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

Elena Volodina

University of Gothenburg, Sweden

Abstract

SweLL – **Swe**dish **L**earner **L**anguage – 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], COPLE2 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*

Huminfra Conference 2024, Gothenburg, 10–11 January 2024.

elena.volodina@svenska.gu.se (E. Volodina)

infrastructure for Swedish as a second language,[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* [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 refects 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* [15] → since 2021: *SweLL-pilot corpus*
- 2017: *SweLL project*[1] (finished in 2021)
- 2018: *SweLL portal* [16]
- 2019: *SweLL corpus* [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 Reseach 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 [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 [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 [21], was followed up by König et al. in 2022 [22] and is still ongoing [23]. Paquot et al. [23] identify eight groups of metadata – administrative, corpus design, learner, text, task, annotation, annotator and transcriber[2] – with multiple subcategories divided into obligatory and optional. In both

[1] https://spraakbanken.gu.se/en/projects/swell

[2] Work-in-progress document available here: https://tinyurl.com/L2metadataV2

	SweLL-pilot				SweLL-gold		
	SpIn	**SW1203**	**TISUS**	**total**	**original**	**normalized**	**Total**
Year of collection	2012 -2016	2012 -2013	2006	**2006 -2016**	2017 -2020		**2006 -2020**
Nr tokens	46 911	52 518	60 632	**160 061**	147 842	(151 851)	**307 903**
Nr sentences	4 302	3 145	3 422	**10 869**	7 807	(8 137)	**18 676**
Nr essays	256	141	105	**502**	502	(502)	**1 004**
Nr A1 essays	59	0	0	**59**	Beginner: 289		N/A
Nr A2 essays	143	0	0	**143**			N/A
Nr B1 essays	46	40	0	**86**	Intermediate: 45		N/A
Nr B2 essays	2	71	32	**105**			N/A
Nr C1 essays	0	23	73	**96**	Advanced: 168		N/A
Nr C2 essays	0	7	0	**7**			N/A
Nr Unknown	6	0	0	**6**			N/A

Table 1
Overview of SweLL-pilot and SweLL-gold statistics per subcorpus

SweLL corpora, most of the obligatory metadata are considered,[4567] 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 independened 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 independed corrected version, and corrections were *correct-annotated*[3] 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 autoomatic 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

[3] In 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

	Metadata	Privacy	Normalized	Correct-annotated	CEFR labeled
SweLL-pilot	Harmonized	Anonymized	no	no	yes
SweLL-gold	Harmonized	Pseudonymized	yes	yes	no

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* corpus[7] [17] was developed within the *SweLL project*[1] [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)';

[7] SweLL-gold metadata: https://spraakbanken.github.io/swell-release-v1/Metadata-SweLL

(2) absense or presence of a corrected version of the original essay ('Normalized'); (3) absense or presence of the manually assigned labels for corrections ('Correct-annotated'); and (4) absense 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[8] [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.[9] 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[10] [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];

[8] https://mormor-karl.github.io/

[9] https://sunet.artologik.net/gu/swell

[10] https://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

[1] K. Tenfjord, P. Meurer, K. Hofland, The ASK corpus: A language learner corpus of Norwegian as a second language, in: LREC'06, 2006, pp. 1821–1824.

[2] A. Lüdeling, M. Walter, E. Kroymann, P. Adolphs, Multi-level error annotation in learner corpora, in: Proceedings of corpus linguistics, volume 1, Citeseer, 2005, pp. 14–17.

[3] M. Reznicek, A. Lüdeling, C. Krummes, F. Schwantuschke, Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.0, Humboldt-Universität zu Berlin, Berlin, Germany, 2012.

[4] A. Boyd, J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schöne, B. Štindlová, C. Vettori, The MERLIN corpus: Learner Language and the CEFR, in: LREC'14, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014.

[5] A. Mendes, S. Antunes, M. Janssen, A. Gonçalves, The COPLE2 corpus: a learner corpus for Portuguese., in: LREC'16, 2016.

