Universal Dependencies Meets Second Language Acquisition: the Case of Swedish

SPRÅKBANKENTEXT
A research infrastructure for language data and a language technology research unit

UNIVERSITY OF GOTHENBURG

Arianna Masciolini – Språkbanken Text, University of Gothenburg, Sweden

Why use UD for L2 corpora?

- existing parsers allow for **faster, semi-automatic annotation**
- rich morphosyntactic UD annotation supports the **study of L2 grammatical patterns**
- UD provides a uniform annotation layer that enables **cross-lingual comparisons**:
  - between a learner's L1 and L2
  - between different L2s
  - between standard and learner language

L1-L2 Treebanks

Parallel treebanks where **learner sentences** are **paired with *correction hypotheses***.

L1-L2 treebanks can be an even better basis for:

- **grammatical error retrieval and analysis**
- automatic **feedback comment generation**

UD Treebanks of Second Language

language	name	sentences	status	parallel
Chinese	CFL	451	released	no
English	ESL	5124	retired*	**yes**
English	ESLSpok	2320	released	no
Italian	Valico	398	released	**yes**
Korean	KSL	7530	released*	no
Russian	?	500	in progress	**yes**
Swedish	SweLL	~5000	in progress	**yes**

* available for download but not part of the latest UD release

Overall Project Goals

- improving **UD guidelines for L2** (Swedish) **treebanks**
- creating an **L1-L2 Swedish treebank**
- training **parsing models for L2 material**
- developing **tools for parallel (L1-L2) treebanks**

SweLL-UD: a Treebank of L2 Swedish

Source Corpus

SweLL-gold, aka the Swedish Learner Language corpus*:

- **genre**: essays (miscellaneous topics)
- **learners**: adult L2 Swedish students with various language backgrounds and proficiency levels
- **annotation**: manual correction, error tagging, pseudonymization and normalization (minimal edits)
- **size**: 502 essays (→ 5000+ parallel sentences)
- **license**: CLARIN-ID -PRIV -NORED -BY
  (but the data can be redistributed as long as some metadata is removed and full essays cannot be reconstructed)

* part of the Manually Annotated Corpora CLARIN Resource Family

Project Status and Plan

1. **preprocessing**:
   - sentence pair extraction (✓)
   - automatic pre-annotation (✓)
2. **manual validation** of a 500-sentence test set || **guidelines development** (ongoing)
3. **test set** release (planned in 2025) - 2 versions:
   - sentence-shuffled version at universaldependencies.org
   - full-essay version with all metadata released as the source corpus
4. gradual annotation and release of a **development and training set**

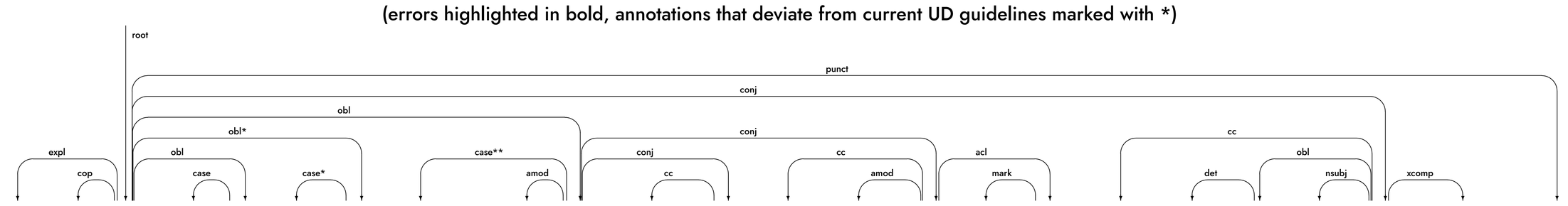
Example

Metadata

proficiency level: beginner; **L1**: Tagalog; **best writing language**: English...

Original Sentence

(errors highlighted in bold, annotations that deviate from current UD guidelines marked with *)



Existing Universal + Swedish-specific guidelines cover most ungrammatical fragments of this sentence, but:

- * : foreign construction, annotated borrowing from English guidelines
- ** : mismatch between POS and DEPREL (intentional – we annotate literally at the token level and follow distributional criteria at the syntactic level)

Correction Hypothesis