Universal Dependencies

as an annotation standard for L2 corpora

Arianna Masciolini
Språkbanken Text (University of Gothenburg)

4 settembre 2025

Learner data

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>.
```

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>
```

The problems

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

The solution: UD

- 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**

Universal Dependencies 101

UD annotation in 3 steps

1. **segmentation** (sentences, then words)
2. **word-level annotation** (lemmas, POS tags, morphological features)
3. **syntactic annotation** (dependency relations)

Example

[...] Det bästa i Sverige är naturen. Jag älskar naturen så mycket. Nu har jag vant mig vid att bo i Sverige efter 9 månader.

Step 1: segmentation

Det	bästa	i	Sverige	är	naturen	.
the	*best*	*in*	*Sweden*	*is*	*the.nature*	*.*

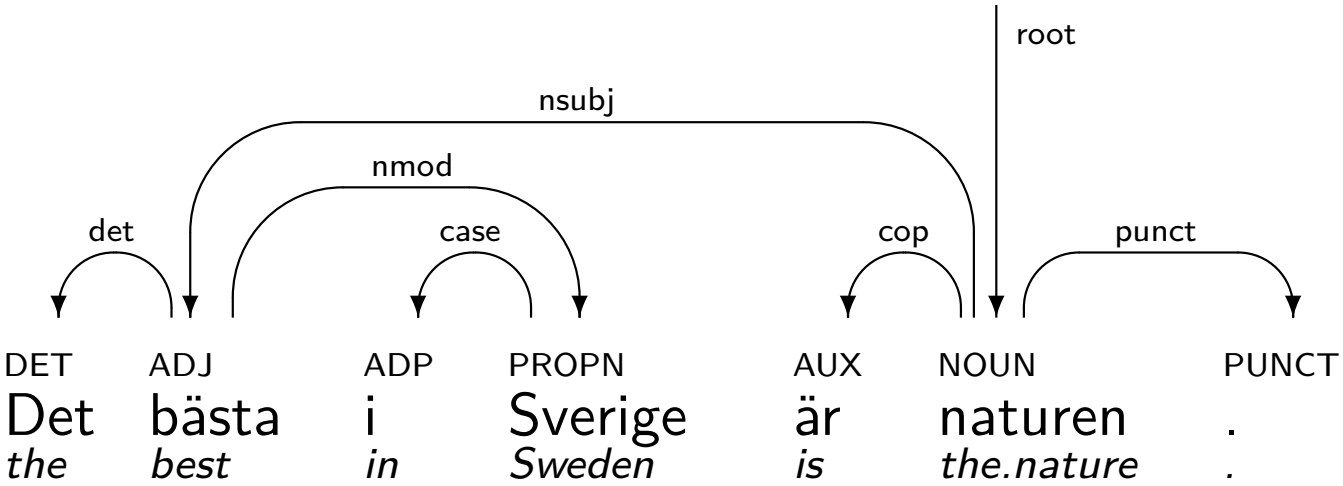
“The best thing in Sweden is nature.”

Step 2: POS tagging

DET	ADJ	ADP	PROPN	AUX	NOUN	PUNCT
Det	bästa	i	Sverige	är	naturen	.
the	*best*	*in*	*Sweden*	*is*	*the.nature*	*.*

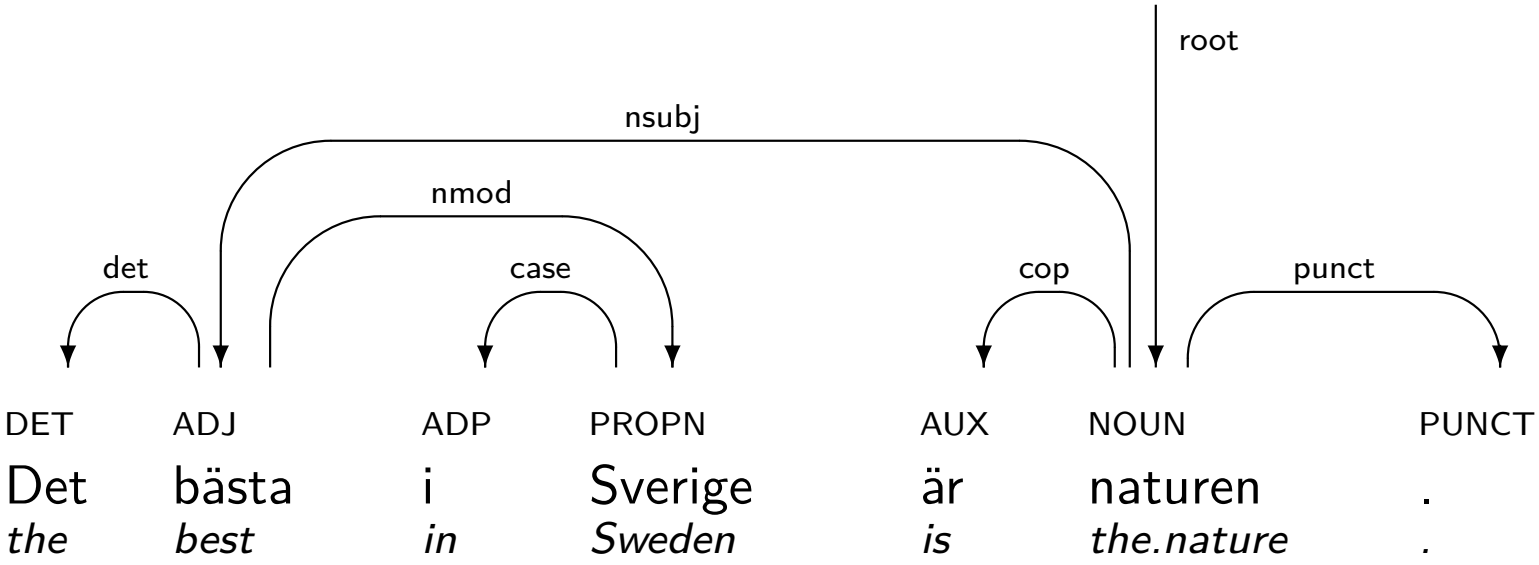
"The best thing in Sweden is nature."

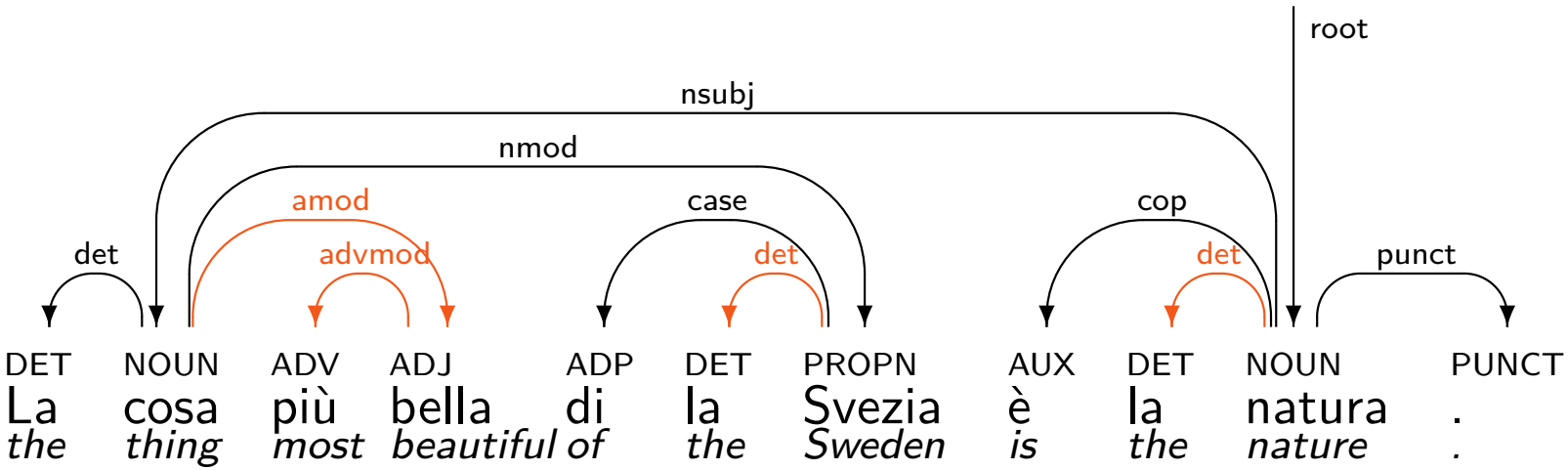
Step 3: syntax



"The best thing in Sweden is nature."

