Cross-lingual approaches to computational SLA: the potential of Universal Dependencies

mid seminar - 21.10.2023

Arianna Masciolini

Supervisors: Elena Volodina and Dana Dannélls

How this presentation works

- slides with a **blue** frame are from my ideas seminar, back in early April 2023[1]
- everything else either
  - happened during the last 1.5 years
  - I have found a better way to explain

[1] possibly with minor modifications

Keywords

- **L2 grammar acquisition**
- **tutorial ICALL**
- **exercise generation**
- self-study $\rightarrow$ **automatic feedback**
- **multilingual**
- **grammar-based**

Keywords'

- **L2 grammar acquisition**
- ~~**tutorial**~~ **ICALL** (language tools)
- ~~**exercise generation**~~
- self-study $\rightarrow$ **automatic feedback** $\rightarrow$ **AWE/FCG**
- **multilingual**
- **grammar-based**, but also data-driven (**learner corpora**)

Keywords'

- **L2 grammar acquisition**
- **ICALL** (language tools)
- self-study $\rightarrow$ **automatic feedback** $\rightarrow$ **AWE/FCG**
- **multilingual**
- **grammar-based**, but also **data-driven** (learner corpora)

Keywords”

- **computational SLA**
- **grammar**
- **cross-linguality**, but also focus on **L2 Swedish**
- (parallel) **UD learner treebanks**

Resarch question

Can leveraging Universal Dependencies help develop cross-lingually applicable tools and methods for computational SLA?

Universal Dependencies 101

What is UD?

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

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.

Example – step 1: segmentation

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

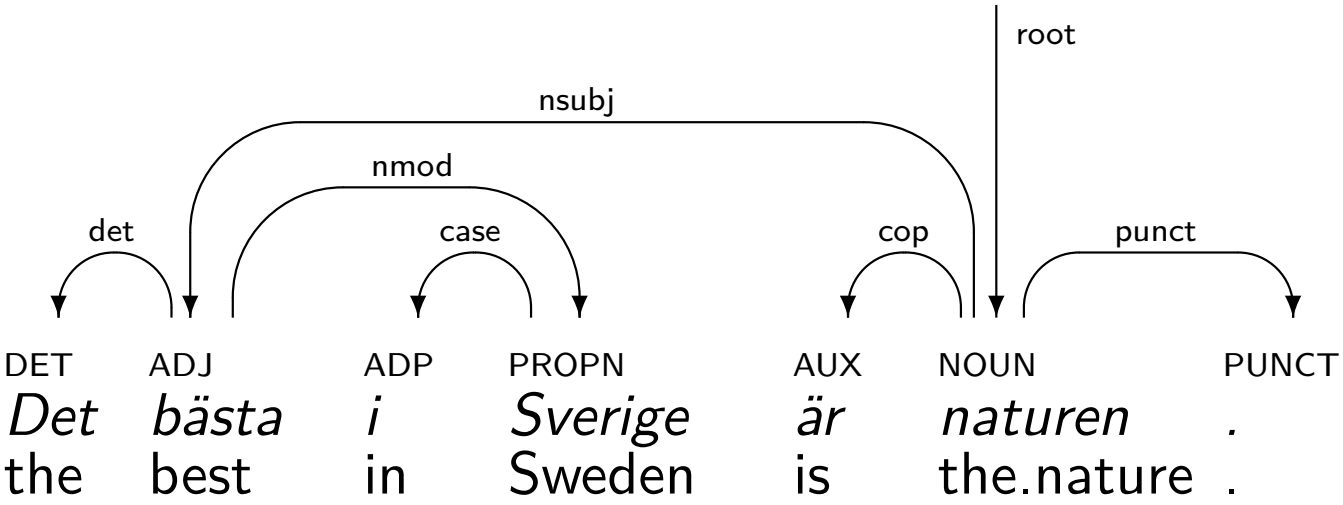
“The best thing in Sweden is nature.”

Example – 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.”

Example – step 3: syntax



“The best thing in Sweden is nature.”

