UD as an annotation standard for learner language

a case study on L2 Swedish

Arianna Masciolini

LT2214 Computational Syntax

VT25

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

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

The solution: UD

- fine-grained morphosyntactic annotation
- parsers
- cross-linguistic consistency $\rightarrow$ possibility to compare:
  - L2 vs. standard
  - L1 vs. L2
  - different L2s

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

- L2 sentences ‖ correction hypotheses
- no explicit error tagging

UD treebanks of learner language

language	name	size	status	parallel
Chinese	CFL	451	released	**yes****
English	ESL	5124	retired*	**yes**
English	ESLSpok	2320	released	no
Italian	Valico	398	released	**yes**
Korean	KSL	12977	released	no
Russian	?	500	WIP	**yes**
Swedish	SweLL	~5000	WIP	**yes**

*available for download but not part of the latest UD release
**only L2 half available

Challenges

expectations	**reality**
fine-grained annotation	when the validator allows that
parsers	don't work terribly well
cross-linguistic consistency	is limited to error-free spans

The `root` of the problem

The UD guidelines are designed with standard language in mind

- should we annotate the intended meaning (correction) and/or the observed language use?
- how to handle mismatches between the characteristics of individual tokens and their use in context?

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

Example 0

Självklart	**att**	**det**	**är**	viktigt	.
of.course	*that*	*it*	*is*	*important*	*.*

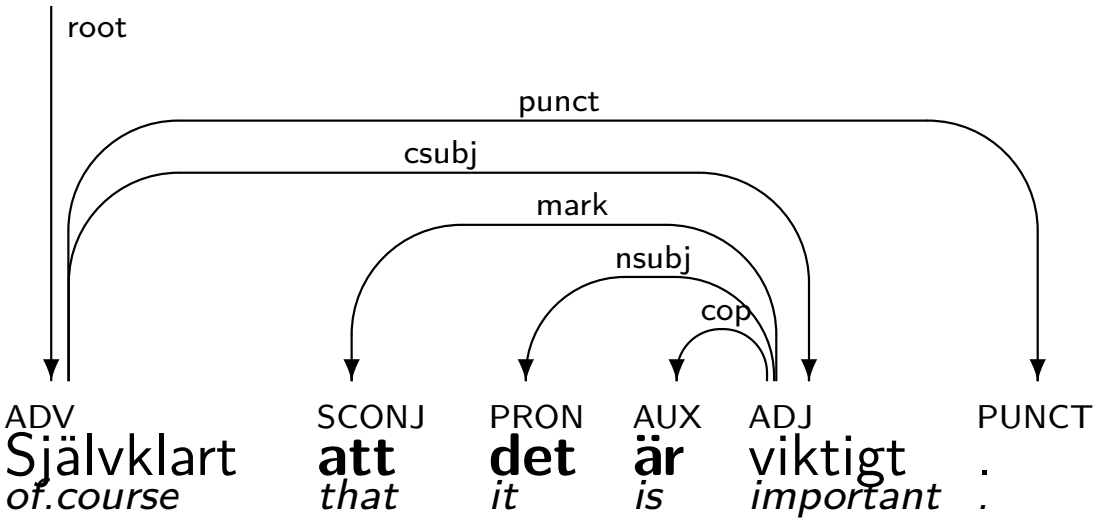
- correction: “Självklart **är det** viktigt.”
- translation: “Of course it is important.”

Example 0

ADV	SCONJ	PRON	AUX	ADJ	PUNCT
Självklart	**att**	**det**	**är**	viktigt	.
of.course	*that*	*it*	*is*	*important*	*.*

- correction: “Självklart **är det** viktigt.”
- translation: “Of course it is important.”

Example 0



- correction: “Självklart **är det** viktigt.”
- translation: “Of course it is important.”

