SweLL-UD: A Treebank of L2 Swedish Essays

1st UniDive Training School, 8-12 July 2024, Technical University of Moldova

Arianna Masciolini

Språkbanken Text, Department of Swedish, Multilingualism, Language Technology, University of Gothenburg, Sweden

UD Treebanks in SLA Research

Benefits of UD treebanks in corpus-based Second Language Acquisition research:

- uniform morphosyntactic annotation layer allowing for **cross-lingual comparisons**:
  - between standard and learner language
  - between a learner's L1 and L2
  - between different L2s
- possibility to carry out **grammatical error analysis** without explicit error tagging e.g. if learner sentences are paired with corrections (*parallel L1-L2 treebank*)
- **semi-automatic annotation** through existing parsers

Existing L2 UD Treebanks

language	name	# sentences	status	parallel
Chinese	CFL	451	released	no
English	ESL	5124	retired*	yes
Italian	Valico	398	released	yes
Korean	KSL	7530	released*	no
Russian	?	500	in progress	yes
Swedish	SweLL	~5000	in progress	yes

* available through GitHub but not part of the latest UD release

The SweLL-UD treebank

Objectives

- ultimate goal: building a **training-scale treebank** of L2 Swedish
- within the next few months: releasing a high-quality **500-sentence test set**
- at the training school: discussing **L2-specific guidelines** and **tooling**

Data

Source corpus: **SweLL-gold** (Swedish Learner Language corpus)

- **text type**: short essays
- **learners**: adult L2 Swedish students with various language backgrounds and proficiency levels
- **annotation**: manual correction + error tagging + pseudonymization
- **size**: 502 essays, ~5000 parallel sentences
- **license**: CLARIN-ID -PRIV -NORED -BY (but the data can be and has been redistributed as long as it is not possible to reconstruct full essays)

Project status and plan

1. **preprocessing**: sentence pair and metadata extraction, conversion to CoNNL-U, shuffling and creation of a train, development and test split ✓
2. **preliminary annotation experiments** with a small annotation team ✓
3. **L1 (corrections) annotation**:
   (a) automatic pre-annotation with an off-the-shelf parser ✓
   (b) manual validation with a larger annotation team (hopefully mixed L1 and L2 speakers, students and researchers)
4. **L2 (originals) annotation**:
   (a) automatic pre-annotation with parser fine-tuned on the manually validated corrections ↑
   (b) manual validation with the same annotation team

TODO: example tree pair from SweLL, glossed

Acknowledgement

TODO: financing for both annotation and summer school, other participants TODO: nationella Språkbanken logo

Funded by the European Union