Treebanking SweLL

the Swedish Learner Language Corpus

Arianna Masciolini

Språkbanken Text

Department of Swedish, Multilingualism, Language Technology

22 October 2025

Introducing SweLL

(aka the Swedish Learner Language corpus)

- **genre**: essays (misc topics)
- **learners**: adult L2 Swedish learners with various language backgrounds and proficiency levels
- **annotation**: error tagging, pseudonymization and normalization (minimal edits)

A look at the data

Swedish (SweLL)

```
<sentence> <w ref="1">"</w> <w ref="2" target_form="Det"
correction_label="L-Ref">Den</w> <w ref="3">är</w>
<w ref="4">en</w> <w ref="5">tredjedel</w>
<w ref="6">av</w> <w ref="7">din</w> <w ref="8">dag</w>
<w ref="9">!</w> </sentence>
```

A look at the data

Swedish (SweLL)

```
<sentence> <w ref="1">"</w> <w ref="2" target_form="Det"
correction_label="L-Ref">Den</w> <w ref="3">är</w>
<w ref="4">en</w> <w ref="5">tredjedel</w>
<w ref="6">av</w> <w ref="7">din</w> <w ref="8">dag</w>
<w ref="9">!</w> </sentence>
```

English (FCE)

```
I also suggest that more plays and films should
<ns type="RV"> <ns type="FV"><i>be taken</i><c>take</c>
</ns> place</ns>.
```

Italian (VALICO)

```
Finse <MC><i>aveva paura</i><c>che aveva paura</c>
</MC> di un <DN><i>rapito</i><c>rapimento</c></DN>.
```

The problems

- lack of interoperability between corpora
- lots of manual annotation needed
- coarse-grained error labels
- exclusive focus on errors

The solution: UD

Universal Dependencies

- a **cross-lingually consistent grammatical annotation scheme**, designed to be
  - human- *and* machine-readable
  - suitabile for both mono- *and* multilingual use cases
- a growing multilingual collection of dependency treebanks (160+ languages and 600+ contributors!)

The solution: UD

- adopting a **shared data format** grants basic interoperability between corpora
- **parsers** help boostrapping the annotation process
- **fine-grained morphosyntactic annotation** allows moving beyond error detection/tagging
- **cross-linguistic consistency*** enables comparisons between:
    - L1 and L2
    - different L2s
    - L2 and TL[1]

[1] especially with **parallel learner treebanks**

UD treebanks of learner language

language	name	sentences	status
Chinese	CFL	451	released
English	ESL	5124	retired
English	ESLSpok	2320	released
Greek	GLCII		in progress
Italian	Valico	398	released
Korean	KSL	12977	released
Russian		500	in progress
Swedish	**SweLL**	~**5000**	**in progress**

UD guidelines do ***not*** cover all relevant interlanguage phenomena and are ***not*** universally adopted across learner treebanks

Learn more:

Arianna Masciolini, Aleksandrs Berdicevskis, Maria Irena Szawerna, and Elena Volodina. *Annotating second language in Universal Dependencies: a review of current practices and directions for harmonized guidelines* (2025)

TL;DR - literal vs. “distributional”

