DaLAJ-GED -
a dataset for Grammatical Error Detection tasks on Swedish

Elena Volodina[1], **Yousuf Ali Mohammad**[1],
Aleksandrs Berdicevskis[1], **Gerlof Bouma**[1], **Joey Öhman**[2]

[1]Språkbanken Text, Department of Swedish, Multilingualism, Language Technology,
University of Gothenburg, name.surname1.surname2@gu.se
[2]AI Sweden, name.surname1.surname2@ai.se

Abstract

DaLAJ-GED is a dataset for linguistic acceptability judgments for Swedish, covering five head classes: lexical, morphological, syntactical, orthographical and punctuation. DaLAJ-GED is an extension of DaLAJ.v1 dataset (Volodina et al., 2021a,b). Both DaLAJ datasets are based on the SweLL-gold corpus (Volodina et al., 2019) and its correction annotation categories.

DaLAJ-GED presented here contains 44,654 sentences, distributed (almost) equally between correct and incorrect ones and is primarily aimed at linguistic acceptability judgment task, but can also be used for other tasks related to grammatical error detection (GED) on a sentence level. DaLAJ-GED is included into the Swedish SuperLim 2.0 collection,[1] an extension of SuperLim (Adesam et al., 2020), a benchmark for Natural Language Understanding (NLU) tasks for Swedish.

This paper gives a concise overview of the dataset and presents a few benchmark results for the task of linguistic acceptability, i.e. binary classification of sentences as either correct or incorrect.

1 Introduction

The DaLAJ dataset has been inspired by the English CoLA dataset (Warstadt et al., 2019) and, like the CoLA dataset, is primarily aimed at linguistic acceptability judgments as a way to check the ability of models to distinguish correct language from incorrect. Other members of the CoLA-family are represented by, among others, RuCoLA for Russian (Mikhailov et al., 2022), NoCoLA for Norwegian (Samuel and Jentoft, 2023), ItaCoLA for Italian (Trotta et al., 2021), CLiMP for Chinese (Xiang et al., 2021) and a few others. Unlike most of the CoLA datasets that contain artificially constructed incorrect sentences, DaLAJ is based on originally written learner essays and learner errors in SweLL-gold corpus (Volodina et al., 2019). The DaLAJ approach as a way to create datasets for linguistic acceptability judgments has been introduced in Volodina et al. (2021a). A follow-up on this approach is presented in Samuel and Jentoft (2023) for Norwegian based on the ASK corpus (Tenfjord et al., 2006).

The Swedish DaLAJ – Dataset for Linguistic Acceptability Judgments – is a part of SuperLim, the Swedish equivalent of the English SuperGLUE (Wang et al., 2019) benchmark for NLU tasks.

2 Dataset description

The DaLAJ-GED dataset contains 44,654 sentences, of which 22,539 are incorrect sentences from the SweLL-gold corpus (Volodina et al., 2019) and 22,115 are correct ones from both SweLL-gold and Coctaill (Volodina et al., 2014) corpora (Table 1).

Split	Correct sent	Incorr. sent	Total sent	Total tokens
Train	17,472	18,109	35,581	603,625
Dev	2,424	2,278	4,702	77,251
Test	2,219	2,152	4,371	72,349
Total	**22,115**	**22,539**	**44,654**	**753,225**

Table 1: Sentence and token counts in DaLAJ-GED

[1]https://spraakbanken.gu.se/resurser/superlim

Elena Volodina, Yousuf Ali Mohammed, Aleksandrs Berdicevskis, Gerlof Bouma and Joey Öhman. DaLAJ-GED - a dataset for Grammatical Error Detection tasks on Swedish. *Proceedings of the 12th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2023)*. Linköping Electronic Conference Proceedings 197: 94–101.

Figure 1: Sample of a DaLAJ-GED sentence in the Huggingface repository for SuperLim.
Literal translation: 'Are they really most important [thing] in the life?'. Expected: *Är de verkligen **det** viktigaste i livet?* 'Are they really **the** most important [thing] in life?'