[6] A. Rosen, J. Hana, B. Vidová Hladká, T. Jelínek, S. Škodová, B. Štindlová, Compiling and annotating a learner corpus for a morphologically rich language: CzeSL, a corpus of non-native Czech, Nakladatelství Karolinum, 2020.

[7] R. Darģis, I. Auzina, K. Levāne-Petrova, I. Kaija, Detailed Error Annotation for Morphologically Rich Languages: Latvian Use Case, in: Human Language Technologies–The Baltic Perspective, IOS Press, 2020, pp. 241–244.

[8] I. Glisic, A. K. Ingason, The nature of Icelandic as a second language: An insight from the learner error corpus for Icelandic, in: CLARIN Annual Conference, 2022, pp. 23–33.

[9] H. Yannakoudakis, T. Briscoe, B. Medlock, A new dataset and method for automatically grading ESOL texts, in: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011, pp. 180–189.

[10] M. Paquot, Corpora and Second Language Acquisition, in: The Routledge Handbook of Corpora and English Language Teaching and Learning, Routledge, 2022, pp. 26–40.

[11] O. Vinogradova, O. Lyashevskaya, Review of Practices of Collecting and Annotating Texts in the Learner Corpus REALEC, in: International Conference on Text, Speech, and Dialogue, Springer, 2022, pp. 77–88.

[12] J. Lindberg, G. Eriksson, CrossCheck-korpusen - en elektronisk L2-korpus för skriven svenska., in: B. de Geer A. Malmberg (Red.), Språk på tvärs Rapport från ASLA:s höstsymposium Södertörn, 11-12 november 2004, Svenska föreningen för tillämpad språkvetenskap. Uppsala 2005, 2004, pp. 89–98.

[13] B. Hammarberg, Introduktion till ASU-korpusen: En longitudinell muntlig och skriftlig textkorpus av vuxna inlärares svenska med en motsvarande del från infödda svenskar., 2010.

[14] B. Megyesi, J. Näsman, A. Palmér, The Uppsala corpus of student writings: Corpus creation, annotation, and analysis, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 3192–3199.

[15] E. Volodina, I. Pilán, I. Enström, L. Llozhi, P. Lundkvist, G. Sundberg, M. Sandell, SweLL on the rise: Swedish learner language corpus for European reference level studies, in: Proceedings of the 10th Language Resources and Evaluation Conference (LREC), Portorož, Slovenia, 2016.

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

for Corpus Construction, in: CLARIN Annual Conference, 2022, pp. 101–108.

[17] E. Volodina, L. Granstedt, A. Matsson, B. Megyesi, I. Pilán, J. Prentice, D. Rosén, L. Rudebeck, C.-J. Schenström, G. Sundberg, et al., The SweLL language learner corpus: From design to annotation, Northern European Journal of Language Technology (NEJLT) 6 (2019) 67–104.

[18] E. W. Stemle, A. Boyd, M. Jansen, T. Lindström Tiedemann, N. Mikelić Preradović, A. Rosen, D. Rosén, E. Volodina, Working together towards an ideal infrastructure for language learner corpora, Widening the Scope of Learner Corpus Research (2019).

[19] E. Volodina, M. Janssen, T. L. Tiedemann, N. M. Preradovic, S. K. Ragnhildstveit, K. Tenfjord, K. de Smedt, Interoperability of second language resources and tools, in: Proceedings of the CLARIN Annual Conference, 2018, pp. 90–94.

[20] A. König, J.-C. Frey, E. W. Stemle, Exploring reusability and reproducibility for a research infrastructure for l1 and l2 learner corpora, Information 12 (2021) 199.

[21] S. Granger, M. Paquot, Core Metadata [Schema] for Learner Corpora Draft 1.0, 2017.

[22] A. König, J.-C. Frey, E. W. Stemle, A. Glaznieks, M. Paquot, Towards standardizing LCR metadata, in: Book of Abstracts from the Learner Corpus Research Conference, Italy, 2022.