The U in UD


root
nsubj
nmod
det
case
cop
punct
DET ADJ ADP PROPN AUX NOUN PUNCT
Det bästa i Sverige är naturen .
the best in Sweden is the.nature .


root
nsubj
nmod
amod
case
cop
det
advmod
det
det
punct
DET NOUN ADV ADJ ADP DET PROPN AUX DET NOUN PUNCT
La cosa più bella di la Svezia è la natura .
the thing most beautiful of the Sweden is the nature .

The CoNNL-U format

```
# text = Det bästa i Sverige är naturen.
# text_en = The best thing in Sweden is nature.
1   Det den DET SG-DEF  Definite=Def|Gender=Neut|...    2   det _   _
2   bästa   bra ADJ SPL-DEF Case=Nom|Definite=Def|...   6   nsubj   _   _
3   i   i   ADP _   _   4   case    _   _
4   Sverige Sverige PROPN   SG-NOM  Case=Nom    2   nmod    _   _
5   är  vara    AUX PRES-ACT    Mood=Ind|Tense=Pres|... 6   cop _   _
6   naturen natur   NOUN    SG-DEF-NOM  Case=Nom|Definite=Def|...   0   root
7   .   .   PUNCT   Period  _   6   punct   _   _
```

Table view

ID	FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL
1	Det	den	DET	Definite=Def Gender=Neut…	2	det
2	bästa	bra	ADJ	Case=Nom Definite=Def…	6	nsubj
3	i	i	ADP	_	4	case
4	Sverige	Sverige	PROPN	Case=Nom	2	nmod
5	är	vara	AUX	Mood=Ind Tense=Pres…	6	cop
6	naturen	natur	NOUN	Case=Nom Definite=Def…	0	root
7	.	.	PUNCT	_	6	punct

UD for L2 corpora

L1-L2 treebanks

L1-L2 Parallel Dependency Treebank as Learner Corpus

John Lee, Keying Li, Herman Leung
Department of Linguistics and Translation
City University of Hong Kong
jsylee@cityu.edu.hk, keyingli3-c@my.cityu.edu.hk, leung.hm@gmail.com

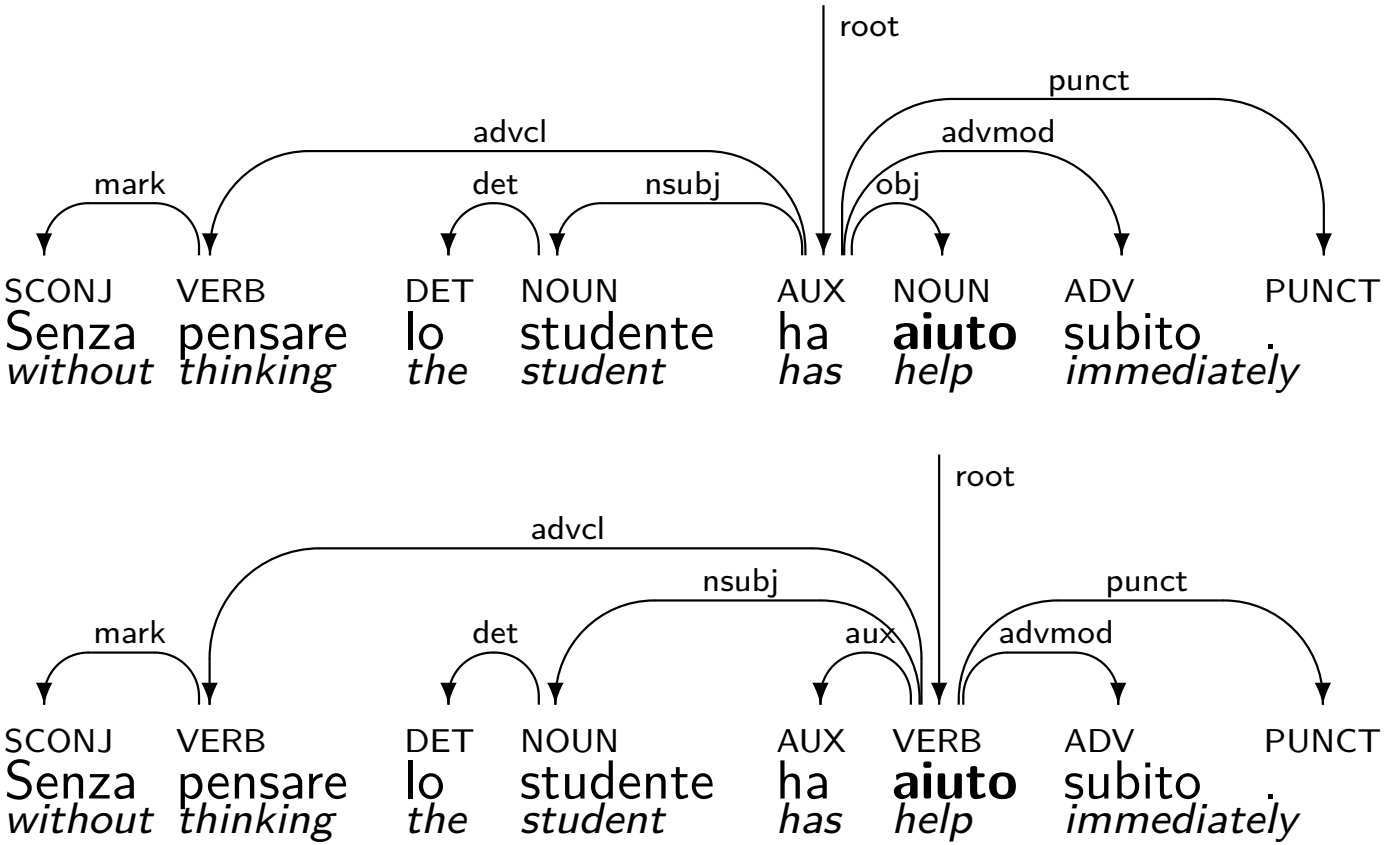
- learner productions // correction hypotheses
- aka *parallel learner treebanks*

UD treebanks of learner language

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

- UD guidelines:
  - do ***not*** cover all interlanguage phenomena
  - are ***not*** universally adopted across learner treebanks
- grammatical errors are ***not*** treated uniformly across treebanks

Literal vs. distributional



What is literal annotation?

en lång **bus** **resa**

What is literal annotation?

