SweLL-UD: A Treebank of L2 Swedish Essays

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

Arianna Masciolini with Maria Irena Szawerna and Elena Volodina

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

UD Treebanks in SLA Research

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

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

Existing Second Language 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

- end 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: further experimenting with **annotation of L2 essays** and discussing **open questions**

Data

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

- **genre**: essays (various topics)
- **learners**: adult L2 Swedish students with various language backgrounds and proficiency levels
- **annotation**: manual correction, error tagging and pseudonymization
- **size**: 502 essays (5000+ parallel sentences)
- **license**: CLARIN-ID -PRIV -NORED -BY (but the data can and has been redistributed as long as it is impossible to reconstruct full essays)

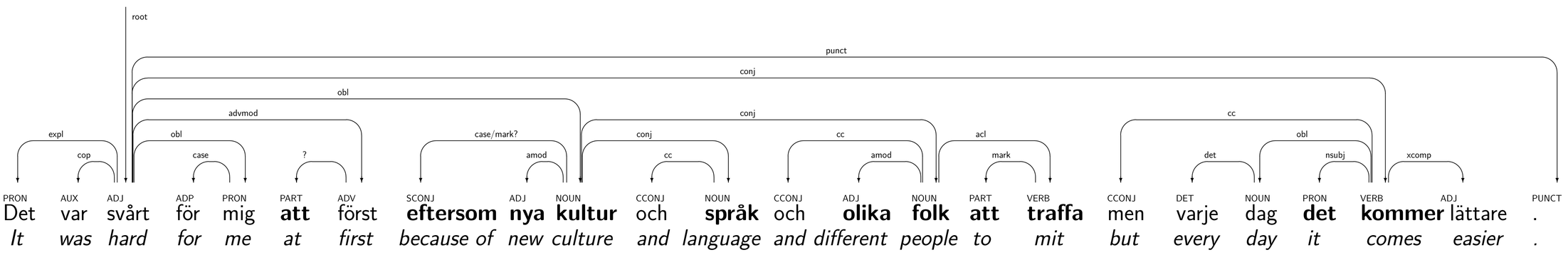
Project status and plan

1. **preprocessing**: sentence pair and metadata extraction, CoNNL-U format conversion, shuffling and creation of train, development and test splits ✓
2. **preliminary annotation experiments** with a small annotation team ✓
3. **L1 (corrections) annotation**:
   - automatic pre-annotation with a pretrained UDPipe model ✓
   - manual validation with a larger annotation team including both L1 and L2 Swedish speakers
4. **L2 (originals) annotation**:
   - automatic pre-annotation with a custom (fine-tuned) model
   - manual validation with the larger annotation team

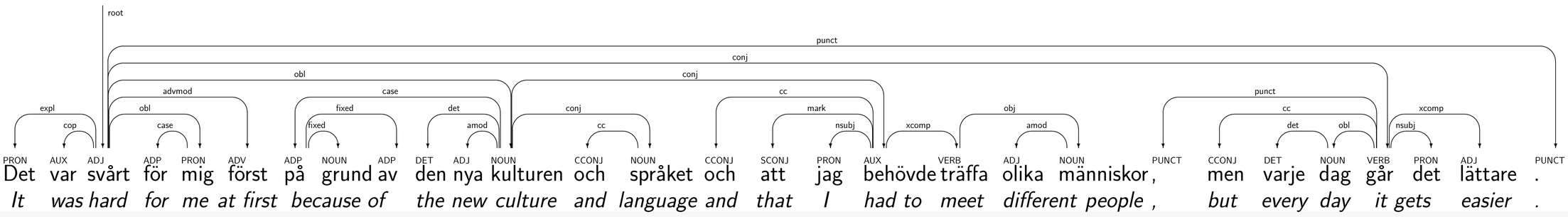
Open questions

- **tooling** - desiderata:
  - tree visualization
  - no installation required (as with Eevee)
  - parallel L1-L2 editing (or at least viewing, as with STUnD)
- **L2-specific guidelines**:
  - when annotating an original learner sentence, to what extent should the syntactic annotation of its correction be taken into account?
  - what is a good balance between the general principle of *literal reading* and keeping the two trees as structurally similar as possible?
  - should the existing L1-specific guidelines be applied when annotating constructionsn that the learners borrow from their L1?

Original sentence from a learner essay (errors highlighted in bold)



Corresponding correction hypothesis



Acknowledgement

Participation to the training school is funded by UniDive, while the annotation project is supported by the Swedish National Research Infrastructure Nationella Språkbanken, funded jointly by the Swedish Research Council (2018–2024, contact 2017-00626) and the ten participating partner institutions.