SweLL-UD: A Treebank of L2 Swedish Essays

Arianna Masciolini
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg, Sweden
arianna.masciolini@gu.se

April 29, 2024

Keywords treebank, L2 Swedish, learner essays, corrections

1 Background

Over the last few years, Universal Dependencies (UD) has grown to include manually annotated and/or validated treebanks of second language (L2) Chinese [4], English [1, 3] and Italian [2]. UD learner corpora have obvious advantages in terms of cross-lingual consistency, allowing for comparisons both between learner and standard language and different L2s, but also in terms of ease of annotation, which greatly benefits from the availability of increasingly fast and reliable parsers.

In particular, most of the aforementioned treebanks consist of sentences extracted from learner essays and paired with grammatical corrections, which also come with UD annotation. This format, proposed by [5] under the name *L1-L2 parallel dependency treebank*[1], is intended as a replacement for explicit error labelling, as it makes it possible to carry out fine-grained error analysis by comparing learner productions to the corresponding corrections.

The author of this proposal put this idea to the test, developing a tree query-based error retrieval system [6, 7] and a treebank-driven method for automatic error pattern extraction [8]. Their effectiveness and, as a consequence, the validity of the "L1-L2 approach" as a whole, is however greatly dependent on annotation quality: while promising results were obtained on manually validated treebanks such as [1] and [2], preliminary experiments on sentence-correction pairs parsed with state-of-the-art tools show that automatic annotation of ungrammatical language is often too inconsistent for reliable error retrieval and analysis.

All this motivates further annotation efforts in this direction: new L1-L2 treebanks of any size can already be studied with the existing tools, while larger-scale datasets could be used to train parser models specifically meant for L2 productions.[2]

2 Objectives

The ultimate goal of this project is to build an L1-L2 treebank of Swedish learner essays sufficiently large and high-quality to train (or fine-tune) an L2 Swedish parser. In the shorter term, however, we aim at releasing a fully manually validated test set, comparable to [4] and [2].

This proposal is mostly concerned with this first step, which crucially includes the development of Swedish L2 annotation guidelines. As mentioned above, such a treebank can already be valuable for small-scale cross-lingual studies. Furthermore, the hope is that this first annotation sub-project will result in insights useful for the faster development of a training-scale treebank. We might, for instance, come to the conclusion that the quality of automatic annotation is sufficiently good for corrected sentences and/or that the overwhelming majority of the necessary manual edits in original learner sentences concern the erroneous segments, leaving plenty of room for automation.

[1] In the expression "L1-L2 treebank", "L1" (first language) refers to the fact that corrections are assumed to be native-like.
[2] Note that, of the currently available L1-L2 treebanks, only [1] has a training split.

3 Data

Data for this project comes from the SweLL Swedish Learner Language corpus, a collection of 502 manually pseudonymized, error-tagged and corrected essays written by adult learners of L2 Swedish [9].

The final treebank, SweLL-UD, will consist of 5000+ sentence-correction pairs extracted from the essays, 80% of which will initially be set aside for a future training set. The remaining 20% will be further halved into a development and a test split, leaving us with about 500 test sentence-correction pairs to annotate during the first stage of the project.

Despite pseudonymization, the full SweLL dataset is not publicly available for privacy reasons and one of the prerequisites for releasing portions of it is to avoid distributing complete essays. As a consequence, the treebank will consist of sentence pairs in random order and sensitive metadata will be omitted. Since UD analysis is at the sentence level, this is not a too big limitation, even if a minority of grammatical corrections are only meaningful in a larger context.

4 Project status and plan

Creating the initial 500-sentence test set involves 4 steps:

1. **data preprocessing**: sentences pairs and relevant metadata are extracted from the SweLL corpus, converted to CoNNL-U format, shuffled and split into a train, development and test set;
2. **L1 annotation**: corrections are automatically annotated with a pretrained model and manually validated by a small team of students and researchers at the University of Gothenburg;
3. **L2 annotation**: corrections are automatically annotated with an *ad-hoc* parser (see below) and manually validated by the same annotators according to L2 Swedish-specific guidelines, to be defined in parallel with the first two steps;
4. **treebank release**, accompanied by a paper providing a description of the resource and experience report of the annotation project.