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

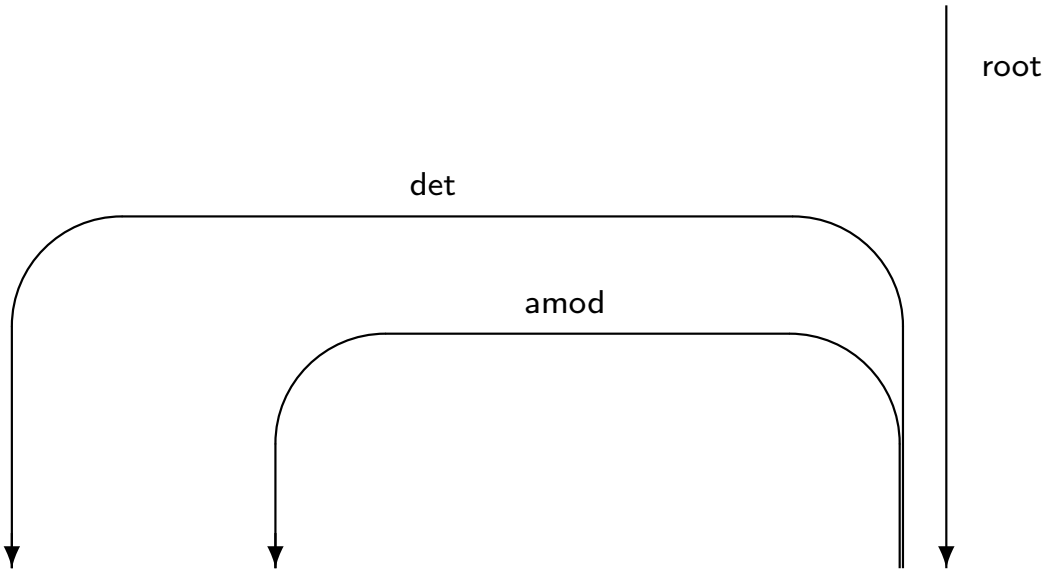
What is literal annotation?

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

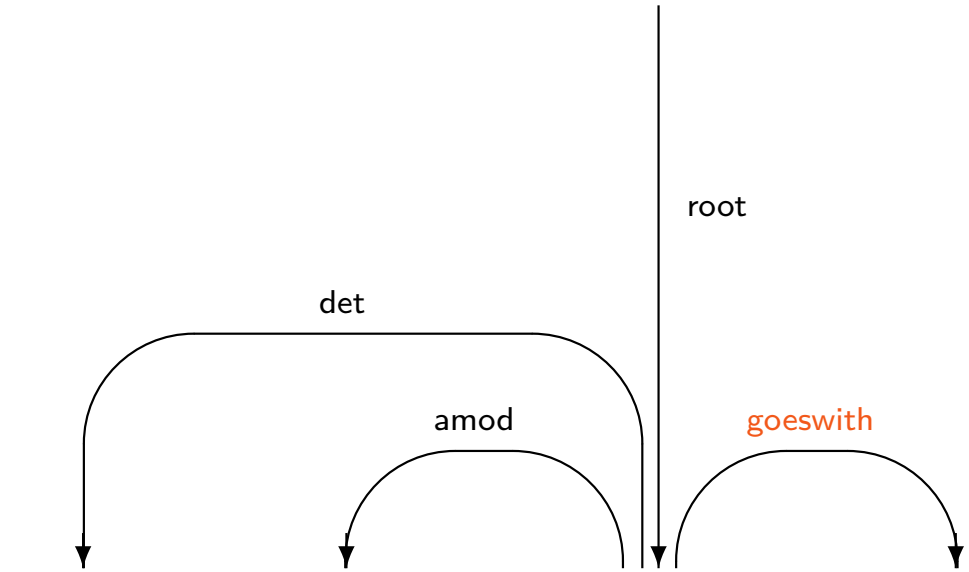


DET	ADJ	NOUN
en	lång	**bussresa**
a	*long*	*bus.trip*

Correction-aware annotation

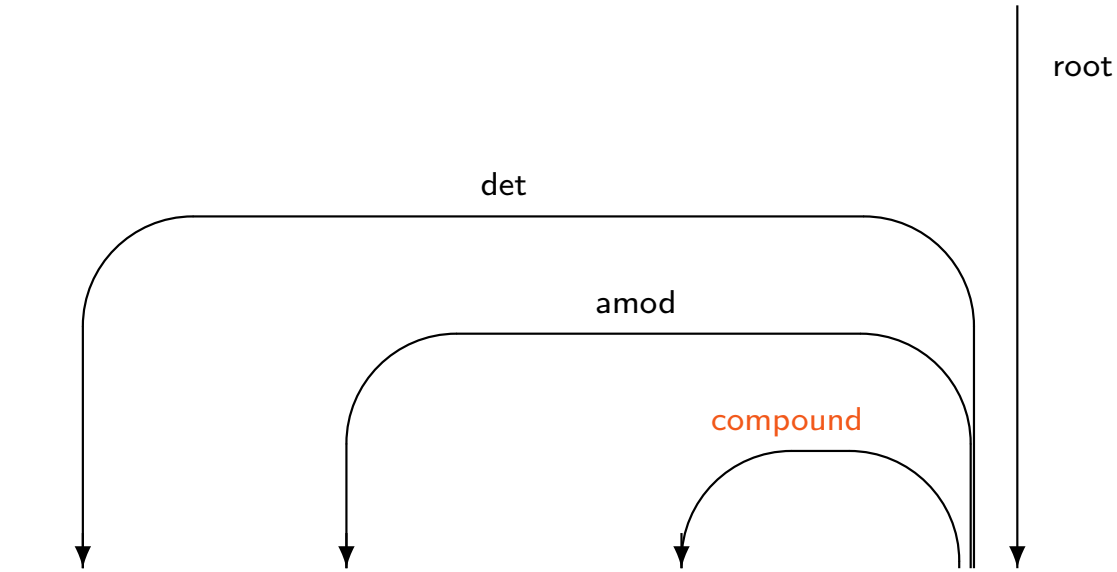
DET	ADJ	NOUN	NOUN
en	lång	**bus**	**resa**
a	*long*	*bus*	*trip*