Example: 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

Example: CoNNL-U

```
# 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   _   _
```

Why use UD for L2 corpora?

- existing parsers allow for **faster annotation**
- rich morphosyntactic annotation supports the **study of L2 grammatical patterns**
- u-categories facilitate **cross-lingual comparisons** between:
  - a learner's L1 and L2
  - different L2s
  - standard and learner language

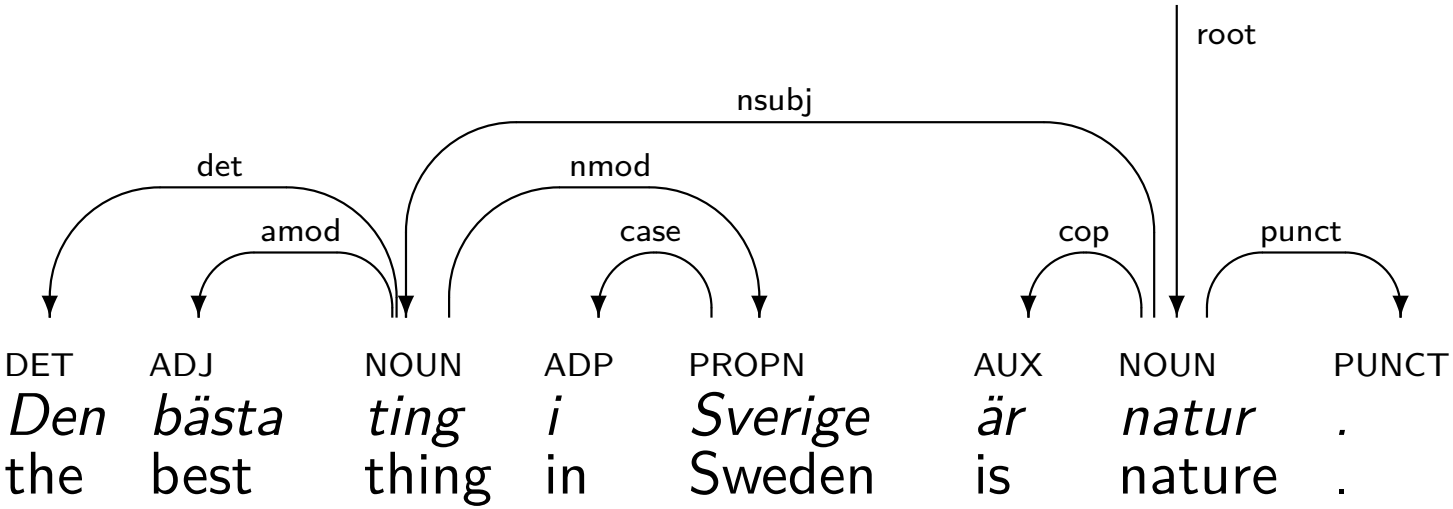
L1-L2 treebanks

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

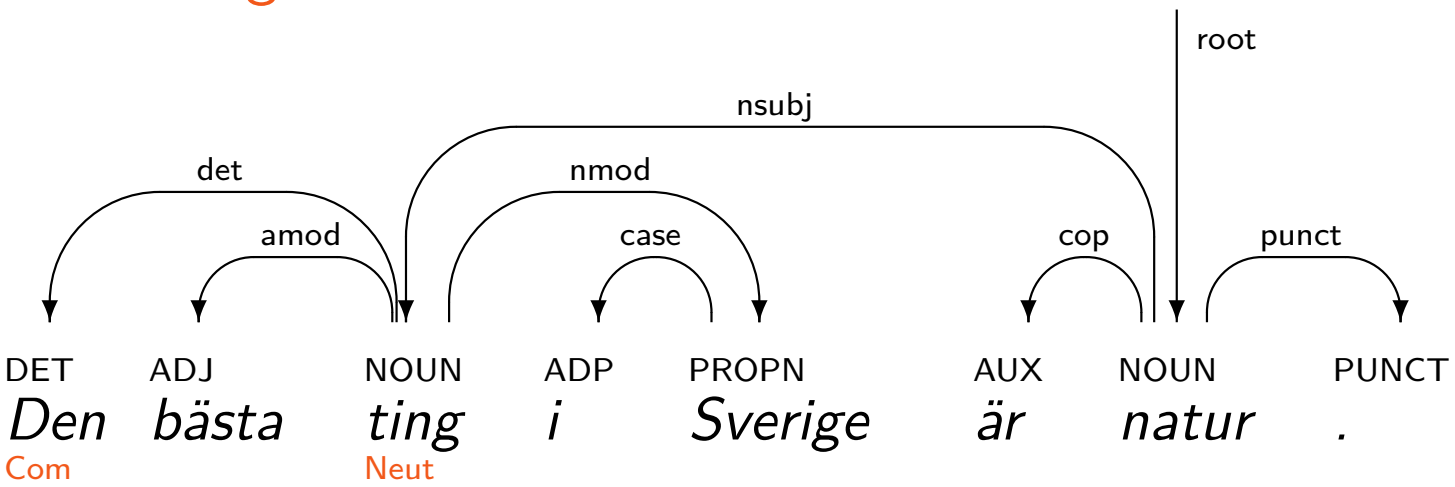
- L2 sentences // correction hypotheses
- no explicit error tagging, just **UD annotation**
  - better **interoperability** between learner corpora

Example


root
nsubj
det
nmod
amod
case
cop
punct
DET ADJ NOUN ADP PROPN AUX NOUN PUNCT
Den bästa ting i Sverige är natur .
the best thing in Sweden is nature .

