DaLAJ - a dataset for linguistic acceptability judgments for Swedish

Elena Volodina[1], Yousuf Ali Mohammed[1], Julia Klezl[2]
University of Gothenburg, Sweden
[1]name.surname1.surname2@svenska.gu.se
[2]gusklezju@student.gu.se

Abstract

We present DaLAJ 1.0, a **Da**taset for **L**inguistic **A**cceptability **J**udgments for Swedish, comprising 9 596 sentences in its first version. DaLAJ is based on the SweLL second language learner data (Volodina et al., 2019), consisting of essays at different levels of proficiency. To make sure the dataset can be freely available despite the GDPR regulations, we have sentence-scrambled learner essays and removed part of the metadata about learners, keeping for each sentence only information about the mother tongue and the level of the course where the essay has been written. We use the normalized version of learner language as the basis for DaLAJ sentences, and keep only one error per sentence. We repeat the same sentence for each individual correction tag used in the sentence. For DaLAJ 1.0 four error categories of 35 available in SweLL are used, all connected to lexical or word-building choices. The dataset is included in the SwedishGlue benchmark.[1] Below, we describe the format of the dataset, our insights and motivation for the chosen approach to data sharing.

1 Introduction

Grammatical and linguistic acceptability is an extensive area of research that has been studied for generations by theoretical linguists (e.g. Chomsky, 1957), and lately by cognitive and computational linguists (e.g. Keller, 2000; Lau et al., 2020; Warstadt et al., 2019). Acceptability of sentences is defined as "the extent to which a sentence is permissible or acceptable to native speakers of the language." (Lau et al., 2015, p.1618), and there have been different approaches to studying it. Most work views acceptability as a binary phenomenon: the sentence is either acceptable/ grammatical or not (e.g. Warstadt et al., 2019). Lau et al. (2014) show that the phenomenon is in fact gradient and is dependent on a larger context than just one sentence. While most experiments are theoretically-driven, the practical value of this research has been also underlined, especially with respect to language learning and error detection (Wagner et al., 2009; Heilman et al., 2014; Daudaravicius et al., 2016).

Datasets for acceptability judgments require linguistic samples that are unacceptable, which requires a source of so-called negative examples. Previously, such samples have been either manually constructed, artificially generated through machine translation (Lau et al., 2020), prepared by automatically distorting acceptable samples e.g. by deleting or inserting words or inflections (Wagner et al., 2009) or collected from theoretical linguistics books (Warstadt et al., 2019). Using samples produced by language learners has not been mentioned in connection to acceptability and grammaticality studies. However, there are obvious benefits of getting authentic errors that automatic systems may meet in real-life. Another benefit of reusing samples from learner corpora is that they often contain not only corrections, but also labels describing the corrections. The major benefit, though, is that (un)acceptability judgments come from experts, i.e. teachers, assessors or trained assistants, and are therefore reliable.

[1]SwedishGlue (Swe. SuperLim) is a collection of datasets for training and/or evaluating language models for a range of Natural Language Understanding (NLU) tasks.

Elena Volodina, Yousuf Ali Mohammed and Julia Klezl 2021. DaLAJ – a dataset for linguistic acceptability judgments for Swedish. *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)*. Linköping Electronic Conference Proceedings 177: 28–37.

Categories	Explanation	A-lev	B-lev	C-lev	Total
O-Comp	Problem with compounding	252	62	232	546
L-Der	Word formation problem (derivation or compounding)	193	124	404	721
L-FL	Non-Swedish word corrected to Swedish word	46	17	26	89
L-W	Wrong word or phrase	1157	562	1723	3442
Total		**1648**	**765**	**2385**	**4798**

Table 1: Dataset overview, with number of sentences per correction tag, level and in total

Approximate level	Nr essays	Nr labels
A:Beginner	289	11 180
B:Intermediate	45	5 119
C:Advanced	168	12 986
Total	**502**	**29 285**

Table 2: Statistics over the SweLL data

2 Dataset description

We use the error-annotated learner corpus SweLL (Volodina et al., 2019) as a source of "unacceptable" sentences and select sentences containing corrections of the type that is of relevance to the SwedishGlue benchmark[2] (Adesam et al., 2020).

