016, we can see a steady increase in interest to (1) development of automatic tools and approaches, such as classification of essays, lexical complexity prediction, error detection and correction [42, 43, 44, 45, 46, 47, 48, 49, 50, 51]; (2) data-driven linguistic studies on vocabulary and grammar scopes in second language learning, grammatical patterns at different levels of linguistic development, etc. [52, 46, 53, 54, 55, 56]; (3) novel approaches to feedback generation [57, 50];

8https://mormor-karl.github.io/
9https://sunet.artologik.net/gu/swell
10https://spraakbanken.gu.se/korp/
(4) methodological studies, such as, pseudonymization of research data, effects of errors on the performance of automatic tools, fairness and bias in language assessment [58, 59, 26], etc.

A number of derivative resources have been developed since 2016 based on the two SweLL corpora, such as wordlists for language learners – SweLLex [52] and later Sen*Lex [45] – for studies on lexical competences of L2 learners; DaLAJ [60] for studies on linguistic acceptability [61], CoDeRooMor [62] for studying derivational morphology of Swedish, MuClaGED [63] for error classification, synthetic datasets imitating real-life errors [64] and many others. The Swedish MultiGED dataset\(^{11}\) based on SweLL-gold has been used for the MultiGED shared task [48] and we plan new shared tasks based on the SweLL corpora in the near future.

5. Future directions

We are expecting both short-term and long-term impact from the two corpora described in this article on the fields of Swedish as a Second language, Learner Corpus Research (nationally and internationally), and NLP- and AI-based approaches to L2 Swedish.

First of all, we intend to promote the use of the datasets among NLP researchers through organization of multilingual shared tasks.\(^{12}\)

Second, we will work towards extending authentic learner datasets through setting on-the-fly pseudonymization algorithms for continuous collection of essays directly from schools.

In parallel, we will also work on generation of synthetic datasets with basis in the current SweLL data, for example experimenting with GPT models to generate mock learner essays at different levels of proficiency, using real-life essays as samples, or generating error datasets using linguistic patterns observed in the SweLL-gold data.

Finally, we will search for possibilities to harmonize the two SweLL corpora (and potential other subcorpora that will be added to the SweLL infrastructure module) between each other through normalization and correction annotation of SweLL-pilot and CEFR-labeling of SweLL-gold. We do not exclude that these steps will be performed automatically (with subsequent manual proofreading) after we have experimented with the automatic approaches to normalization, correction annotation and CEFR labeling.

Acknowledgments

Work on the article has been supported by Nationella språkbanken and Huminfra, both funded by the Swedish Research Council (2018-2024, contract 2017-00626; 2022-2024, contract 2021-00176) and their participating partner institutions. Work on SweLL-pilot was supported by the Center for Language Technology at the University of Gothenburg (SpIn and SW1203), and by Ebba Danelius stiftelse för sociala och kulturella ändamål and Granholms stiftelse, at Stockholm University (TISUS). Work on SweLL-gold has been supported by the infrastructure grant from the Swedish Riksbankens Jubileumsfond SweLL – research infrastructure for Swedish as a second language, grant IN16-0464:1.

\(^{11}\)https://github.com/spraakbanken/multiged-2023

\(^{12}\)https://spraakbanken.gu.se/en/compsla
References

[16] Y. A. Mohammed, A. Matsson, E. Volodina, Annotation Management Tool: A Requirement