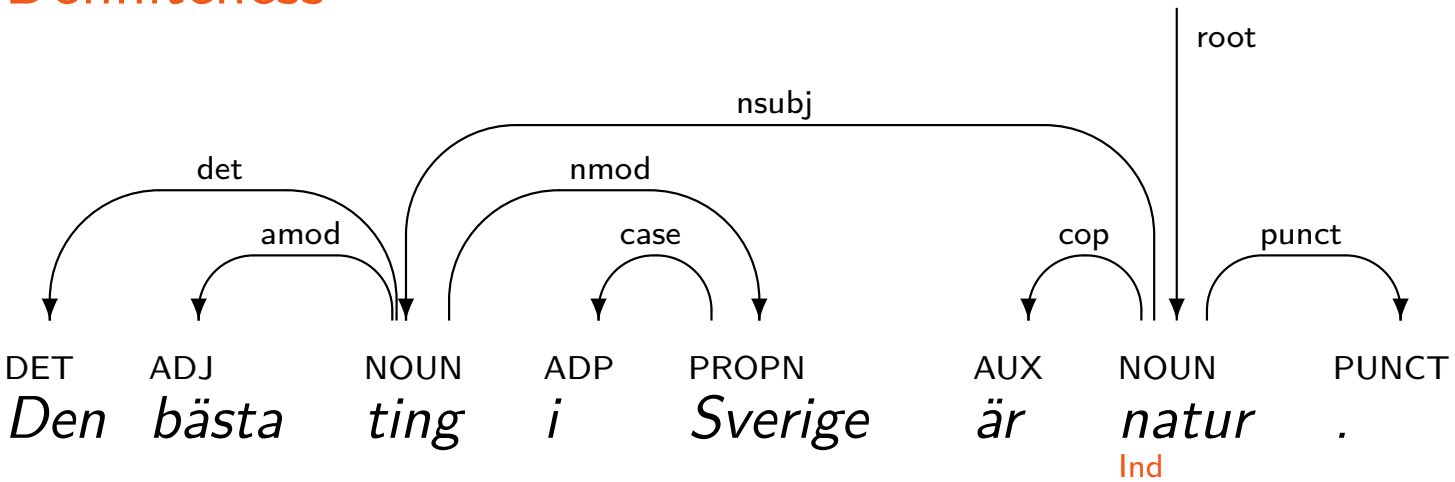
Example

Gender agreement



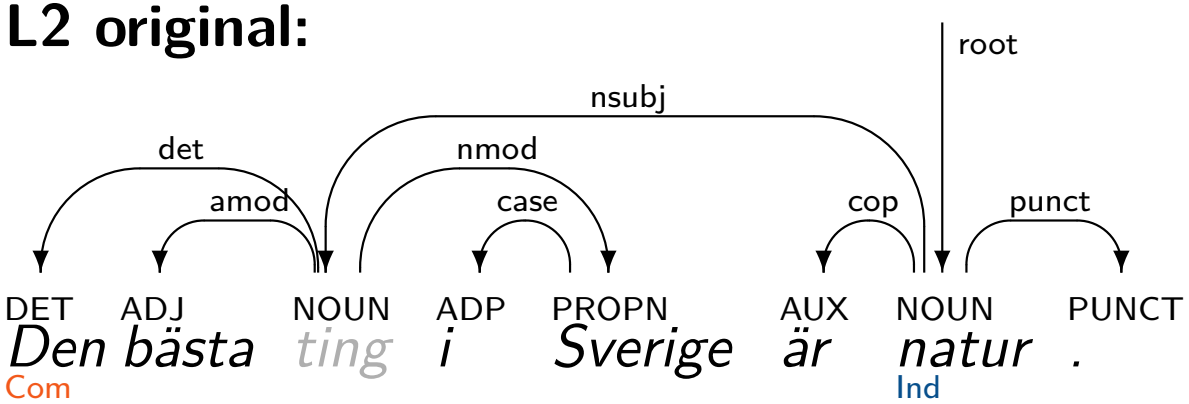
Example

Definiteness

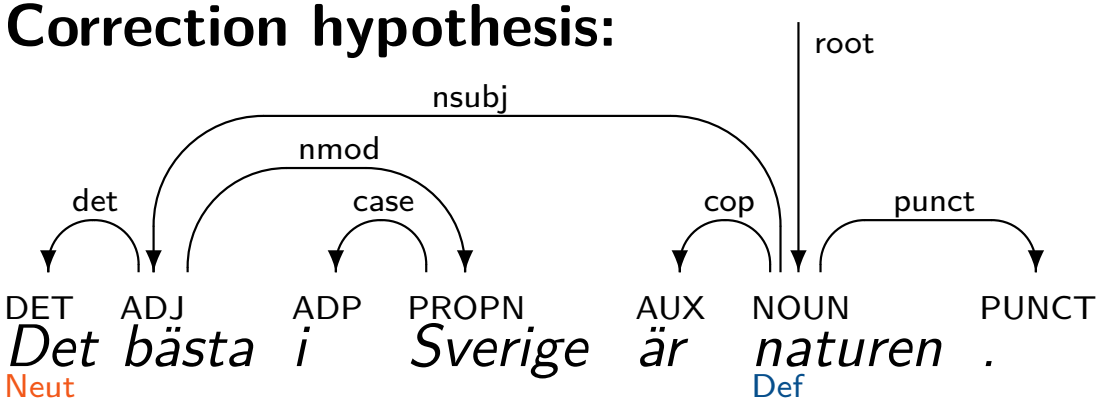


Example

L2 original:



Correction hypothesis:



L2 treebanks in UD

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

Treebanking SweLL

Treebanking SweLL

SweLL-UD: A Treebank of L2 Swedish Essays

1st UniDive Training School, 8-12 July 2024, Technical University of Moldova

Arianna Masciolini with Maria Irena Szawerna and Elena Volodina

Språkbanken Text, Department of Swedish, Multilingualism, Language Technology, University of Gothenburg, Sweden

UD Treebanks in SLA Research

Advantages of UD treebanks in corpus-based Second Language Acquisition research:

- uniform morphosyntactic annotation layer allowing for qualitative and quantitative **cross-lingual comparisons**:
  - between standard and learner language
  - between a learner's L1 and L2
  - between different L2s
- possibility to carry out **grammatical error retrieval and analysis** and **automatic feedback generation** without explicit error tagging e.g. if learner sentences are paired with corrections (*parallel L1-L2 treebank*)
- **semi-automatic annotation** with the help of the existing UD parsers