Column	**Explanation/values**	**Example**
Sentence		`Är de verkligen viktigaste i livet?`
Label	`correct` or `incorrect`	`incorrect`
Error span: start	character index, as counted from 0 in the sentence	`16`
Error span: stop	character index, as counted from 0 in the sentence; half-open range	`16` (in this case, the range [16, 16] denotes an empty string)
Confusion pair: incorrect span	string representing the error token(s) or empty	
Confusion pair: correction	string representing the correct version	`det`
Error label	one or more error labels describing the same error segment. Values: `P`unctuation, `O`rthography, `L`exical, `M`orphology, `S`yntax)	`M`
Education level	`Nybörjare`, `Fortsättning` , `Avancerad` ('Beginner', 'Intermediate', 'Advanced')	`Fortsättning`
L1	mother tongue(s), full names in Swedish	`Polska` ('Polish')
Data source	DaLAJ/SweLL or Coctaill	`DaLAJ/SweLL gold`

Table 2: DaLAJ-GED columns using the example from Figure 1

Each learner-written sentence is associated with the writer's mother tongue(s) and information about the level of the course at which the essay was written. Perhaps unsurprisingly, the number of fully correct sentences in the learner essays is lower than the number of sentences that contain some mistake. To compensate for this imbalance, we added correct sentences from the Coc-