Example 1

Jag	hade	**emotskänslor**	fast	jag	**var**	**vänta**	det
I	*had*	*againstfeelings*	*although*	*I*	*was*	*wait*	*that*

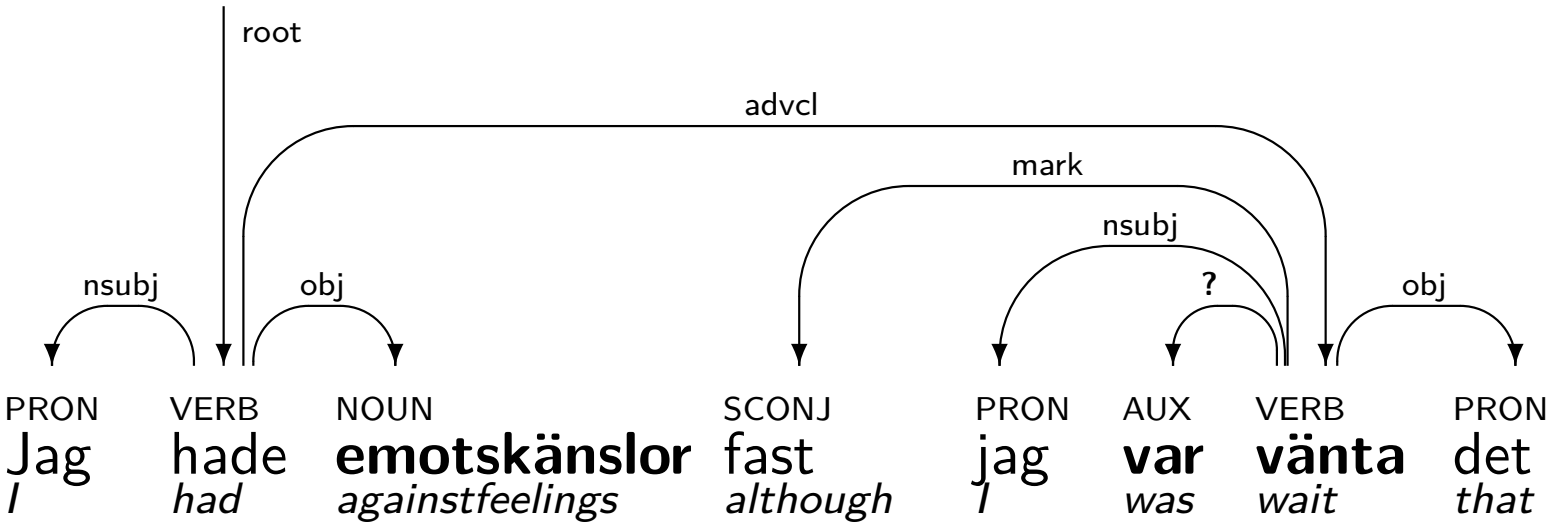
- correction: “Jag hade **motstridiga känslor** fast jag **hade väntat mig** det”
- translation: “I had mixed feelings although I was expecting that”

Example 1

PRON	VERB	NOUN	SCONJ	PRON	AUX	VERB	PRON
Jag	hade	**emotskänslor**	fast	jag	**var**	**vänta**	det
I	*had*	*againstfeelings*	*although*	*I*	*was*	*wait*	*that*