In the current version, four *lexical error types* are included into the DaLAJ dataset (see Section 2.2). The resulting dataset contains 4 798 sentence pairs (correct-incorrect), where the two sentences in each sentence pair are identical to each other except for one error. In total, DaLAJ 1.0 contains 9 596 sentences (which is a sum of unacceptable sentences and their corrected "twin" sentences). To compare, Lau et al. (2014) use a dataset of 2 500 sentences and Warstadt et al. (2019) have about 10 700 sentences for a similar task. We have a possibility to extend the DaLAJ dataset by other correction types (spelling, morphological or syntactical) in future versions. The full SweLL dataset contains 29 285 correction tags, of which 25 878 may become relevant for the current task (omitting punctuation, consequence and unintelligibility correction tags).

2.1 The source corpus

The SweLL data (Volodina et al., 2019) has been collected over four years (2017-2020) from adult learners of Swedish from formal educational settings, such as courses and tests. The collection contains about 680 pseudonymized essays in total, with 502 of those manually normalized (i.e. rewritten to standard Swedish) and annotated for the nature of the correction (aka error annotation). Table 2 shows the statistics over SweLL in number of essays and correction tags per level. Levels of the sentences correspond to the level of the course that learners were taking when they wrote essays. The essays represent several levels, namely:

`A` - beginner level
`B` - intermediate level
`C` - advanced level

The data is saved in two versions: the original and the normalized, with correction labels assigned to the links between the two versions. The 502 corr-annotated essays contain 29 285 corrections distributed over 35 correction tags, as listed in Appendix A.

2.2 Selection of (un)grammatical sentences

The linguistic acceptability task in the SwedishGlue is described as a natural language understanding (NLU) task conceptualized as binary judgments from a perspective relevant for research on language learning, language planning etc. (Adesam et al., 2020). Semantic aspects of the sentence are the main focus of this task. This deviates from the type of language included into the CoLA dataset available through GLUE (Warstadt et al., 2019), where also morphological and syntactic violations are included. In DaLAJ 1.0, we have selected four correction types from the SweLL corpus that would maximally correspond to the need of semantic interpretation of the context, namely `L-W`, `L-Der`, `L-FL`, `O-Comp` (Rudebeck and Sundberg, 2020), described below.

L-W: Wrong word or phrase. The `L-W` tag represents the correction category *wrong word or phrase*. It is used when a word or phrase in the original text has been replaced by another word or

[2]SwedishGlue is a collection of datasets for training and/or evaluating language models for a range of Natural Language Understanding (NLU) tasks.

original sentence	corrected sentence	error indices (original)	corrected indices	error-corr pair	error label	L1	Approximate level
Förr i tiden kunde vi byta en sak till en annan .	Förr i tiden kunde vi byta en sak mot en annan .	34-37	34-36	till--mot	L-W	Polska	B:Fortsättning
Jag kan ta några exempel av betydelsen av pengar .	Jag kan ta några exempel på betydelsen av pengar .	25-26	25-26	av--på	L-W	Somaliska	B:Fortsättning
Och där efter köper vi nästan allt vi behöver .	Och därefter köper vi nästan allt vi behöver .	4-12	4-11	där efter--därefter	O-Comp	Somaliska	B:Fortsättning
För det andra är det nyckeln av livet .	För det andra är det nyckeln till livet .	29-30	29-32	av--till	L-W	Somaliska	B:Fortsättning
Det är svårt ibland men det kommer inte på engång .	Det är svårt ibland men det kommer inte på en gång .	43-48	43-49	engång--en gång	O-Comp	Somaliska	B:Fortsättning
Bostäder i D-hemland är ett lite hett topik att diskutera .	Bostäder i D-hemland är ett lite hett ämne att diskutera .	38-42	38-41	topik--ämne	L-FL	Ungerska	A:Nybörjare
Bostäder i D-hemland är ett lite hett ämne att diskussera .	Bostäder i D-hemland är ett lite hett ämne att diskutera .	47-56	47-55	diskussera--diskutera	L-FL	Ungerska	A:Nybörjare
Många lägenheter och hus är gamla och har dålig energi-niveå .	Många lägenheter och hus är gamla och har dålig energinivå .	48-59	48-57	energi-niveå--energinivå	O-Comp	Ungerska	A:Nybörjare
Man betalar mycket i vintern , vilket är likadant som i Sverige .	Man betalar mycket på vintern , vilket är likadant som i Sverige .	19-19	19-20	i--på	L-W	Ungerska	A:Nybörjare
Man betalar mycket på vintern , vilket är det samma som i Sverige .	Man betalar mycket på vintern , vilket är likadant som i Sverige .	42-50	42-49	det samma--likadant	L-W	Ungerska	A:Nybörjare

Figure 1: An excerpt from the dataset

phrase in the normalized version. It is placed on units which are exchanged rather than corrected. For example,

`Alla blir *busiga med sociala medier →`
`Alla blir upptagna med sociala medier`

which may be verbatim translated as

`Everyone is *naughty with social media →`
`Everyone is busy with social media`

Note the Engligh influence on the use of the word `*busiga` to convey the meaning that someone is `*busy` (Swe `upptagen`), the Swedish word `busig` meaning `mischievous, naughty`.