Correction-aware annotation



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

Correction-aware annotation

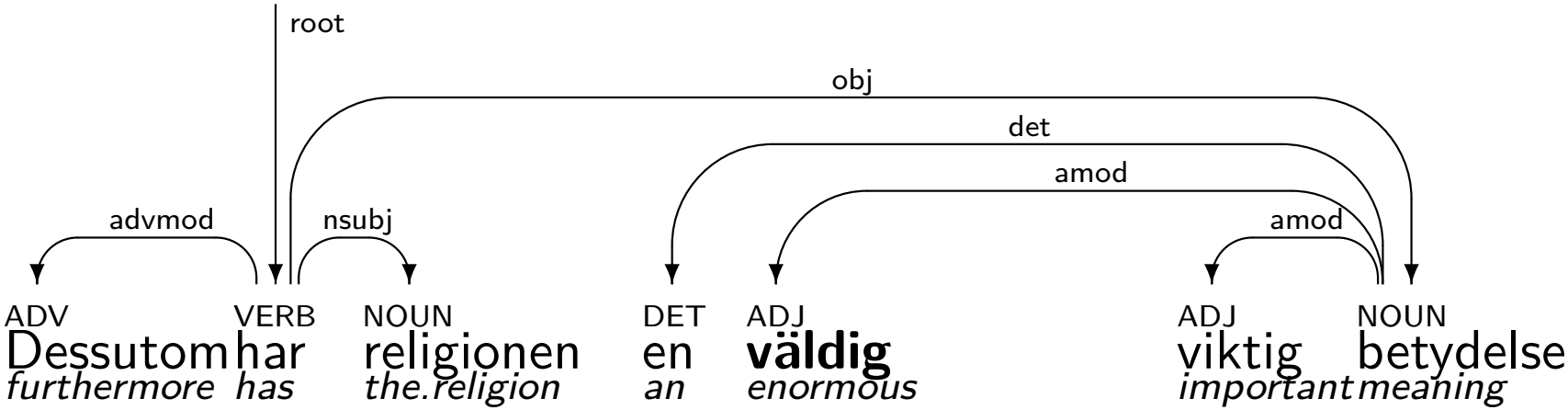


DET	ADJ	NOUN	NOUN
en	lång	**bus**	**resa**
a	*long*	*bus*	*trip*

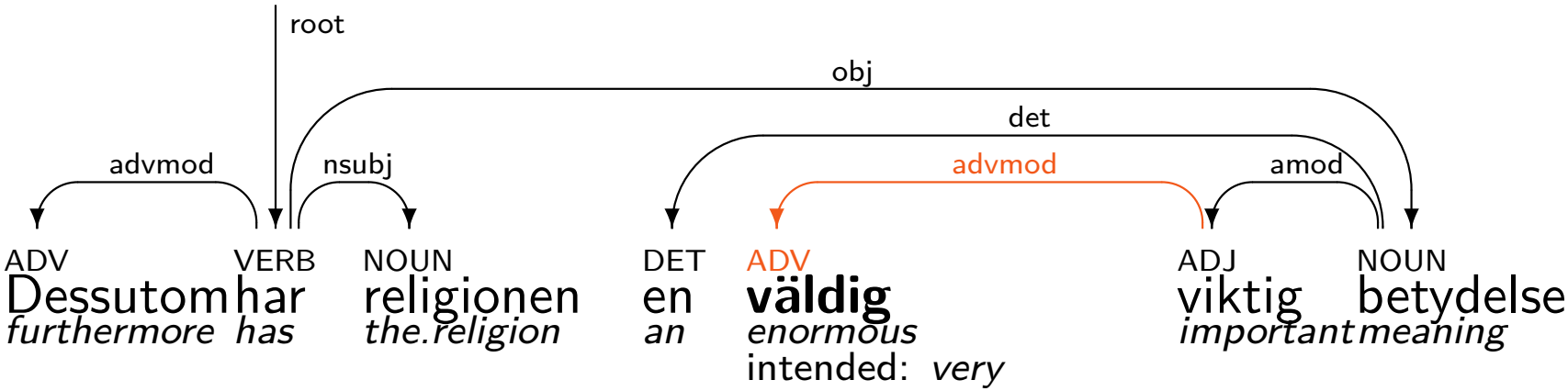
Informativeness

ADV	VERB	NOUN	DET	ADJ		ADJ	NOUN
Dessutom	har	religionen	en	**väldig**		viktig	betydelse
furthermore	*has*	*the.religion*	*an*	*enormous*		*important*	*meaning*

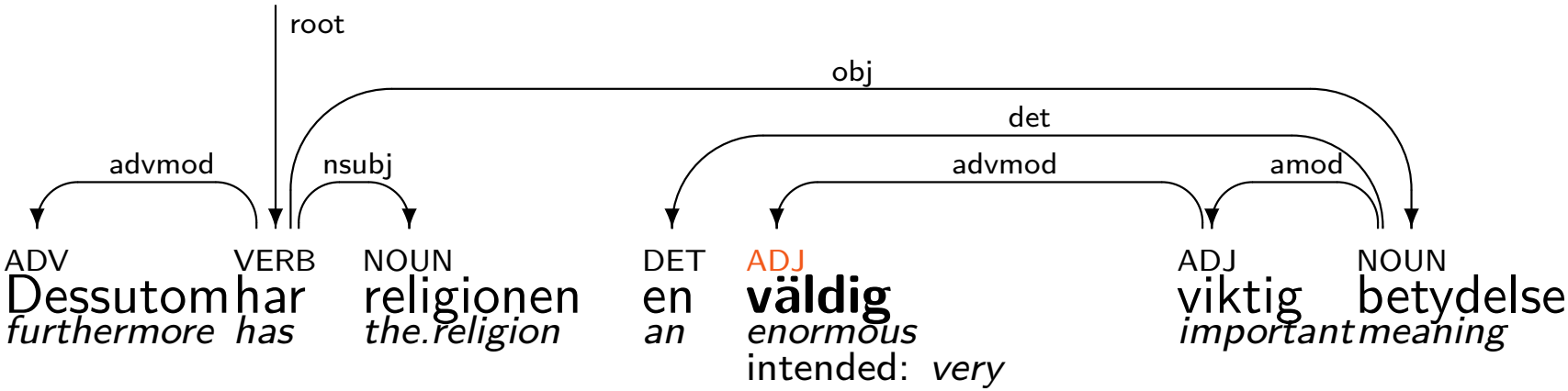
Informativeness



Informativeness



Informativeness

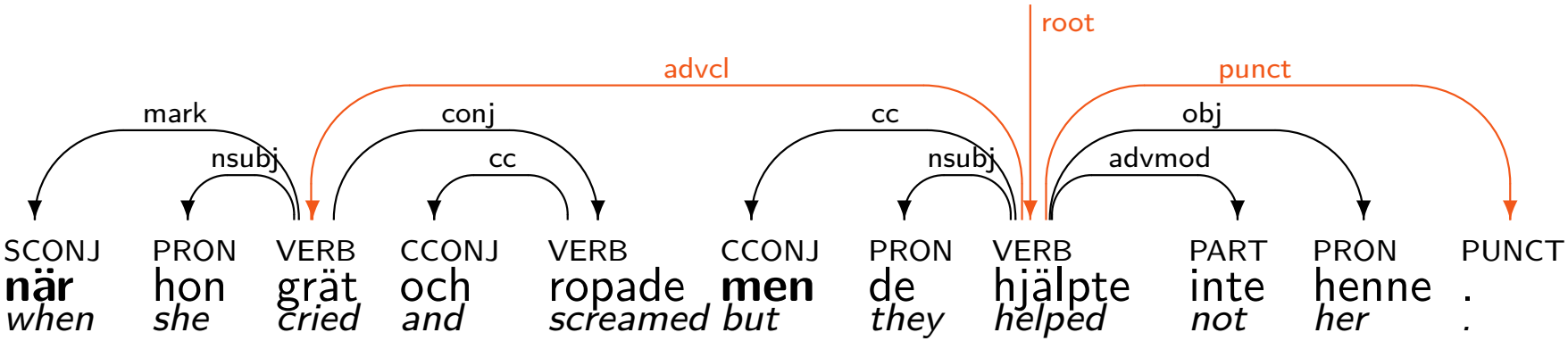


Consistency

SCONJ	PRON	VERB	CCONJ	VERB	CCONJ	PRON	VERB	PART	PRON	PUNCT
när	hon	grät	och	ropade	**men**	de	hjälpte	inte	henne	.
when	*she*	*cried*	*and*	*screamed*	*but*	*they*	*helped*	*not*	*her*	*.*