en långing **bus** **resa**

Literal annotation

NUM/DET/...	ADJ	NOUN	NOUN/VERB
en	lång	**bus**	**resa**

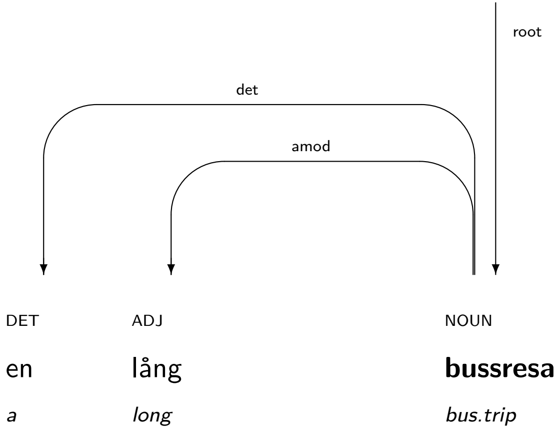
Literal annotation

DET	ADJ	NOUN	NOUN
en	lång	**bus**	**resa**
a	*long*	*?*	*trip*

Correction-aware annotation

DET	ADJ	NOUN
en	lång	**bussresa**
a	*long*	*bus.trip*

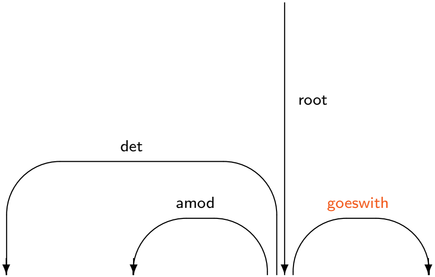
Correction-aware annotation



Correction-aware annotation

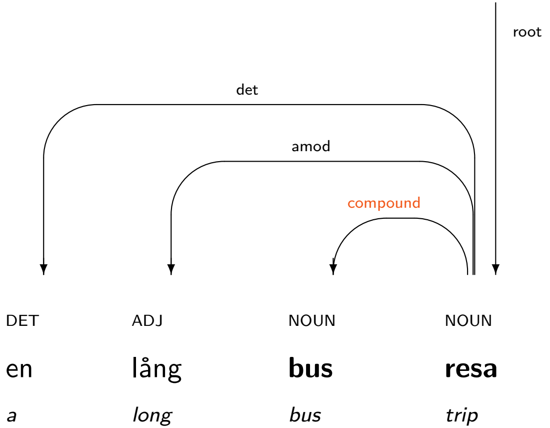
DET	ADJ	NOUN	NOUN
en	lång	**bus**	**resa**
a	*long*	*bus*	*trip*

Correction-aware annotation



DET	ADJ	NOUN	NOUN
en	lång	**bus**	**resa**
a	*long*	*bus*	*trip*

Correction-aware annotation



root

det

amod

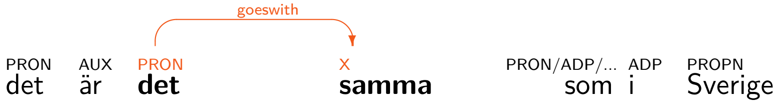
compound

DET	ADJ	NOUN	NOUN
en	lång	**bus**	**resa**
a	*long*	*bus*	*trip*

Transfer-aware annotation

PRON	AUX	PRON/DET	ADJ	PRON/ADP/...	ADP	PROPN
det	är	**det**	**samma**	som	i	Sverige

Transfer-aware annotation



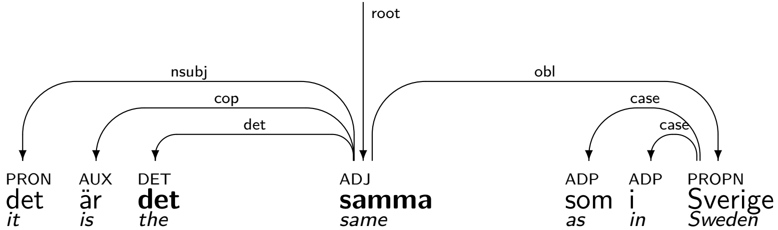
goeswith

PRON	AUX	PRON	X	PRON/ADP/...	ADP	PROPN
det	är	**det**	**samma**	som	i	Sverige

Transfer-aware annotation

PRON	AUX	DET	ADJ	ADP	ADP	PROPN
det	är	**det**	**samma**	som	i	Sverige
it	*is*	*the*	*same*	*as*	*in*	*Sweden*

Transfer-aware annotation



Annotators

- Sasha (L1: Russian)
- Maria (L1: Polish)
- Arianna (L1: Italian)

And soon:

- Caroline (L1: French)

Where?

`github.com/UniversalDependencies/UD_Swedish-SweLL`[2]

[2] just not yet!

When?

- test set (500 sentences): first release hopefully on November 15 as part of UD 2.17
- dev set (another 500 sentences): as part of UD 2.18
- train set (4000 sentences!): we'll see about that

Tackar *all* som hjälpte oss!

SPRÅKBANKENTEXT

UniDive