L-Der: Word formation. The `L-Der` tag represents the correction category *deviant word formation*. It is used for corrections of the internal morphological structure of word stems, both with regard to compounding and to derivation.

The `L-Der` tag is exclusively used for links between one-word units (not necessarily one-token units, since a word may mistakenly be written as two tokens), where the normalized word has kept at least one root morpheme from the original word, but where another morpheme has been removed, added, exchanged or had its form altered. For example,

`De är *stressiga på grund av studier →`
`De är stressade på grund av studier`

which may be translated as

`They are *stressy because of the studies →`
`They are stressed because of the studies`

Note that `*stressiga` uses an existing derivation affix `-ig(a)`, which is wrong in this context, instead of the correct suffix `-ade, stressade`.

L-FL: Foreign word corrected to Swedish. The `L-FL` tag is used for *words from a foreign (non-Swedish) language* which have been corrected to a Swedish word. It may also be applied to words which have certain non-Swedish traits due to influence from a foreign language. For example,

`Jag och min *family →`
`Jag och min familj`
`English: I and my family`

O-Comp: Spaces and hyphens between words. The `O-Comp` tag is used for corrections which involve the removal of a space between two words which have been interpreted as making up a compound in the normalized text version, or, more rarely, the adding of a space between two words.

It may also be used for corrections regarding the use of hyphens in *compounds*. Some examples,

```
Jag kände mig *jätte *konstig →
Jag kände mig jättekonstig
English: I felt very strange
```

Distribution of the correction tags in the DaLAJ 1.0 dataset is shown in Table 1.

2.3 Data format

The task of linguistic acceptability judgments is traditionally performed on the *sentence level*, where each sentence includes *maximum one deviation*. In real life learner-written sentences may contain several errors, but it has been shown that training algorithms on samples with focus on one error only produces better results than when mixing several errors in one sentence; extending the context to a paragraph may further improve the results (Katinskaia and Yangarber, 2021). Paragraphs in learner data, however, are not predictable or well defined, and on several occasions in the SweLL data entire essays consist of one paragraph only. Including in the DaLAJ dataset full paragraphs, in certain cases equivalent to full essays, entails risks of revealing author identities through indications of author-related events or other identifiers despite our meticulous work on pseudonymization of essays (Volodina et al., 2020; Megyesi et al., 2018). We assess, therefore, that we have no possibility to include paragraphs into the dataset due to the restrictions imposed by the GDPR, so we follow the generally accepted standard of single sentences with single deviations.

For each correction label used in the corpus data, we take the corrected target sentence and preserve only one erroneous segment in it to make it "unacceptable". This means that the same sentence can be repeated several times in the dataset, with different segments/deviations being in the focus. Positive samples are represented by the corrected sentences. We have data in a tab separated file format, with eight columns, namely:

1. Original (i.e. unacceptable) sentence, e.g. `Men pengarna är inte *alls` (Eng. But money is not *at all)
2. Corrected sentence, e.g. `Men pengarna är inte allt` (Eng. But money is not everything)
3. Error string indices, e.g. `21-24`
4. Correct string indices, e.g. `21-24`
5. Error-correction pair, e.g. `alls-allt`
6. Error label, e.g. `L-W`
7. Mother tongue(s) (L1), e.g. `Somali`
8. Approximate level, e.g. `B:Intermediate`

Figure 1 shows an excerpt from the dataset. Note that some of the sentences in the "Corrected sentence" column are repeated more than once. The corresponding original sentences contain a new error focus each time. The dataset is (by default) balanced with respect to the number of correct and incorrect samples, however, correct samples contain a number of duplicates which should be complimented by a corresponding number of unique correct sentences, which is something we will add in the next release of the dataset. The dataset is not equally balanced as far as number of sentences per level or per correction code are concerned, which is a more challenging problem.

CoLA dataset authors have explicitly tested that the vocabulary used in their dataset belongs to the 100 000 most frequent words in the language (Warstadt et al., 2019). In the case of DaLAJ, we have not done any such investigation since we believe that the vocabulary used by second language learners cannot be so advanced as to be outside the 100K most frequent words.