[23] M. Paquot, A. König, E. W. Stemle, J.-C. Frey, A core metadata schema for L2 data, in: Book of Abstracts from the EuroSLA Conference 2023, 2023.

[24] E. EU Commission, General Data Protection Regulation., Official Journal of the European Union, 59, 1-88., 2016. URL: https://gdpr-info.eu/(Accessed2019-11-19).

[25] E. Volodina, Y. A. Mohammed, S. Derbring, A. Matsson, B. Megyesi, Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 357–369.

[26] E. Volodina, S. Dobnik, T. L. m Tiedemann, X.-S. Vu, Grandma Karl is 27 years old–research agenda for pseudonymization of research data, in: 2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService), IEEE, 2023, pp. 229–233.

[27] M. Wirén, A. Matsson, D. Rosén, E. Volodina, Svala: Annotation of second-language learner text based on mostly automatic alignment of parallel corpora, in: CLARIN Annual Conference, Pisa, Italy, 8-10 October, 2018, Linköping University Electronic Press, 2019, pp. 222–234.

[28] L. Rudebeck, G. Sundberg, On the other side of the error tag: The nature and functions of the corrected texts, in: Book of Abstracts from the Learner Corpus Research Conference 2022, Italy, 2022, p. 103.

[29] M. Hammarstedt, A. Schumacher, L. Borin, M. Forsberg, Sparv 5 User Manual, Technical Report, Göteborg, 2022.

[30] Š. Arhar Holdt, I. Kosem, Šolar, the developmental corpus of Slovene (2023).

[31] L. Borin, M. Forsberg, J. Roxendal, Korp – the corpus infrastructure of Språkbanken, in: Proceedings of LREC 2012. Istanbul: ELRA, volume Accepted, 2012, p. 474–478.

[32] K. Krippendorff, Content analysis: An introduction to its methodology, Sage publications, 2018.

[33] E. Volodina, B. Megyesi, M. Wirén, L. Granstedt, J. Prentice, M. Reichenberg, G. Sundberg, A friend in need?: Research agenda for electronic Second Language infrastructure, in:

Swedish Language Technology Conference (SLTC) 2016, 2016.

[34] E. Volodina, B. Megyesi, SweLL transcription guidelines, L2 essays (2021). URL: https://gupea.ub.gu.se/handle/2077/69429.

[35] B. Megyesi, L. Rudebeck, E. Volodina, SweLL pseudonymization guidelines, 2021. URL: http://hdl.handle.net/2077/69431.

[36] L. Rudebeck, G. Sundberg, SweLL correction annotation guidelines (2021). URL: https://gupea.ub.gu.se/handle/2077/69434.

[37] L. Rudebeck, G. Sundberg, M. Wirén, SweLL normalization guidelines (2021). URL: https://gupea.ub.gu.se/handle/2077/69432.

[38] R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics, Computational linguistics 34 (2008) 555–596.

[39] B. Megyesi, L. Granstedt, S. Johansson, J. Prentice, D. Rosén, C.-J. Schenström, G. Sundberg, M. Wirén, E. Volodina, Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish, in: Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning, LiU Electronic Press, Stockholm, Sweden, 2018, pp. 47–56. URL: https://aclanthology.org/W18-7106.

[40] M. Perc, The Matthew effect in empirical data, Journal of The Royal Society Interface 11 (2014) 20140378. doi:https://doi.org/10.1098/rsif.2014.0378.

[41] A. Søgaard, Should We Ban English NLP for a Year?, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 5254–5260.

[42] I. Pilán, E. Volodina, T. Zesch, 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, 2016, pp. 2101–2111.

[43] I. Pilán, E. Volodina, Investigating the importance of linguistic complexity features across different datasets related to language learning, in: Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing, 2018, pp. 49–58.

[44] D. Alfter, E. Volodina, Towards single word lexical complexity prediction, in: Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications, 2018, pp. 79–88.

[45] D. Alfter, Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective, Data Lingvistica 31, University of Gothenburg, 2021.

