

Treebanking SweLL

the Swedish Learner Language Corpus

Arianna Masciolini

Språkbanken Text

Department of Swedish, Multilingualism, Language Technology

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>.

See 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
 - ❖ suitable 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 bootstrapping 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 TL¹

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

lång

bus

resa

Literal annotation



NUM/DET/... ADJ

en

NOUN

lång

NOUN/VERB

resa

Literal annotation



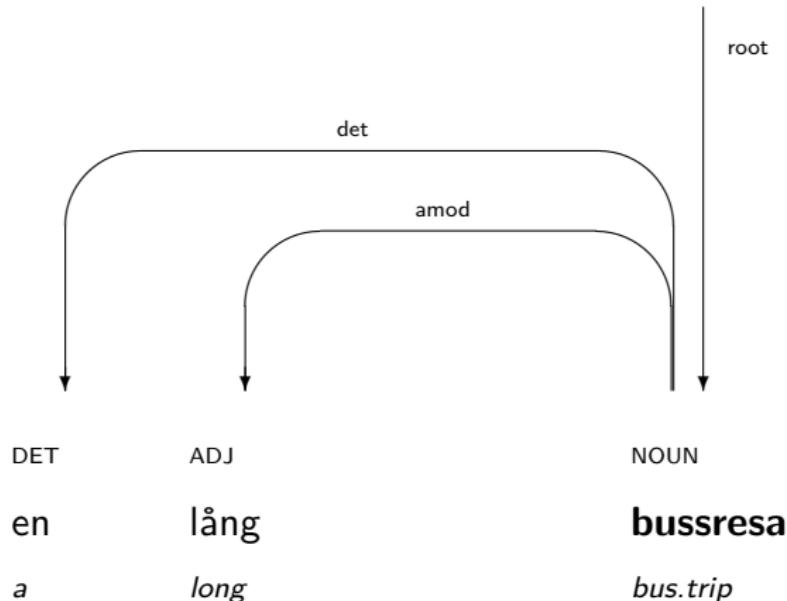
DET	ADJ	NOUN	NOUN
en	lång	bus	resa
a	<i>long</i>	?	<i>trip</i>

Correction-aware annotation



DET	ADJ	NOUN
en	lång	bussresa
a	<i>long</i>	<i>bus.trip</i>

Correction-aware annotation

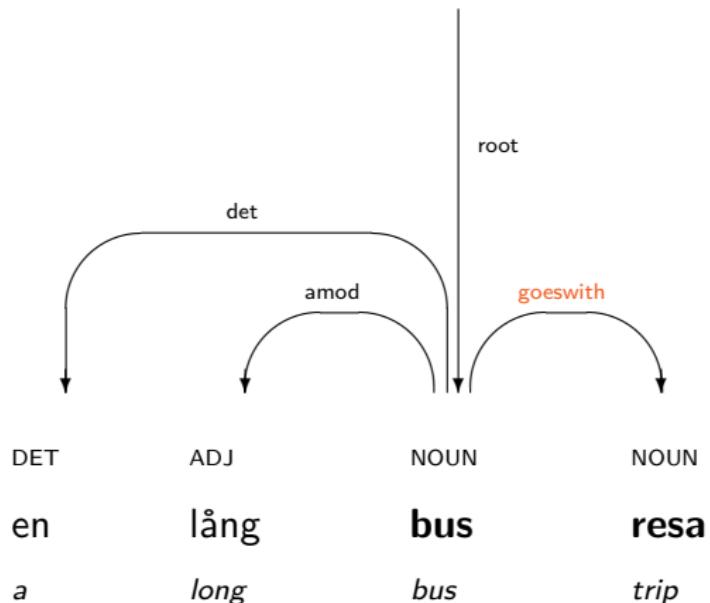


Correction-aware annotation

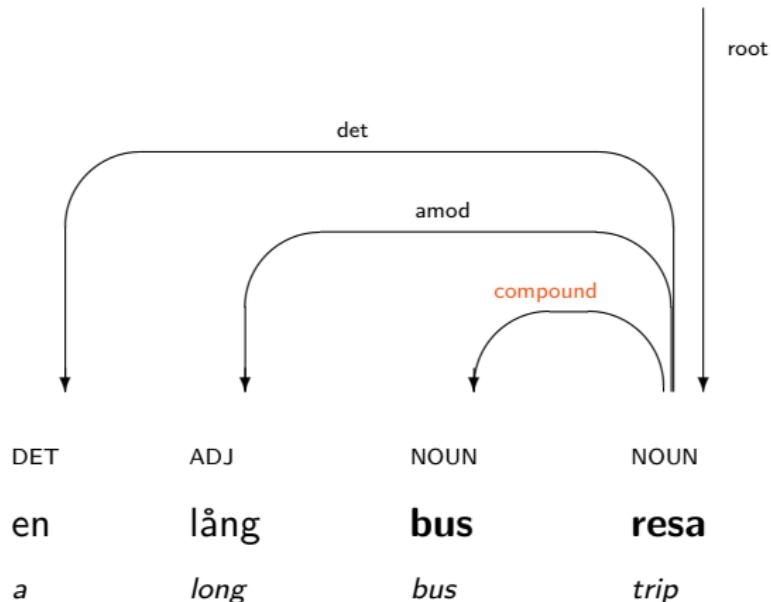


DET	ADJ	NOUN	NOUN
en	lång	bus	resa
a	<i>long</i>	<i>bus</i>	<i>trip</i>

Correction-aware annotation



Correction-aware annotation



Transfer-aware annotation



PRON AUX PRON/DET
det är **det**

ADJ
samma

PRON/ADP/... ADP
som i

PROPN
Sverige

Transfer-aware annotation



Transfer-aware annotation

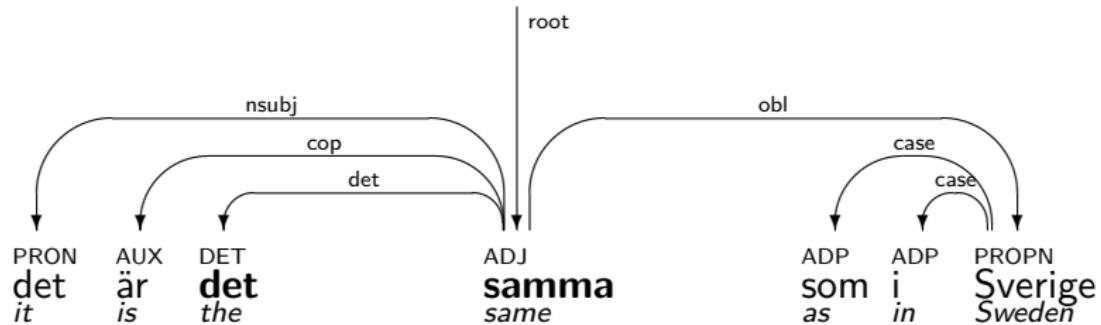


PRON AUX DET
det är **det**
it *is* *the*

ADJ
samma
same

ADP ADP PROPN
som i Sverige
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-SweLL²](https://github.com/UniversalDependencies/UD_Swedish-SweLL)

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