Initial experiments on the dataset, data splits and first baselines are reported in an extended version of this article, available at arXiv.org. The DaLAJ 1.0 dataset is freely available at the SwedishGlue webpage.[3]

3 First analysis

We see multiple advantages to use the proposed format for L2 data. Apart from a potential to share the data with wider community of researchers, it also (1) helps expand the data (each original sentence potentially generating several sentences) and (2) helps focus on one error only, facilitating fine-grained analysis of model performance as well as human evaluation of model predictions.

Our analysis has suggested, that the DaLAJ 1.0 dataset needs to be cleaned in several ways. First, the SweLL corpus contains a number of essays where learners add reference lists by the end of essays. Naturally, punctuation in reference lists is non-standard, among others not always containing full stop which sabbotages sentence segmentation. Besides, references are syntactically elliptical and do not fit into the standard language. We would

[3] https://spraakbanken.gu.se/en/resources/swedishglue

need to clean the dataset of all such sentences to ensure more objective training and testing.

Second, some sentences contain "hanging" titles or e-mail headers. Those hanging elements have not been separated by a full stop in the original essays, and have been prefixed to the next following sentence, which, again, can interfere with model training, e.g. `(Swe) En B-institution-entusiast Hej Segerstad kommun ! > (Eng) A B-institute-enthusiast Hi Segerstad municipality !`

Yet another observed weakness of the DaLAJ 1.0 dataset, is that the positive sentences are repetitive. Since the models need to be trained on unique samples, we plan to exchange the non-unique ones with other sentences. Luckily, positive samples are easier to find than negative ones. We plan to use a corpus of L2 coursebooks graded for levels of proficiency, COCTAILL (Volodina et al., 2014), to replace duplicate sentences with the ones of equivalent level, and as far as possible, having similar linguistic features and length. Another potential source of in-domain positive sentences are SweLL sentences that do not contain any correction tags. However, such sentences are not many, and we would still need to use COCTAILL sentences or some other correct sentences.

The described changes will be introduced in DaLAJ 1.1 and in the test test for DaLAJ 1.0.

Finally, there is an important difference between the type of sentences used in CoLA and DaLAJ datasets. CoLA sentences are constructed manually for linguistic course books exemplifying various theoretically important linguistic features, and do not require wider context to interpret; whereas DaLAJ sentences are torn out of their natural context, and contain anaphoric references and elliptical structures. However, the applied value of training (machine learning) algorithms on DaLAJ sentences is higher than CoLA sentences (as we imagine that) since such models can be used in language learning context for writing support.

4 Reflections on access to learner data

Datasets and corpora collected from (second) language learners contain private information represented both on the metadata level and - depending on the topic - in the texts. Presence of personal information makes those datasets non-trivial to share with the public in a FAIR[4] way (Frey et al., 2020; Volodina et al., 2020), to say nothing of a potential to use such data for *shared tasks*. This is rather unfortunate since collection and preparation of such corpora is an extremely time-consuming and expensive process. Language learner datasets can seldom boast big sizes appropriate for training data-greedy machine learning algorithms, and could therefore benefit from aggregating data from several sources - provided they are accessible. Access to such data, besides, ensures transparency of the research and stimulates its fast development (MacWhinney, 2017; Marsden et al., 2018).

As data owners, we have to face two contradictory forces: one requiring open sharing, and the other preventing it. Among advocates for sharing data openly we see

• national and international funding agencies, e.g. Swedish Research Council[5] or European Commission[6], requiring guarantees from grant holders that any produced data will be made available for other researchers,

• national and international infrastructures, e.g. Clarin[7] or SLABank,[8] and

• updated journal policies (e.g. The Modern Language Journal).[9]

On the more restrictive side, we have national Ethical Review Authorities[10] and the General Data Protection Regulation, GDPR (Commission, 2016), described shortly below.

The Swedish Ethical Review Authority currently requires that we keep the original data (e.g. hand-written/ non-transcribed/ non-pseudonymized essays) for ten years after the project end so that researchers, who may question the trustworthiness of the original data handling, can require access to the original data for inspection. This means that the data owners need to keep mappings between learner names and their corpus IDs to make it possible to link de-identified and pseudonymized essays to their original versions.