Existing Second Language UD Treebanks

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

* available through GitHub but not part of the latest UD release

The SweLL-UD treebank

Objectives

- end goal: building a **training-scale treebank** of L2 Swedish
- within the next few months: releasing a high-quality **500-sentence test set**
- at the training school: further experimenting with **annotation of L2 essays** and discussing **open questions**

Data

Source corpus: **SweLL-gold** (Swedish Learner Language corpus)

- **genre**: essays (various topics)
- **learners**: adult L2 Swedish students with various language backgrounds and proficiency levels
- **annotation**: manual correction, error tagging and pseudonymization
- **size**: 502 essays (5000+ parallel sentences)
- **license**: CLARIN-ID -PRIV -NORED -BY (but the data can and has been redistributed as long as it is impossible to reconstruct full essays)

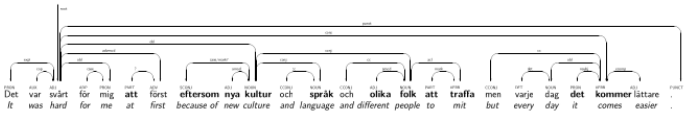
Project status and plan

1. **preprocessing**: sentence pair and metadata extraction, CoNNL-U format conversion, shuffling and creation of train, development and test splits ✓
2. **preliminary annotation experiments** with a small annotation team ✓
3. **L1 (corrections) annotation**:
   - automatic pre-annotation with a pretrained UDPipe model ✓
   - manual validation with a larger annotation team including both L1 and L2 Swedish speakers
4. **L2 (originals) annotation**:
   - automatic pre-annotation with a custom (fine-tuned) model
   - manual validation with the larger annotation team

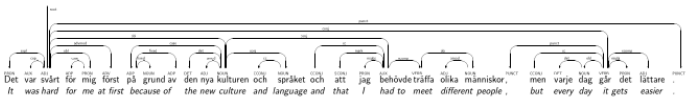
Open questions

- **tooling** - desiderata:
  - tree visualization
  - no installation required (as with Eevee)
  - parallel L1-L2 editing (or at least viewing, as with STUnD)
- **L2-specific guidelines**:
  - when annotating an original learner sentence, to what extent should the syntactic annotation of its correction be taken into account?
  - what is a good balance between the general principle of *literal reading* and keeping the two trees as structurally similar as possible?
  - should the existing L1-specific guidelines be applied when annotating constructionsn that the learners borrow from their L1?

Original sentence from a learner essay (errors highlighted in bold)



Corresponding correction hypothesis



Acknowledgement

Participation to the training school is funded by UniDive, while the annotation project is supported by the Swedish National Research Infrastructure Nationella Språkbanken, funded jointly by the Swedish Research Council (2018–2024, contact 2017-00626) and the ten participating partner institutions.

Universal Dependencies Meets Second Language Acquisition: the Case of Swedish

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**

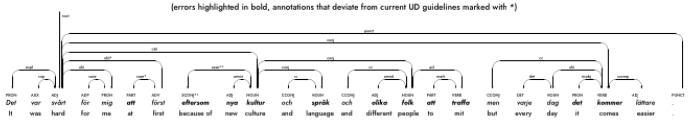
Example

Metadata

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

Original Sentence

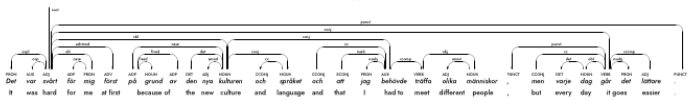
(errors highlighted in bold, annotations that deviate from current UD guidelines marked with *)



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



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

Project plan

1. **preprocessing**: sentence pair extraction and automatic pre-annotation (completed)
2. **manual validation** of a 500-sentence **test set** // **guidelines development** (ongoing)
3. **test set release** (planned for May 2025) - 2 versions:
   - sentence-shuffled version at (UD)
   - full-essay version with all metadata (SweLL v2?)
4. incremental **annotation and release** of a **dev** and a **train set**