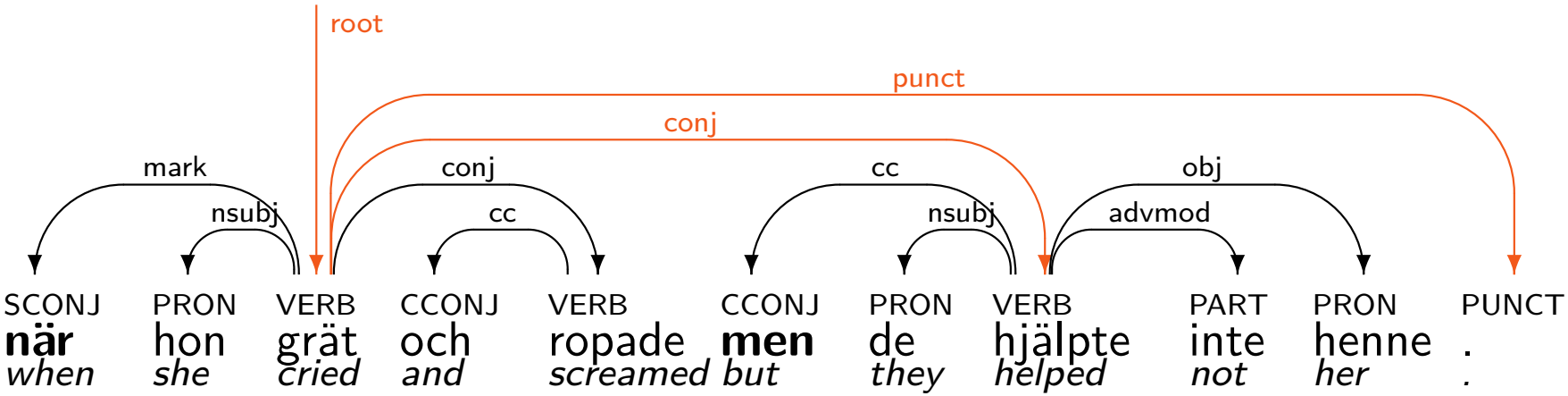
- när hon grät och ropade ~~men~~ hjälpte de inte henne.
  (*when she cried and screamed they did not help her.*)
- ~~när~~ hon grät och ropade, men de hjälpte inte henne.
  (*she cried and screamed, but they did not help her.*)

Consistency



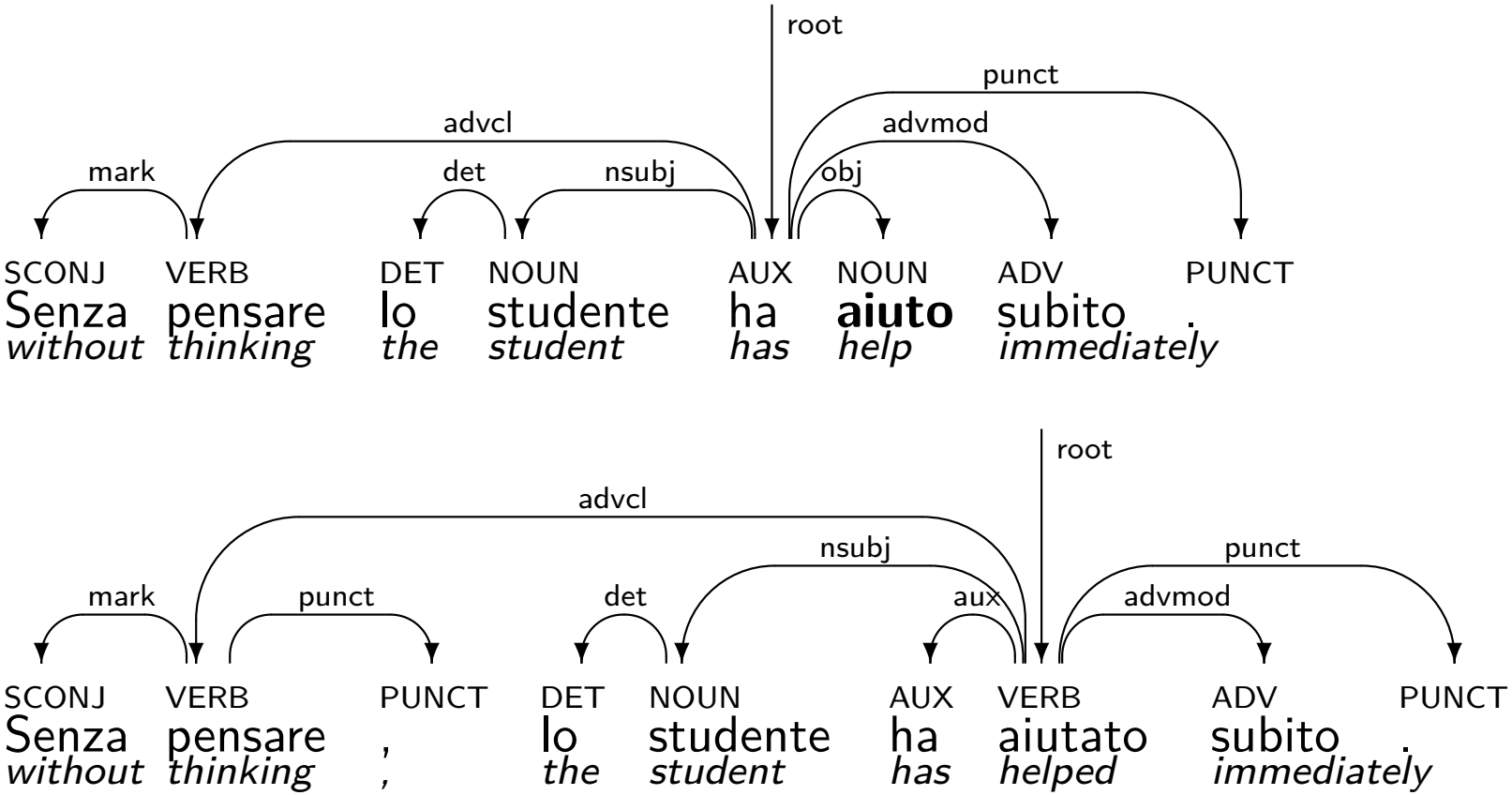
- när hon grät och ropade ~~men~~ hjälpte de inte henne.
  (*when she cried and screamed they did not help her.*)
- ~~när~~ hon grät och ropade, men de hjälpte inte henne.
  (*she cried and screamed, but they did not help her.*)