General Data Protection Regulation sets certain

[4]FAIR: Findable, Accessible, Interoperable, Reusable (Wilkinson et al., 2016)

[5]https://www.vr.se/english/mandates/open-science/open-access-to-research-data.html

[6]https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-access_en

[7]https://www.clarin.eu/

[8]https://slabank.talkbank.org/

[9]https://onlinelibrary.wiley.com/journal/15404781

[10]https://www.government.se/government-agencies/the-swedish-ethics-review-authority-etikprovningsmyndigheten/

limitations on the data where personal data occurs, among others:

• learner identities should be protected, e.g. pseudonymized or de-identified;

• data need to be removed if any of the data providers (=learners) requests that;

• users that are granted access to the data should have affiliation inside Europe; and

• questions that users can work with are limited to the ones stated in the consent forms, in the case of SweLL encompassing research on and didactic applications for language learning.

To meet these requirements, data owners need to administer data access through an application form, where applicants have to be asked about their geographical location and research questions, and need to be informed about the limitations of spreading data to unauthorized users, etc. Users outside Europe can file an application to the university lawyers who have to consider them on a case-to-case basis. The GDPR applies to the data as long as a mapping of learner names with their corpus IDs (as required by the Ethical Review Authorities) is not destroyed. At a certain point of time (currently 10 years) the mapping key will be destroyed and the data will no longer be under the GDPR protection.

In both cases, a 10-year quarantine is obligatory. The restrictions above do not seem to hamper most of the potential EU-based researchers from getting access to the data in its entirety, especially researchers working with qualitative analysis of the data inside a limited project group, e.g. Second Language Acquisition researchers or researchers on language assessment. However, when it comes to the NLP field, the most effective way to stimulate research is to organize *shared tasks* or provide access to testing and evaluation datasets without any extra administration, as it is, for example, done in the GLUE[11] and SuperGLUE[12] benchmarks (Wang et al., 2018, 2019).

From the above it follows that data owners need to keep a promise to the funding agencies to make the data open, and at the same time, to follow the legislation and keep the data locked within Europe and only for research questions dealing with language learning. Being representatives of a "trapped researcher" group, we have been considering how to make learner data available for a wider audience. For a range of NLP tasks we suggest, thus, sharing L2 data in a sentence scrambled way with limited amount of socio-demographic metadata, for example for error detection & correction tasks. The DaLAJ dataset is a proof-of-concept attempt in this direction.

Ultimately, the education NLP community working with L2 datasets would win by setting up a benchmark with available (multilingual) datasets in the same way as GLUE benchmark is doing for Natural Language Understanding (NLU) tasks.

5 Concluding remarks

We have presented a new dataset for Swedish which can be used for a variety of tasks in Natural Language Processing (NLP) or Second Language Acquisition (SLA) contexts. We see our contributions both with regards to the dataset, as well as with suggesting a format for L2 datasets that may allow sharing learner data more openly in the GDPR age.

In the near future, we will test binary linguistic acceptability classification on the current selection of correction categories, and on the full SweLL dataset (all correction tags), per error category and level, establishing baselines for this task on this dataset. We plan to correlate the classification results with correction categories, levels and L1s. Further, we plan to apply models, trained on DaLAJ, to real learner data containing multiple errors per sentence, to assess the effect of data manipulation (i.e. original essays > DaLAJ format) on algorithm training. Proofreading the dataset and addressing identified weaknesses and errors is another direction for the future work.

In some more distant future we would like to organize shared tasks using DaLAJ. Apart from binary classification for linguistic acceptability judgments, we see a potential of using DaLAJ dataset (in extended version to cover the full correction tagset) for a range of other tasks, including:

• error detection (identification of error location)

• error classification (labeling for error type)

• error correction (generating correction suggestions)

• first language identification (given samples written by learners, to identify their mother tongues)

• classification of sentences by the level of proficiency of its writers, and other potential tasks.

[11] https://gluebenchmark.com/

[12] https://super.gluebenchmark.com/

Acknowledgments

This work has been supported by *Nationella Språkbanken* – jointly funded by its 10 partner institutions and the Swedish Research Council (dnr 2017-00626), as well as partly supported by a grant from the Swedish Riksbankens Jubileumsfond (SweLL - research infrastructure for Swedish as a second language, dnr IN16-0464:1).

References

Yvonne Adesam, Aleksandrs Berdicevskis, and Felix Morger. 2020. SwedishGLUE–Towards a Swedish Test Set for Evaluating Natural Language Understanding Models. *Research Reports from the Department of Swedish. GU-ISS-2020-04.*

Noam Chomsky. 1957. *Syntactic structures*. The Hague: Mouton.