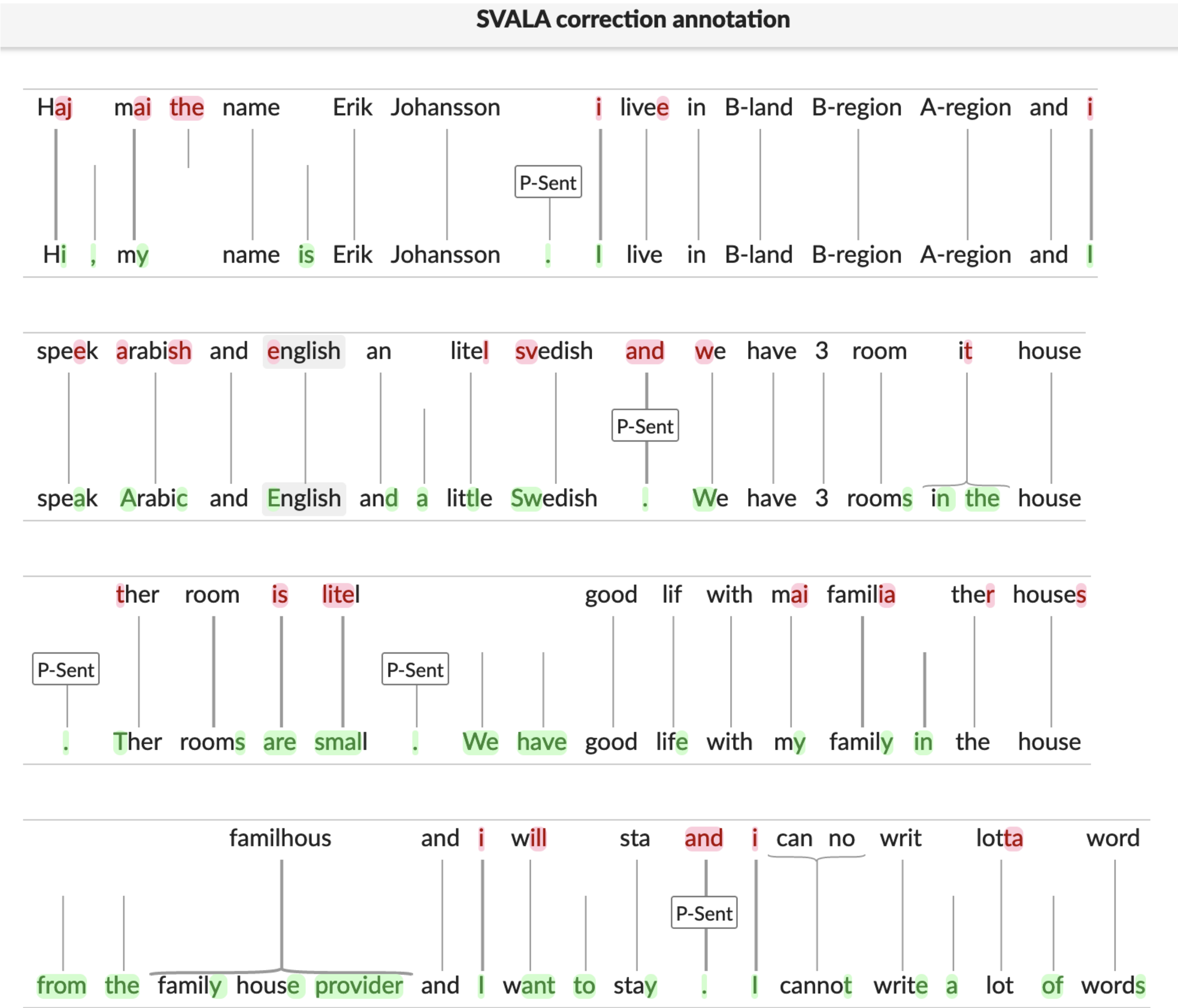

Figure 2: A mock-up translation of an original SweLL-gold sentence. Note the one-to-many (1-to-5) relation between the number of sentences in the original (the top row) and the number of sentences in the target version (the second row). Label `P-Sent` indicates a punctuation correction leading to a sentence split or merge.

taill corpus of coursebooks aimed at second language learners of Swedish (Volodina et al., 2014), keeping the same distribution over beginner-intermediate-advanced levels as among the incorrect sentences. For that, CEFR labels (CoE, 2001) used in Coctaill, have been grouped into (approximate) levels:

- beginner: A1-A2 levels;
- intermediate: B1-B2 levels;
- advanced: C1 level (C2 missing in Coctaill).

This version of DaLAJ is an official improved variant of the previously tested experimental version presented in Klezl et al. (2022).

DaLAJ-GED is distributed as part of Superlim 2.0[2] in a `jsonl` format (primarily), but is also available in tab-separated `tsv` format. See Figure 1 and Table 2 for a description of items / columns in the `jsonl` / `tsv` representations. The example sentence *Är de verkligen viktigaste i livet?* can be literally translated as 'Are they really most important [thing] in life?' and is missing an obligarory definite article (determiner) *det*. A correct Swedish counterpart would be *Är de verkligen **det** viktigaste i livet?* 'Are they really **the** most important [thing] in life?'). The incorrect token is thus an empty string (i.e. the correct token *det* is omitted).

[2]https://github.com/spraakbanken/SuperLim-2

2.1 Source corpora

The **SweLL-gold corpus** (Volodina et al., 2019), used as a source of incorrect sentences, is an error-annotated corpus of learner Swedish. It contains

Current		Replacement suggestion
A-,B-,C-,D-	geoplats	Fafjällen, Undberget, Baraön, Lokomitt
A-,B-,C-,D-	hemland	Brasil, Spanien, Irak, Kina
A-,B-,C-,D-	institution	Volvodrömmen, Linsbiblioteket, Forkecentralen, Bungavård
A-,B-,C-,D-	land	Danmark, Mongoliet, Sudan, Peru
A-,B-,C-,D-	plats	Burocentrum, Andeplats, Storetorg, Bungafors
A-,B-,C-,D-	skola	Buroskola, Andeskola, Storeskola, Bungahjulet
A-,B-,C-,D-	region	Sydlunda, Undered, Hanskim, Bungalarna
A-,B-,C-,D-	stad	Oslo, Paris, Bagdad, Caracas
A-,B-,C-,D-	svensk-stad	Sydden, Norrebock, Rosaborg, Ögglestad
A-,B-,C-,D-	linjen	buss

Table 3: Pseudonymized strings and suggestion for their replacement

502 essays written by adult learners of Swedish at different levels of proficiency (beginner, intermediate, advanced) and representing 81 unique mother tongues in 117 unique combinations of 1-4 languages. The essays represent different topics and genres, some examples being ”Describe your lodging”, ”My first love”, ”Discuss marriage and other lifestyles”, book and film reviews, etc.[3] All essays have been first pseudonymized, then rewritten to represent correct language (i.e normalized) and finally differences between the original and normalized versions were annotated with correction labels (aka error labels).