At the time of writing, data preprocessing is almost complete. Annotating corrections should be relatively straightforward thanks to the existing language-specific annotation guidelines[3] and pretrained parsers, whose performance on normative language is satisfactory at least on the standard test sets.

When it comes to the L2 half of the corpus, on the other hand, there are two difficulties: 1. that standard parser are expected to perform worse, resulting in a more time-consuming and labour-intensive validation process, and 2. that consistently dealing with grammatical errors will require the development of additional guidelines. To mitigate the first problem, we plan on using gold-annotated corrections to fine-tune a domain-specific parser.[4] When it comes to the development of L2 Swedish guidelines, on the other hand, much can be learned from previous L2 annotation efforts.[5] The main shared principle is that of *literal reading* (syntactic analysis is, as much as possible, based on observed language usage and surface features rather than on the author's assumed intended meaning). This general principle and its exceptions, however, need be declined differently for different languages.

4.1 SweLL-UD at the 1st UniDive training school

While the author has already tutored two cohorts of students in basic UD and has sufficient confidence to train less experienced annotators for step 2 of the annotation project, exchanging ideas with more experienced treebank developers would be invaluable for this crucial guideline development phase. Furthermore, the training school would provide an opportunity for the author, who has never been part of any large-scale annotation effort, to further familiarize herself with the corpus annotation infrastructure and the treebank release process. The plan is to come to the training school with a fully preprocessed corpus and having already carried out some exploratory annotation experiments with part of the annotation team, so as to have in mind which L2 phenomena require special attention and what practical problems we encounter in terms of tooling and annotation setup.

[3] `universaldependencies.org/sv`

[4] This is based on preliminary experiments conducted by the author on other languages, to appear at the 17th Workshop on Building and Using Comparable Corpora (BUCC) under the title *Bootstrapping the Annotation of UD Learner Treebank*.

[5] The Italian L1-L2 treebank, for instance, has comprehenive annotation guidelines available at `github.com/ElisaDiNuovo/VALICO-UD_guidelines`

References

[1] Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. Universal Dependencies for learner English. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany, August 2016. Association for Computational Linguistics.

[2] Elisa Di Nuovo, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino, and Cristina Bosco. VALICO-UD: Treebanking an Italian learner corpus in Universal Dependencies. *IJCoL. Italian Journal of Computational Linguistics*, 8(8-1), 2022.

[3] Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. A dependency treebank of spoken second language English. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45, Seattle, Washington, July 2022. Association for Computational Linguistics.

[4] John Lee, Herman Leung, and Keying Li. Towards Universal Dependencies for learner Chinese. In Marie-Catherine de Marneffe, Joakim Nivre, and Sebastian Schuster, editors, *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71, Gothenburg, Sweden, May 2017. Association for Computational Linguistics.

[5] John Lee, Keying Li, and Herman Leung. L1-L2 parallel dependency treebank as learner corpus. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 44–49, Pisa, Italy, September 2017. Association for Computational Linguistics.

[6] Arianna Masciolini. A query engine for L1-L2 parallel dependency treebanks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 574–587, Tórshavn, Faroe Islands, may 2023. University of Tartu Library.

[7] Arianna Masciolini and Márton A Tóth. STUnD: ett Sökverktyg för Tvåspråkiga Universal Dependencies-trädbanker. In *Proceedings of the Huminfra Conference*, pages 95–109, Gothenburg, Sweden, 2024.

[8] Arianna Masciolini, Elena Volodina, and Dana Dannélls. 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)*, pages 585–597, Toronto, Canada, jul 2023. Association for Computational Linguistics.

[9] Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, et al. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104, 2019.