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

Vidas Daudaravicius, Rafael E Banchs, Elena Volodina, and Courtney Napoles. 2016. A report on the automatic evaluation of scientific writing shared task. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 53–62.

Jennifer-Carmen Frey, Alexander König, and Darja Fišer. 2020. Creating a learner corpus infrastructure: Experiences from making learner corpora available. In *ITM Web of Conferences*, volume 33, page 03006. EDP Sciences.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180.

Anisia Katinskaia and Roman Yangarber. 2021. Assessing Grammatical Correctness in Language Learning. In *Proceedings of the Sixteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Frank Keller. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Ph.D. thesis.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2014. Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1618–1628.

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

Emma Marsden, Luke Plonsky, A Gudmestad, and A Edmonds. 2018. Data, open science, and methodological reform in second language acquisition research. *Critical reflections on data in second language acquisition*, 51:219–228.

Beéta Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén, and Elena Volodina. 2018. Learner corpus anonymization in the age of gdpr: Insights from the creation of a learner corpus of swedish. In *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018) at SLTC, Stockholm, 7th November 2018*, 152, pages 47–56. Linköping University Electronic Press.

Lisa Rudebeck and Gunlög Sundberg. 2020. Correction annotation guidelines. SweLL project. Technical report.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, et al. 2019. The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology*, 6:67–104.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beáta Megyesi. 2020. 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*, pages 357–369.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 128–144.

Joachim Wagner, Jennifer Foster, Josef van Genabith, et al. 2009. Judging grammaticality: Experiments in

sentence classification. *Calico Journal*, 26(3):474–490.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Appendices

Appendix A. Overview of all correction types in the source corpus

Categories	Explanation	A-lev	B-lev	C-lev	Total
Orthography (3 codes)					*4381*
O	Spelling error	1746	754	769	3269
O-Cap	Upper / lower case	264	73	229	566
O-Comp	Problem with compounding	252	62	232	546
Lexical level (4 codes)					*4876*
L-Der	Word formation problem (derivation or compounding)	193	124	404	721
L-FL	Non-Swedish word corrected to Swedish word	46	17	26	89
L-Ref	Choice of anaphoric expression	214	112	298	624
L-W	Wrong word or phrase	1157	562	1723	3442
Morphological level (8 codes)					*8005*
M-Adj/adv	Adjective word form corrected to adverb form	45	13	46	104
M-Case	Nominative vs genitive / accusative cases	46	64	147	257
M-Def	Definiteness: articles; form of nouns and adjectives	1056	362	1550	2968
M-F	Grammatical category changed, form kept	168	50	82	300
M-Gend	Gender problem (neuter / uter)	370	131	452	953
M-Num	Number problem (plural / singular)	351	157	523	1031
M-Other	Other corrections, incl. comparative forms of adjectives	55	28	33	116
M-Verb	Verb forms; auxiliaries	984	489	803	2276

Figure 2: Overview of all correction types in the SweLL corpus, part 1

Syntactical level (11 codes)					*7696*
S-Adv	Word order: Adverbial placement	235	131	419	785
S-Comp	Compound vs multi-word expressions; lex-synt restructuring	32	8	96	136
S-Clause	Change of basic clause structure; synt function of components	387	210	532	1129
S-Ext	Extensive and complex correction / restructuring	133	65	112	310
S-FinV	Word order: Finite verb placement	283	142	276	701
S-M	Word missing (i.e. added in the target)	719	375	810	1904
S-Msubj	Subject missing (i.e. added in the target)	175	74	185	434
S-Other	Other syntactical correction	20	20	40	80
S-R	Word redundant (i.e. removed in the target)	501	235	687	1423
S-Type	Change of phrase type / part of speech	209	111	275	595
S-WO	Word order: other	67	40	92	199
Punctuation (4 codes)					*1834*
P-M	Punctuation missing (added in the target)	643	312	879	1834
P-R	Punctuation redundant (removed in the target)	133	85	226	444
P-Sent	Sentence segmentation	6	7	26	39
P-W	Wrong punctuation	127	66	244	437
Other (5 codes)					*1573*
C	Consistency correction, necessitated by another correction	397	205	606	1208
Cit-FL	Non-Swedish word kept (i.e. no correction in the target)	14	0	40	54
Com!	Comments for the corpus users	50	1	58	109
OBS!	Internal temporary comments to annotators	9	7	49	65
X	Unintelligible string	93	27	17	137
TOTAL					*29 285*

Figure 3: Overview of all correction types in the SweLL corpus, part 2