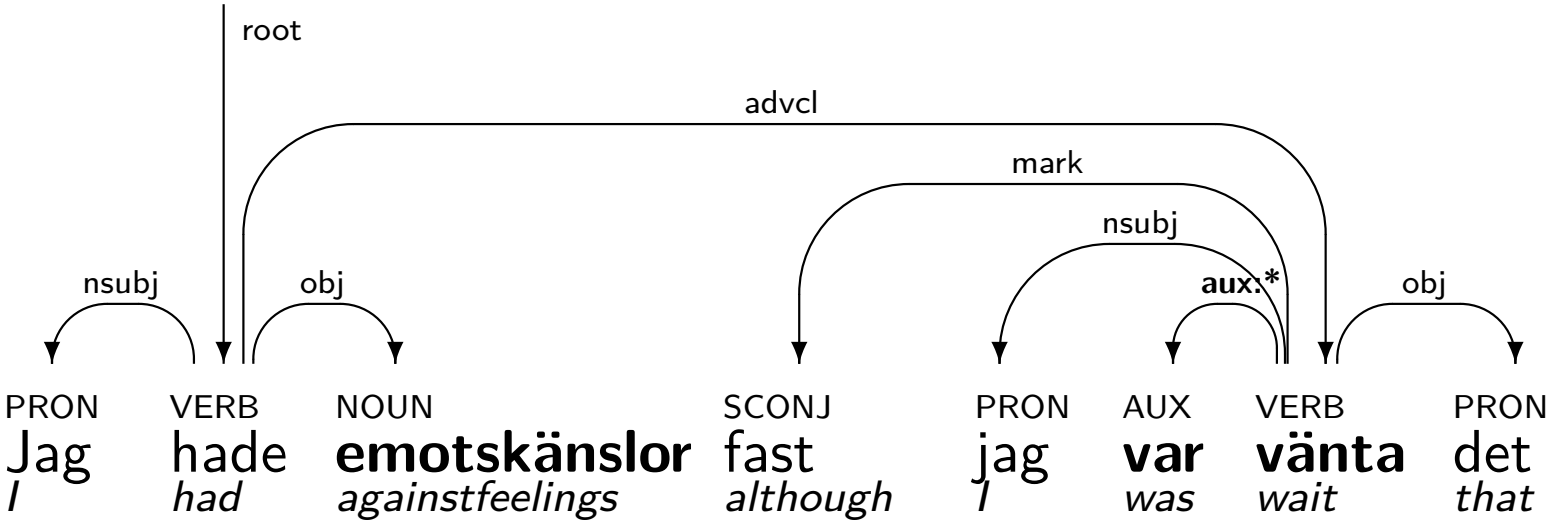
- correction: “Jag hade **motstridiga känslor** fast jag **hade väntat mig** det”
- translation: “I had mixed feelings although I was expecting that”

Example 1



- correction: "Jag hade **motstridiga känslor** fast jag **hade väntat mig** det"
- translation: "I had mixed feelings although I was expecting that"

Example 1



- correction: "Jag hade **motstridiga känslor** fast jag **hade väntat mig** det"
- translation: "I had mixed feelings although I was expecting that"

Example 2

en	lång	**bus**	**resa**
a	*long*	*bus*	*trip*

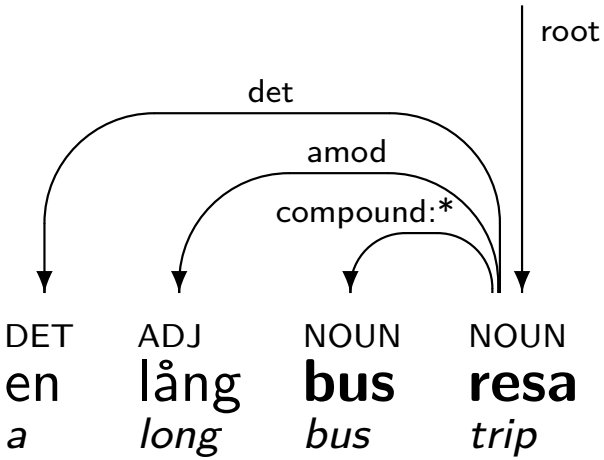
- correction: “en lång **bussresa**”
- translation: “a long bus trip”