Consistency

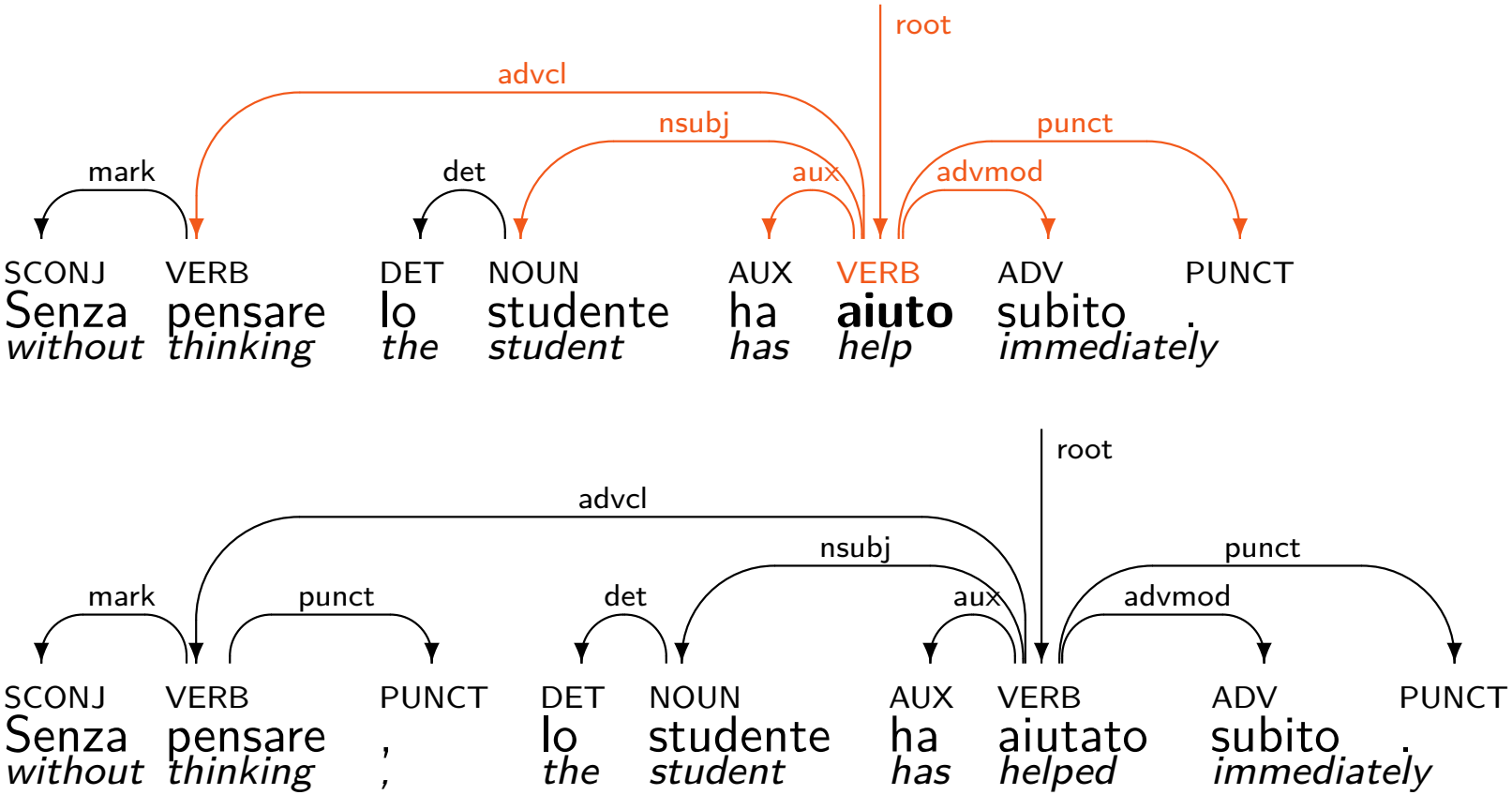


- när hon grät och ropade ~~men~~ hjälpte de inte henne.
  (*when she cried and screamed they did not help her.*)
- ~~när~~ hon grät och ropade, men de hjälpte inte henne.
  (*she cried and screamed, but they did not help her.*)

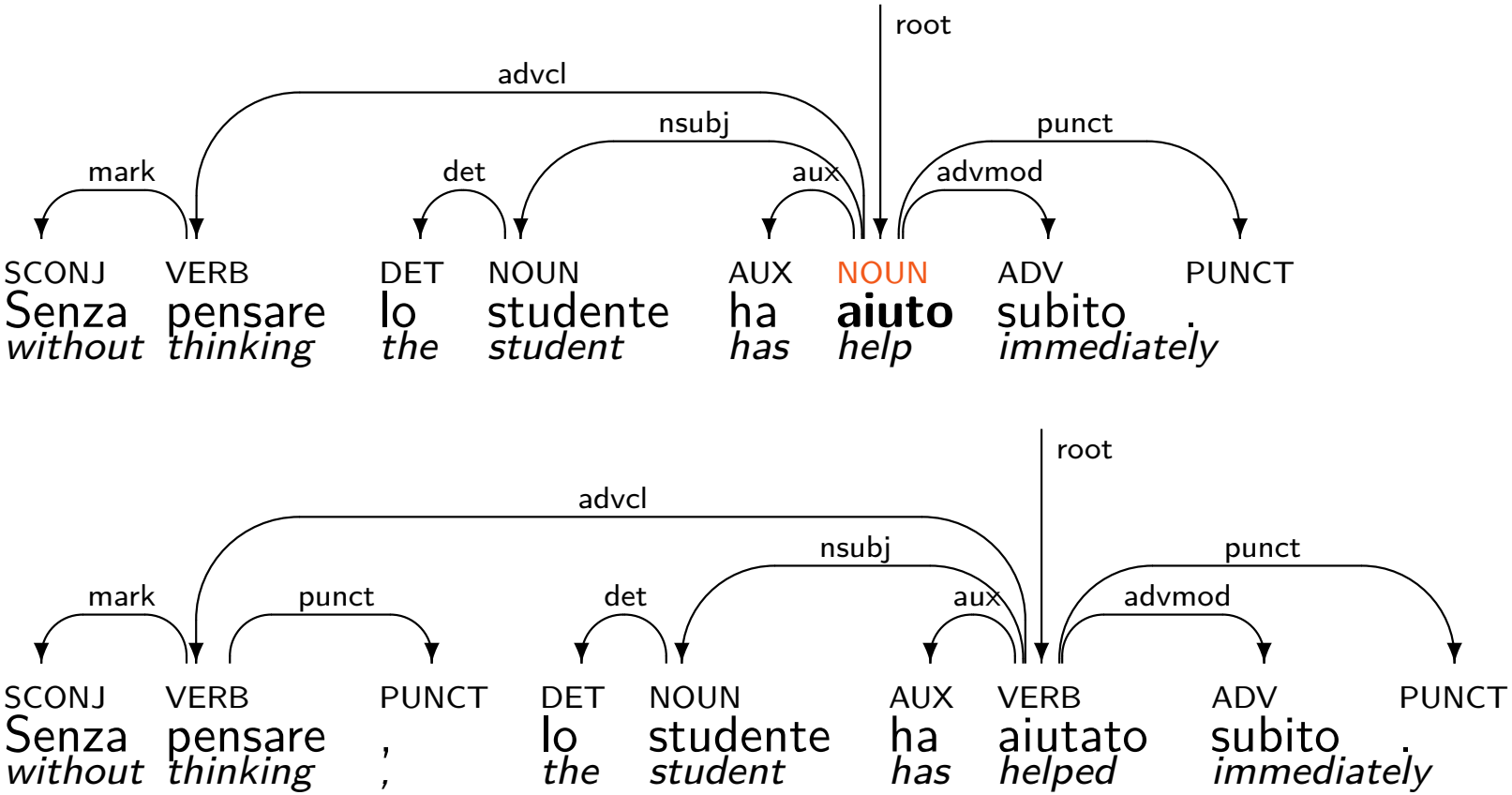
Consistency



Consistency



Consistency



Transfer-aware annotation

AUX/VERB	VERB	ADP	DET	NOUN
Sono	**cerca**	**di**	la	città

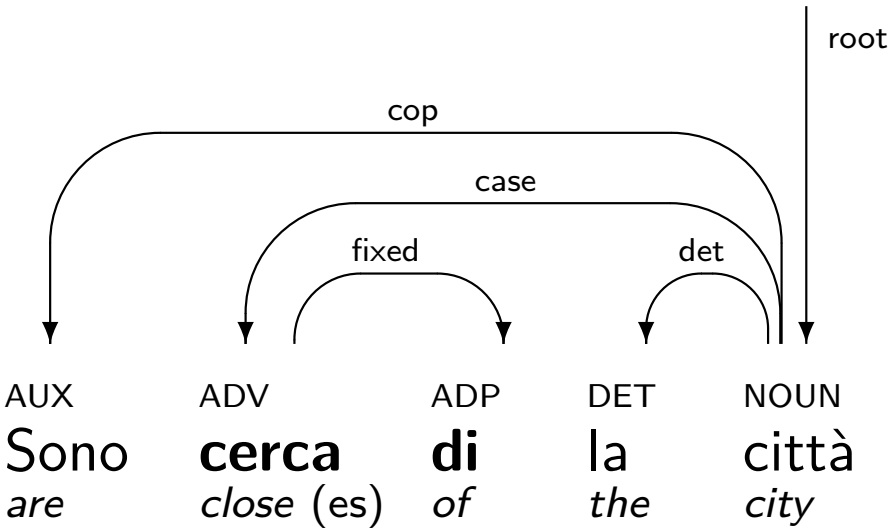
Transfer-aware annotation

AUX/VERB	ADV	ADP	DET	NOUN
Sono	**cerca**	**di**	la	città
están	*cerca*	*de*	*la*	*ciudad*