The **COCTAILL corpus** (Volodina et al., 2014), used as a source of correct sentences for DaLAJ-GED, is a corpus of textbooks used for teaching Swedish to adult second language learners. Each chapter in each textbook is annotated with CEFR labels (A1, A2, B1, B2, C1). The labels are projected to all texts used in each particular chapter, and subsequently to all sentences used in those texts. Texts represent various topics and various genres, including narratives, dialogues, fact texts, instructions, etc.

2.2 Preparation steps

For DaLAJ, only 1-to-1 mappings between original and corrected sentences in SweLL-gold (Volodina et al., 2019) have been used, i.e. where segmentation at the sentence level was unambiguos. Cases like the one mocked in Figure 2 were excluded from DaLAJ. Sentences containing labels `X` (unintelligible string) and `Unid` (unidentified type of correction) were also excluded. Note that the sentences are presented in random order to prevent the possibility to restore original essays – which is a prerequisite for sharing the dataset openly.

To generate several one-error DaLAJ sentences from multi-error original SweLL sentences, we started from the normalized/corrected sentences and projected one error from the original sentences at a time. This means that every incorrect sentence taken from SweLL occurs as many times in DaLAJ as the number of errors it contains. Sometimes, the same token/segment could be described by a cluster of error tags, which were then projected as a group to the single error segment, e.g. *Jag i Stockholm borr* (’I in Stockholm leave’), where *leave* (correct version ’live’) is both misspelled (label `O`) and has word order problem with the placement of a finite verb (label `S-FinV`). All resulting incorrect sentences therefore have exactly one error segment with one or more labels describing that error segment. As such, DaLAJ sentences are neither original, nor artificial, and are best described as hybrid ones.

In a post-processing step, we paid special attention to a class of errors called *consistency corrections* in the SweLL-gold annotation (label: `C`). This label was assigned when a correction was a follow-up of another correction. For example, when a sentence-initial mistake *I slutligen* ‘In finally’ is corrected to *Slutligen* ‘Finally’, the capitalization of *Slutligen* is in a sense a consequence of the correction of the erroneous preposition, and therefore it is marked as a consistency correction. In out-of-context sentences the `C` category is not self-explanatory. Therefore, we excluded in a few

[3] A summary of corpus characteristics is provided in the metadata file: `https://spraakbanken.github.io/swell-release-v1/Metadata-SweLL`

cases such sentences and replaced the `C` label with a label that describes the error more precisely in others. In case of *slutligen* → *Slutligen*, this is the label `O-Cap` (orthographical correction of capitalization).

Due to anonymization of the learner essays in SweLL, the dataset contains pseudonyms of the form *D-stad* 'D-city', *A-linje* 'A-line', etc. We suspect them to be disruptive for automatic tools. Before using the dataset for training and testing, we suggest, therefore, replacing those pseudonyms with more realistic-looking (sometimes nonsense) names like the ones suggested in Table 3.

The incorrect DaLAJ sentences are split into training, development and test sets, the proportion being approximately 80:10:10 of the whole number of sentences. The development and test sets were manually proofread to ensure the quality.

Finally, the incorrect sentences were complemented with correct ones from the COCTAILL corpus.

3 Tasks

DaLAJ-GED is prepared for several *sentence-level tasks*:

Linguistic Acceptability Judgments is the primary task (and the only official SuperLim task). Given a sentence, detect whether it contains any errors (`incorrect`) or not (`correct`), i.e. the task is to perform binary classification on a sentence level.

Grammatical Error Detection (GED) Given a sentence, detect which token(s) need to be corrected, and provide their start-and-end indices, e.g., the omission of `det` with indices `[16-16)` in the example in Table 2.

Multi-Class GED Given a sentence, classify what types of errors need to be corrected, by head classes (punctuation, orthography, lexical, morphology, syntax [POLMS]), e.g.
`[16,16)` → `M` (Morphological error).

Grammatical Error Correction (GEC) Given the incorrect sentence, rewrite it to obtain a correct version, e.g.