Example 2

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

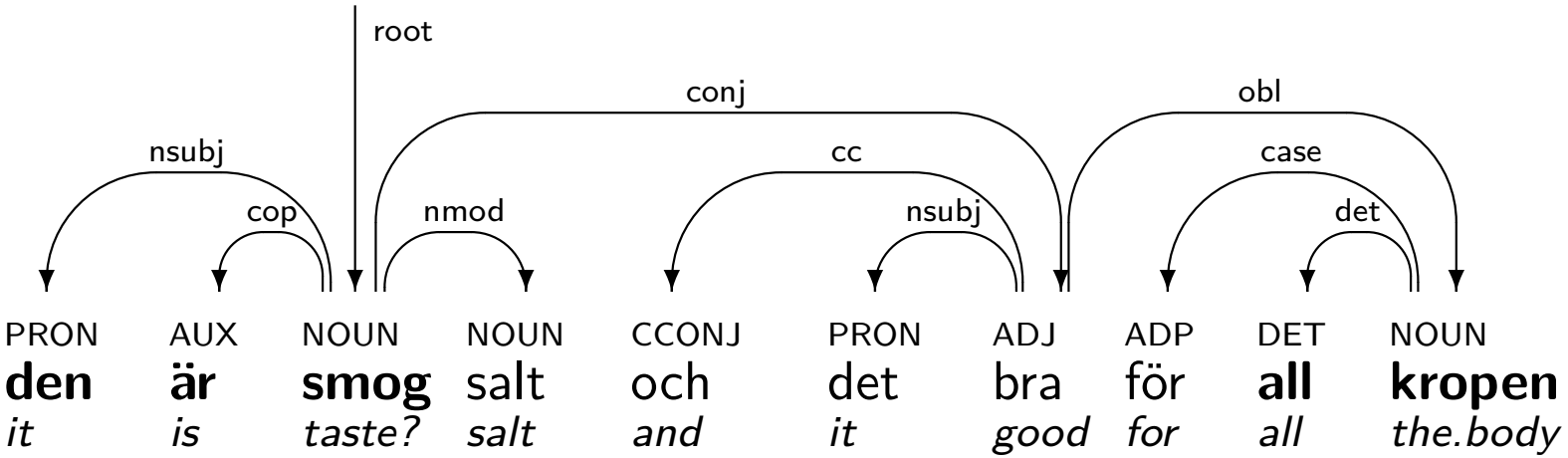
- correction: “en lång **bussresa**”
- translation: “a long bus trip”

Example 2



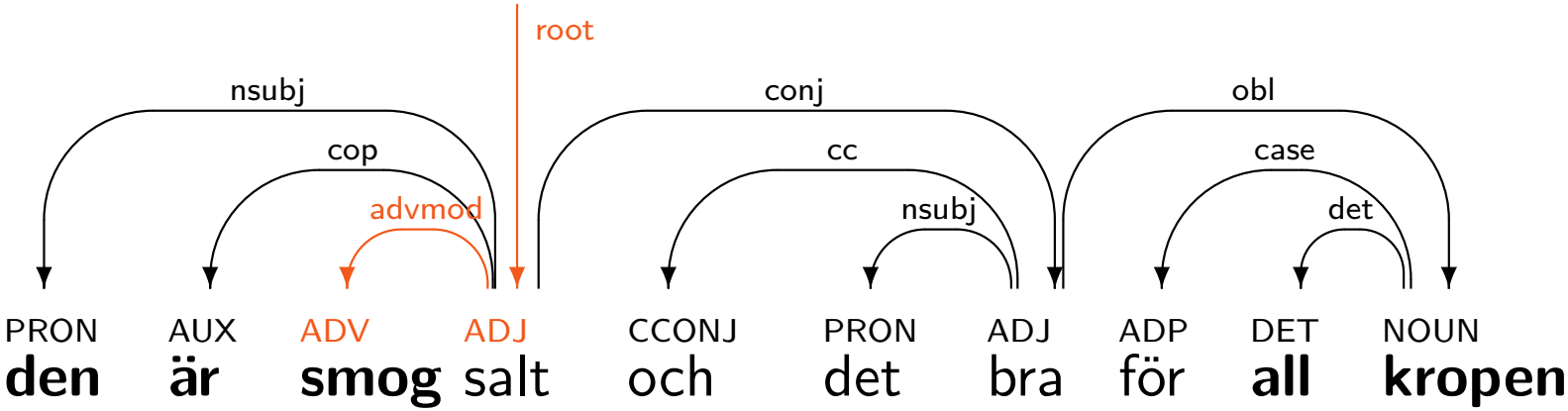
- correction: “en lång **bussresa**”
- translation: “a long bus trip”

Example 3



- correction: "**Det smakar** salt och det **är** bra för **hela kroppen**"
- translation: "it tastes salt and it's good for the whole body"