[46] D. Alfter, T. L. Tiedemann, E. Volodina, Crowdsourcing relative rankings of multi-word expressions: Experts versus non-experts, Nothern European Journal of Language Technology Vol. 7 (2022).

[47] M. Nyberg, Grammatical error correction for learners of Swedish as a second language, 2022.

[48] E. Volodina, C. Bryant, A. Caines, O. De Clercq, J.-C. Frey, E. Ershova, A. Rosen, O. Vinogradova, MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection, in: Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning, 2023, pp. 1–16.

[49] R. Östling, K. Gillholm, M. Kurfalı, M. Mattson, M. Wirén, Evaluation of really good grammatical error correction, arXiv preprint arXiv:2308.08982 (2023).

[50] A. Masciolini, E. Volodina, D. Dannlls, Towards automatically extracting morphosyntactical error patterns from L1-L2 parallel dependency treebanks, in: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), 2023, pp. 585–597.

[51] J. Ehnroth, Y. Park, Correction of Grammatical Errors in Swedish (2023).

[52] E. Volodina, I. Pilán, L. Llozhi, B. Degryse, T. François, SweLLex: second language learners' productive vocabulary, in: Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition, 2016, pp. 76–84.

[53] E. Volodina, D. Alfter, T. L. Tiedemann, Crowdsourcing ratings for single lexical items: a core vocabulary perspective, Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave 10 (2022) 5–61.

[54] T. Lindström Tiedemann, D. Alfter, Y. Ali Mohammed, D. Piipponen, B. Silén, E. Volodina, Multi-word Expressions in Swedish as a second language – typology, annotation and initial results Submitted (2024).

[55] G. Sundberg, J. Prentice, SweLL: En svensk inlärarkorpus, ASLA:s skriftserie/ASLA Studies in Applied Linguistics 30 (2023) 428–453. doi:https://doi.org/10.17045/sthlmuni.24321526428.

[56] E. Volodina, Y. Ali Mohammed, T. Lindström Tiedemann, Swedish Word Family – Construction, Applicability, Strengths and first Experiments (????).

[57] A. Masciolini, A query engine for L1-L2 parallel dependency treebanks, in: Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), University of Tartu Library, Tórshavn, Faroe Islands, 2023, pp. 574–587. URL: https://aclanthology.org/2023.nodalida-1.57.

[58] E. Volodina, Y. Ali Mohammed, S. Derbring, A. Matsson, B. Megyesi, Towards Privacy by Design in Learner Corpora Research: A Case of On-the-fly Pseudonymization of Swedish Learner Essays, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 357–369. URL: https://aclanthology.org/2020.coling-main.32. doi:10.18653/v1/2020.coling-main.32.

[59] E. Volodina, D. Alfter, T. Lindström Tiedemann, M. S. Lauriala, D. H. Piipponen, Reliability of automatic linguistic annotation: native vs non-native texts, in: Selected papers from the CLARINAnnual Conference 2021, Linköping University Electronic Press (LiU E-Press), 2022.

[60] E. Volodina, Y. A. Mohammed, A. Berdičevskis, G. Bouma, J. Öhman, DaLAJ-GED-a dataset for Grammatical Error Detection tasks on Swedish, in: Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning, 2023, pp. 94–101.

[61] J. Klezl, Y. A. Mohammed, E. Volodina, Exploring Linguistic Acceptability in Swedish Learners' Language, in: Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning, 2022, pp. 84–94.

[62] E. Volodina, Y. A. Mohammed, T. L. Tiedemann, CoDeRooMor: A new dataset for non-inflectional morphology studies of Swedish, in: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), 2021, pp. 178–189.

[63] J. C. Moner, E. Volodina, Swedish MuClaGED: A new dataset for Grammatical Error Detection in Swedish, in: Proceedings of the 11th Workshop on NLP for Computer Assisted

Language Learning, 2022, pp. 36–45.

[64] J. C. Moner, E. Volodina, Generation of Synthetic Error Data of Verb Order Errors for Swedish, in: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), 2022, pp. 33–38.