`Är de verkligen viktigaste i livet?`
→
`Är de verkligen` **`det`** `viktigaste i livet?`

4 Acceptability judgments – official SuperLim benchmark

The SuperLim benchmark contains various datasets to evaluate the capability of language models. In this paper we present results for the task of acceptability judgments on the DaLAJ-GED dataset that were produced in the context of the SuperLim projekt.

Table 4 shows the results of the initial baseline models on DaLAJ-GED for the task of linguistic acceptability judgments. The horizontal line separates transformer models (Vaswani et al., 2017; Acheampong et al., 2021) from the more traditional machine learning systems and random baselines.

SuperLim by default uses Krippendorff's α coefficient (Krippendorff, 2004) as its metric for summarizing system performance on the different tasks. Krippendorff's α is a measure of agreement where 1 indicates a perfect score and 0 indicates that the system's predictions are at chance level. Clearly negative scores indicate systematic mispredictions. Krippendorff's α is given in Table 4 together with the standard accuracy metric for reasons of familiarity.

Part of the SuperLim benchmark is a leaderboard website,[4] which makes it possible to compare models and opens for an asynchronous competition focused on Swedish. The results for the baseline models presented here applied to a range of SuperLim tasks are included on this leaderboard. The website also contains a more detailed explanation for the choice of Krippendorff's α.

Each transformer model was fine-tuned as demonstrated in Devlin et al. (2019) on the training split with a binary classification learning objective, using Huggingface with early stopping and a coarse-grained hyperparameter tuning with respect to the development split. The hyperparameter space was inspired by RoBERTa (Liu et al., 2019), see Table 5, with the remaining hyperparameters left as the Huggingface default values. The results indicate that larger models typically perform better and that Swedish pre-trained models perform better than multilingual variants. Moreover, the transformer models significantly outperform traditional systems. A comparison of the α and Accuracy metrics shows that they mostly demonstrate the same picture here, albeit on a different scale. However, for the two worst perform-

[4] `www.example.org` (to be supplied)

Model	α	Acc
KBLab/megatron-bert-large-swedish-cased-165k	**0.753**	**0.877**
KBLab/bert-base-swedish-cased-new	**0.753**	0.876
AI-Nordics/bert-large-swedish-cased	0.745	0.872
KB/bert-base-swedish-cased	0.740	0.870
xlm-roberta-large	0.738	0.869
KBLab/megatron-bert-base-swedish-cased-600k	0.718	0.860
xlm-roberta-base	0.701	0.851
NbAiLab/nb-bert-base	0.644	0.822
SVM	0.518	0.758
Decision Tree	0.269	0.636
Random	0.007	0.503
Random Forest	-0.312	0.498
Majority label (`incorrect`)	-0.340	0.492

Table 4: SuperLim results for a selection of models on DaLAJ-GED task, reported in Krippendorff's alpha coefficient (Superlim's default measure) and accuracy.

Hyperparameter	Value(s)
Learning Rate	{1e-5, 2e-5, 3e-5, 4e-5}
Batch Size	{16, 32}
Warmup Ratio	0.06
Weight Decay	0.1
Max Epochs	10

Table 5: Hyperparameter configuration for fine-tuning transformer models

ing systems, we see very low α-scores, whereas Accuracy hovers around the .5 mark. This is because these models grossly overpredict one of the labels, a characteristic that is punished by α.

The results suggest that the dataset is of a size and quality that is sufficient for neural models. An interesting further comparison could be with human baselines, which is a potential future step.

Replicability Each pre-trained language model is publicly available on Huggingface, with the model names as presented here. The traditional baselines are implemented using the scikit-learn Python library (Pedregosa et al., 2011). Full source code and instructions for reproducing the results are made publicly available on GitHub.[5]

Pre-trained language models Below we provide additional details and references to a few of the most prominent language models in the results. In the official SuperLim benchmark, the best-performing model in terms of the average score is KBLab/megatron-bert-large-swedish-cased-165k.[6] This 340M parameter model is trained and published by KBLab[7] and was trained for 165K steps using a batch size of 8K. It was trained on about 70GB of textual data, consisting mostly of OSCAR (Suárez et al., 2019; Ortiz Suárez et al., 2020) and Swedish newspapers curated by the National Library of Sweden.