Transfer-aware annotation

AUX/VERB	ADV	ADP	DET	NOUN
Sono	**cerca**	**di**	la	città
are	*close* (es)	*of*	*the*	*city*

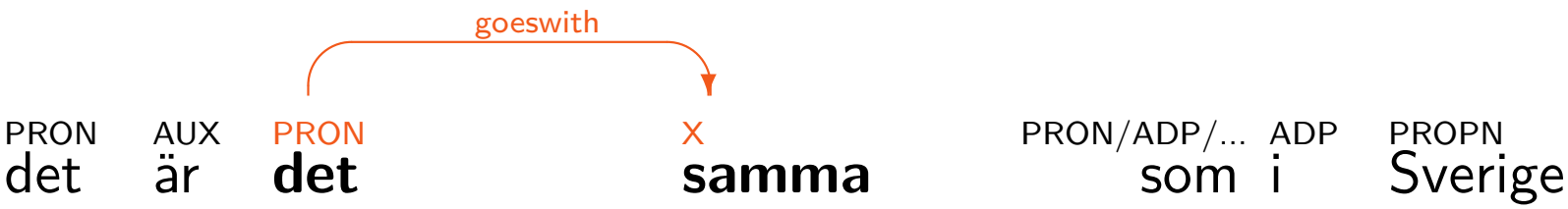
Transfer-aware annotation



Transfer-aware annotation

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

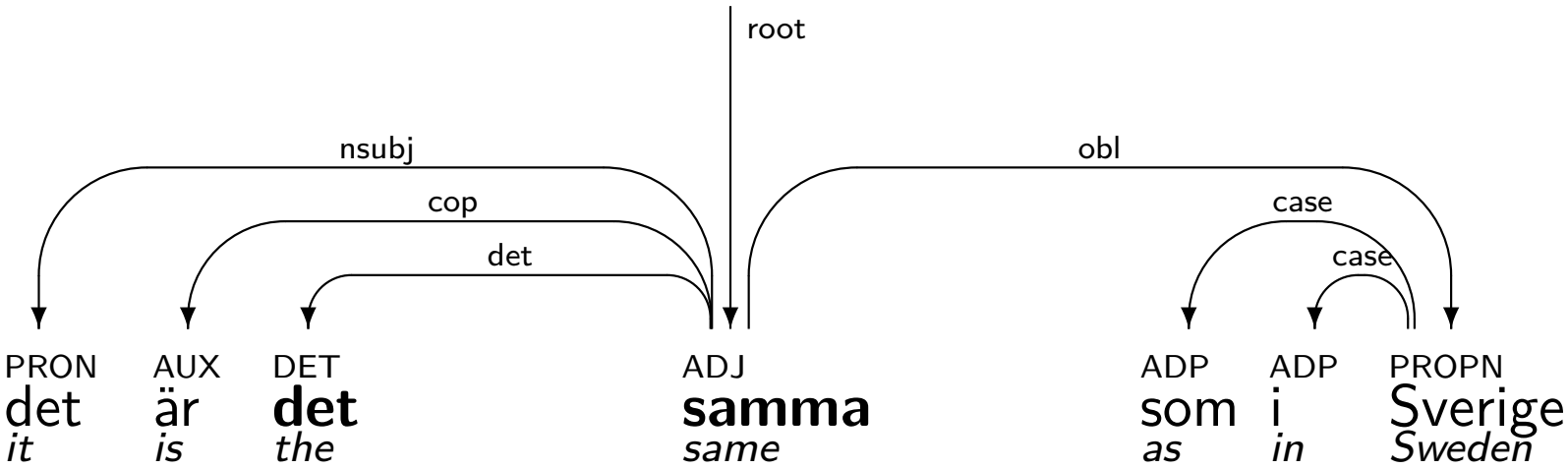
Transfer-aware annotation



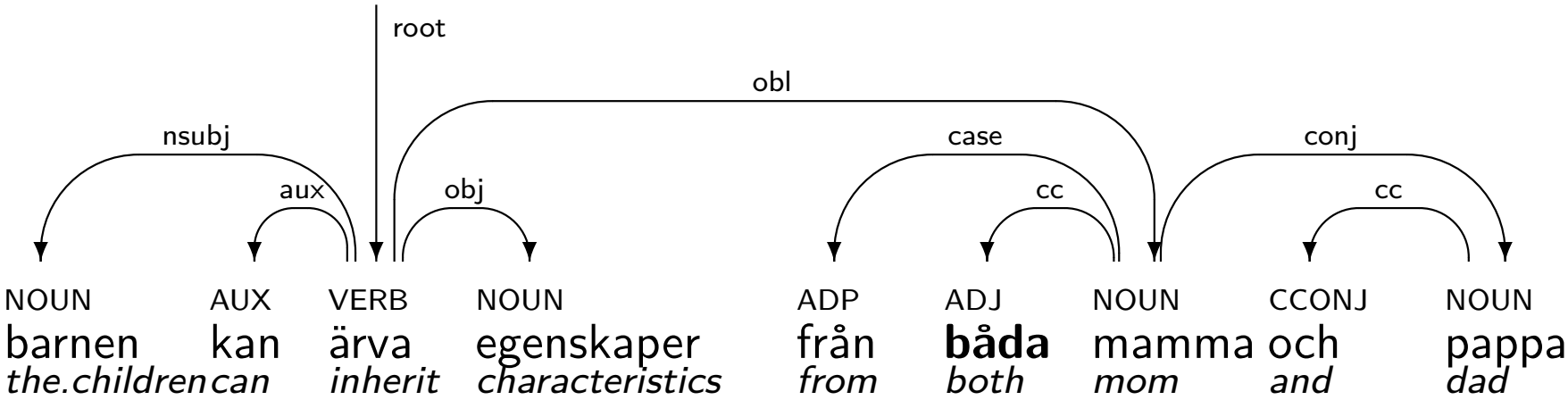
Transfer-aware annotation

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

Transfer-aware annotation

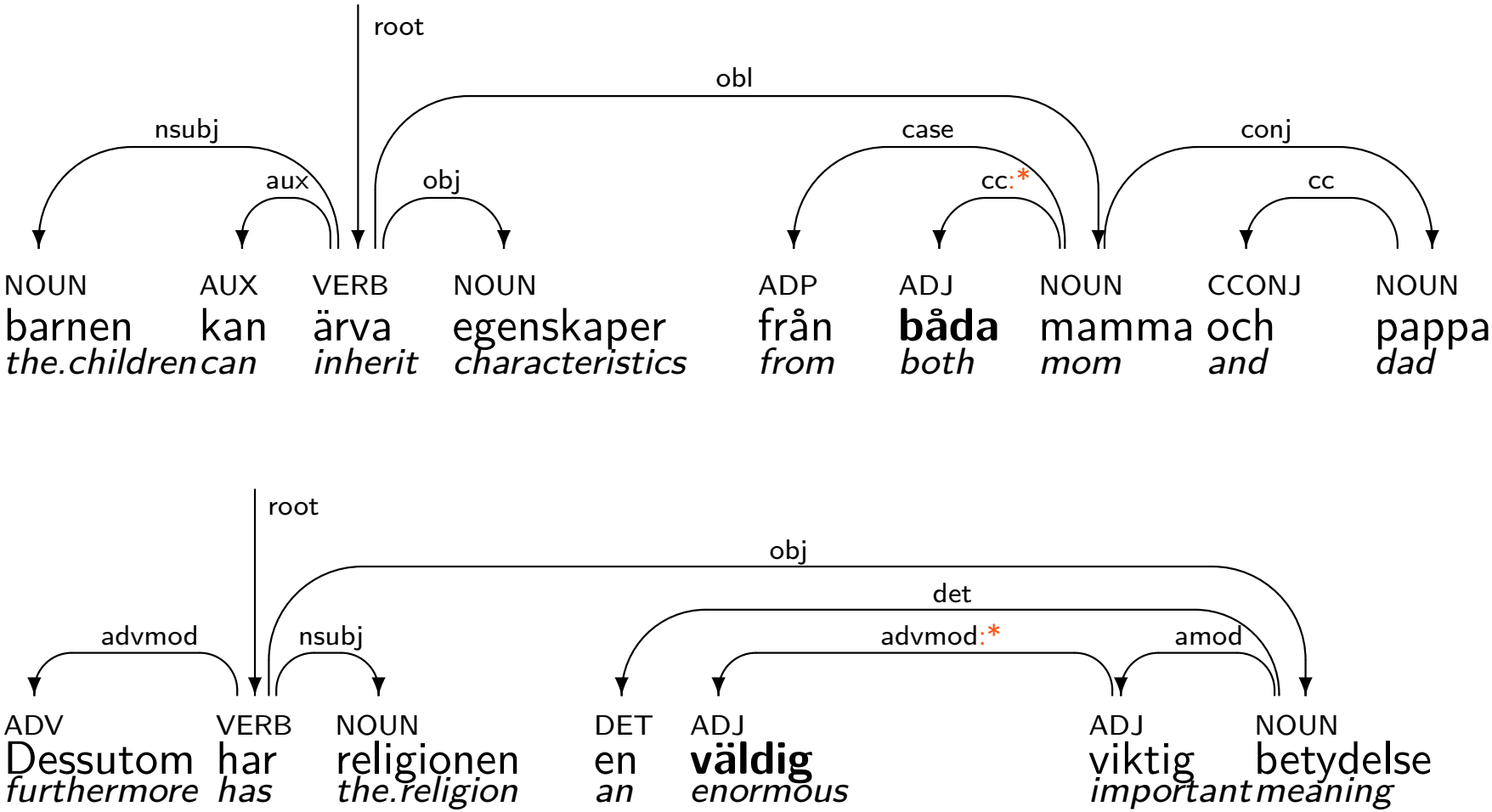


Validation issues




```
[Line 11002 Sent org-548]:
[L3 Syntax rel-upos-cc] 'cc' should not be 'ADJ' ('båda')
```

A new subtype?



Proposal

- **literal** criteria at the token level
  - `UPOS`, `FEATS` and even `LEMMA`
- **correction-aware** syntactic annotation
  - even if the validator complains
- **transfer-aware** annotation of non-idiomatic expressions
  - similar to code-switched analysis of foreign material

(details in 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*, Proceedings of the Eighth Workshop on Universal Dependencies, 2025)

Next steps towards harmonization

1. annotation experiments with other L2 treebank maintainers (English)
2. new guidelines page
3. validator updates

Treebanking SweLL

Source corpus

SweLL-gold, 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)
- **license**: CLARIN-ID -PRIV -NORED -BY

Status

- ~500-sentence **test set** *probably* to be released as part of UD v 1.17 (November 2025)
- ~500-sentence **dev set** *hopefully* to be added in the spring (self-contained project)
- ~4000-sentence **train set** *ideally* to be added over time (automation?)

L2 speakers as annotators

- Aleksandrs Berdičevskis (L1: Latvian & Russian)
- Maria Irena Szawerna (L1: Polish)
- myself (L1: Italian)

Extensions

- A sub-treebank based on the “French SweLL”
- Swedish as a foreign language in France
- annotator: Caroline Grand-Clement (L1: French!)

Tooling for L2 treebanks

L2 parsing is hard!

- Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz, *Universal Dependencies for Learner* ***English***, Proceedings of the 54th Annual Meeting of the ACL, 2016
- Yan Huang, Akira Murakami, Theodora Alexopoulou, and Anna Korhonen, *Dependency parsing of learner* ***English***, International Journal of Corpus Linguistics, 2018
- Elisa Di Nuovo, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino, and Cristina Bosco, *VALICO-UD: Treebanking an* ***Italian*** *Learner Corpus in Universal Dependencies*, IJCoL. Italian Journal of Computational Linguistics, 2022