Example 3: parser output



(obtained with the UDPipe 2 Talbanken 2.15 model)

Our principles

- the validator is a tool, not a goal:
  - ***literal* criteria at the token level**
  - ***distributional* criteria at the syntax level**
  - **borrow from L1** guidelines when necessary
- **correction-aware annotation**: the annotation of learner sentences should be consistent with the semantics of the correction hypothesis

Status

- guidelines and test set (200/500 sentences) WIP
- remaining 5000 + 500 sentences TODO

Status

- guidelines and test set (200/500 sentences) WIP
- remaining 5000 + 500 sentences TODO
  - you are welcome to **participate**!
    - you do *not* have to be a native speaker (in fact, none of the current annotators is)
    - you *might* be able to do this as a course project

Exploring parallel learner treebanks with STUnD

STUnD

- *Sökverktyg för Tvåspråkiga Universal Dependencies-trädbanker*, or
- Search Tool for (parallel) Universal Dependencies Treebanks
- available at `demo.spraakbanken.gu.se/stund` (hopefully)

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)

Use cases

- error retrieval: patterns (queries) $\rightarrow$ trees
- pattern extraction: trees $\rightarrow$ patterns
- feedback comment generation: patterns $\rightarrow$ natural language comments

Sources

In order of appearance

- John Lee, Keying Li, and Herman Leung. *L1-L2 parallel dependency treebank as learner corpus*. In Proceedings of the 15th International Conference on Parsing Technologies, pages 44-49, Pisa, Italy, September 2017. Association for Computational Linguistics
- John Lee, Herman Leung, and Keying Li. *Towards Universal Dependencies for learner Chinese*. In Marie-Catherine de Marneffe, Joakim Nivre, and Sebastian Schuster, editors, Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pages 67-71, Gothenburg, Sweden, may 2017. Association for Computational Linguistics

In order of appearance

- Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. *Universal Dependencies for learner English*. In Katrin Erk and Noah A. Smith, editors, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 737-746, Berlin, Germany, aug 2016. Association for Computational Linguistics.
- 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, 8(8-1), 2022

In order of appearance

- Hakyung Sung and Gyu-Ho Shin. *Constructing a dependency treebank for second language learners of Korean*. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 3747-3758, Torino, Italia, may 2024. ELRA and ICCL
- Hakyung Sung and Gyu-Ho Shin. *Second language Korean Universal Dependency treebank v1.2: Focus on data augmentation and annotation scheme refinement*. In Špela Arhar Holdt, Nikolai Ilinykh, Barbara Scalvini, Micaella Bruton, Iben Nyholm Debess, and Crina Madalina Tudor, editors, Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025), pages 13-19, Tallinn, Estonia, March 2025. University of Tartu Library, Estonia

In order of appearance

- Alla Rozovskaya. *Universal Dependencies for learner Russian*. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 17112-17119, Torino, Italia, may 2024. ELRA and ICCL
- Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, et al. *The SweLL language learner corpus: From design to annotation*. Northern European Journal of Language Technology, 6:67-104, 2019
- Arianna Masciolini. *A query engine for L1-L2 parallel dependency treebanks*. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pages 574–587, Tórshavn, Faroe Islands, May 2023. University of Tartu Library

In order of appearance

- Arianna Masciolini, Elena Volodina, and Dana Dannélls. *Towards automatically extracting morphosyntactical error patterns from L1-L2 parallel dependency treebanks*. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 585-597, Toronto, Canada, jul 2023. Association for Computational Linguistics
- Arianna Masciolini and Márton A Tóth. *STUnD: ett Sökverktyg för Tvåspråkiga Universal Dependencies-trädbanker*. In Proceedings of the Huminfra Conference, pages 95-109, Gothenburg, Sweden, 2024

To appear

- Arianna Masciolini, Herbert Lange and Márton A Tóth. *Exploring parallel corpora with STUnD: a Search Tool for Universal Dependencies*. In the upcoming Huminfra Handbook, Gothenburg, Sweden, **most likely** 2025
- a paper about harmonization of UD guidelines for L2 treebanks (under review)