The second best model, AI-Nordics/bert-large-swedish-cased[8] is of the same size and trained for 600K steps with a batch size of 512. The training data is composed of various sources of internet data and sums to about 85GB.

Among the smaller pre-trained language models, KB/bert-base-swedish-cased[9] (Malmsten et al., 2020) is the greatest performing model, trained on 15-20GB text from a mix of data deposited at the National Library of Sweden and internet data. The model's pre-training consisted of two steps as presented in the original BERT article. First, it was trained 1M steps with a sequence length of 128 and batch size of 512, and then 100K steps with a sequence length of 512 and batch size of 128.

[5]https://github.com/JoeyOhman/SuperLim-2-Testing

[6]https://huggingface.co/KBLab/megatron-bert-large-swedish-cased-165k

[7]https://huggingface.co/KBLab

[8]https://huggingface.co/AI-Nordics/bert-large-swedish-cased

[9]https://huggingface.co/KB/bert-base-swedish-cased

5 Concluding remarks

The contributions of the DaLAJ-GED are twofold. First, efforts like DaLAJ, SuperLim and similar stimulate development of models and approaches to languages other than English, correcting the existing dominance of English in the NLP field (Søgaard, 2022). We expect an increased interest to Swedish NLP following the release of DaLAJ-GED and other SuperLim datasets. The dataset can also be used by researchers who do not have any specific interest in Swedish, but need a high-quality benchmark in order to evaluate transfer learning from another language (e.g. English).

Second, DaLAJ-GED supports the area of automatic method development for Swedish learner language, since it offers not only the data for testing models' general ability to differentiate between correct and incorrect language, but – additionally – offers tasks within second language learning domain for sentence-level grammatical error detection (GED), error classification and error correction (GEC).

DaLAJ-GED complements two other recently released SweLL-gold derivative datasets relevant for second language domain, namely, Swedish MultiGED dataset for error detection on a token level[10] (Volodina et al., 2023) and Swedish MuClaGED dataset for error classification on a token level (Moner and Volodina, 2022). Next steps would be to prepare datasets for feedback generation and for error correction in a larger context than a single sentence as well as in authentic context.

Acknowledgments

The work on the dataset and benchmarking was supported by the Vinnova project *Superlim 2.0*, through the grant 2021-04165. Work on the dataset was also partially funded by a grant from the Swedish Riksbankens Jubileumsfond (*SweLL - research infrastructure for Swedish as a second language*, dnr IN16-0464:1), and 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.

[10]https://github.com/spraakbanken/multiged-2023

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, pages 1–41.

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.

CoE. 2001. *Council of Europe. Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Julia Klezl, Yousuf Ali Mohammed, and Elena Volodina. 2022. Exploring Linguistic Acceptability in Swedish Learners' Language. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 84–94.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian Corpus of Linguistic Acceptability. *arXiv preprint arXiv:2210.12814*.

Judith Casademont Moner and Elena Volodina. 2022. Swedish MuClaGED: A new dataset for Grammatical Error Detection in Swedish. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 36–45.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

1703–1714, Online. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

David Samuel and Matias Jentoft. 2023. NoCoLA: The Norwegian Corpus of Linguistic Acceptability. In *The 24rd Nordic Conference on Computational Linguistics*.

Anders Søgaard. 2022. Should We Ban English NLP for a Year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Kari Tenfjord, Paul Meurer, and Knut Hofland. 2006. The ASK corpus–a language learner corpus of Norwegian as a second language.

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. *arXiv preprint arXiv:2109.12053*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Volodina, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen, and Olga Vinogradova. 2023. MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*.

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, and Julia Klezl. 2021a. DaLAJ–a dataset for linguistic acceptability judgments for Swedish. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 28–37.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021b. DaLAJ-a dataset for linguistic acceptability judgments for Swedish: Format, baseline, sharing. *arXiv preprint arXiv:2105.06681*.

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.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

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

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. *arXiv preprint arXiv:2101.11131*.