- Elena Volodina, David Alfter, Therese Lindström Tiedemann, Maisa Susanna Lauriala, and Daniela Helena Piipponen, *Reliability of automatic linguistic annotation: native vs non-native texts*, Selected papers from the CLARIN Annual Conference 2021, 2022 (**Swedish**)

Lazy experiments

- Arianna Masciolini, *Bootstrapping the Annotation of UD Learner Treebanks*. In Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024, 2024
- Arianna Masciolini, Emilie Francis and Maria Irena Szawerna. *Synthetic Error-Augmented Parsing of Swedish as a Second Language: Experiments with Word Order*, Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024, 2024

TL;DR: need *some* authentic annotated L2 data

STUnD

- *Sökverktyg för Tvåspråkiga Universal Dependencies-trädbanker*, or
- Search Tool for (parallel) Universal Dependencies Treebanks
  - not necessarily learner treebanks
- demo at `demo.spraakbanken.gu.se/stund`
- looooooooooooooooong paper in the upcoming *Huminfra handbook for digital humanities*

Under the hood

1. identify subtree alignments
2. run the query on the LHS treebanks, looking for matching subtres
3. find the corresponding RHS subtree (and check if it matches the RHS-specific patters)

STUnD and beyond

- (error) retrieval: patterns → trees
- (error) pattern extraction:[2] trees → patterns
- feedback comment generation: patterns → natural language comments

[2] see Arianna Masciolini, Elena Volodina, and Dana Dannélls. ***Towards automatically extracting morphosyntactical error patterns from L1-L2 parallel dependency treebanks***, Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, 2023

Tack!

Bonus: CALL applications

Feedback Comment Generation

Parse error patterns, generate natural language sentences:

TREE (AND [POS "NOUN", FEATS_ "Gender=Com"])
[AND [POS "DET", FEATS_ "Gender=Neutr"]]

↓

The determiner's gender is neutrum, but the gender of the noun it refers to is common.

Feedback Comment Generation

Parse error patterns, generate natural language sentences:

TREE (AND [POS "NOUN", FEATS_ "Gender=Com"])
[AND [POS "DET", FEATS_ "Gender=Neutr"]]

↓

OBS: detta substantiv är ett en-ord!

or

Fai alla concordanza tra il nome e il determinante

Feedback Comment Generation

Parse error patterns, generate natural language sentences:

TREE (AND [POS "NOUN", FEATS_ "Gender=Com"])
[AND [POS "DET", FEATS_ "Gender=Neutr"]]

Pay attention to gender agreement!