Conflicting goals

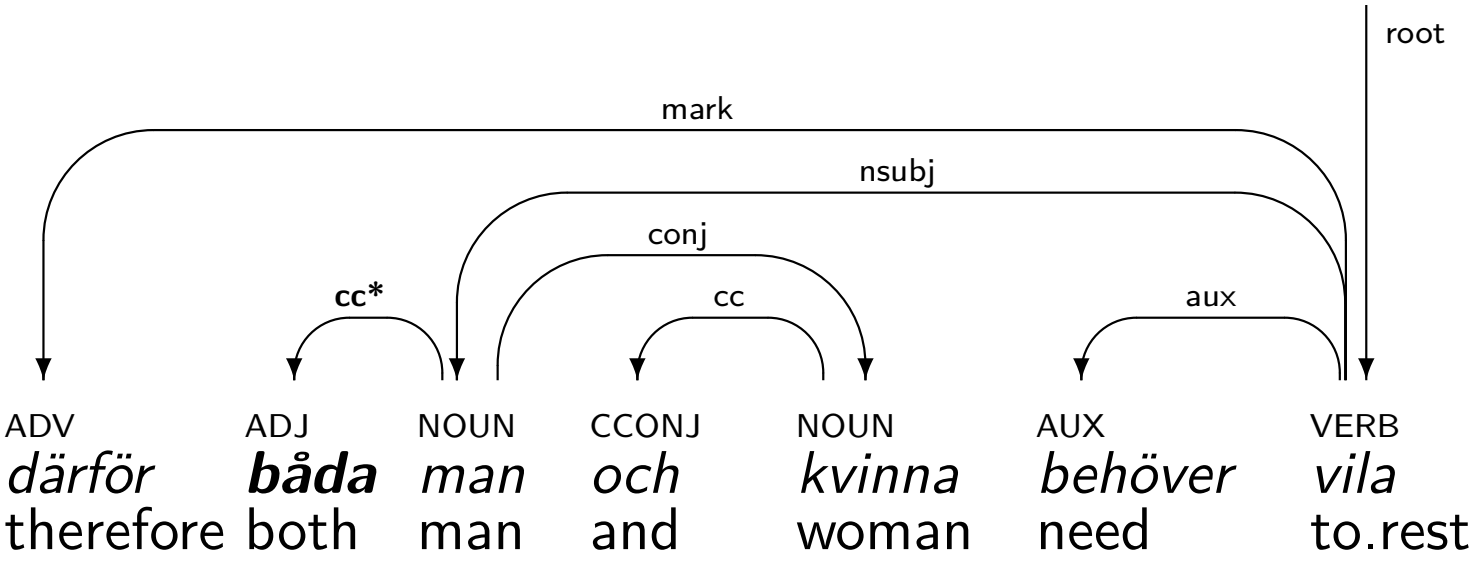
Trees in a learner sentence-correction pair should be:

- **as different as necessary**, so to not "hide" any discrepancies and account for all L2-specific phenomena
- **as similar as possible**, to facilitate qualitative comparisons
- (**acceptable according to existing guidelines**)

Some ideas for...

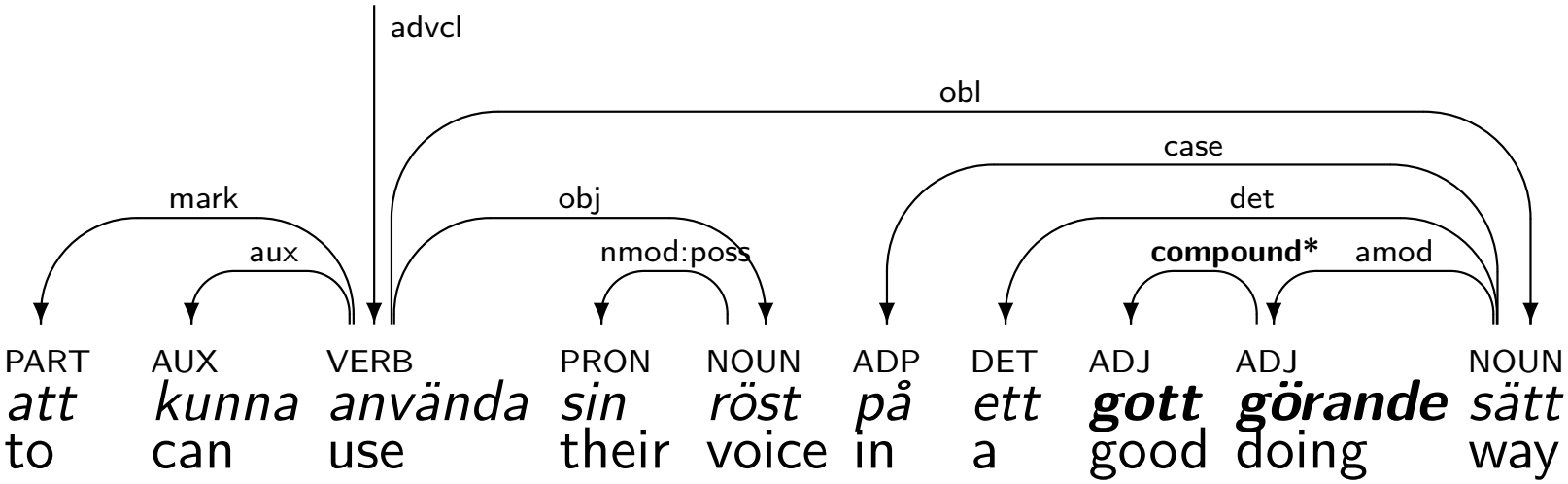
- ...dealing with **ungrammatical segments**:
  - *literal reading* for word-level annotation
  - *distributional criteria* at the syntactic level
- ...annotating **foreign constructions**: borrowing language-specific guidelines from the learner's L1s

Example – ungrammaticality



The word *båda* exists, but it is only an adjective. However, the learner is using it as as the conjunction *både*.

Example – foreign construction



Similar to English *well-meaning*, *good looking*.

From UD-annotated data to feedback comments

Steps

Given a learner sentence:

1. obtain correction hypothesis
2. annotate learner sentence and correction in UD
3. extract error patterns
4. generate feedback comments

Steps

Given a learner sentence:

1. **obtain correction hypothesis**
2. annotate learner sentence and correction in UD
3. extract error patterns
4. generate feedback comments

1. Grammatical Error Correction

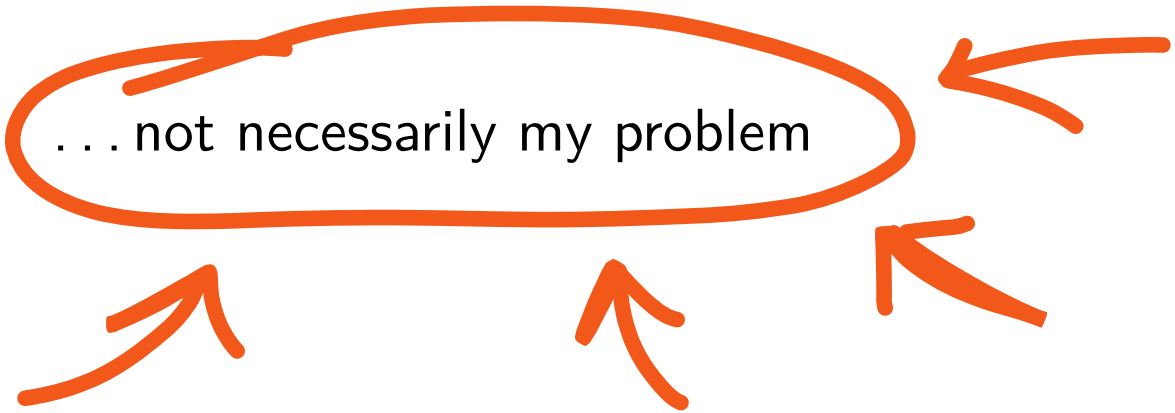
“detta mening korrekt grammatisk?”

↓

“Är denna mening grammatiskt korrekt?”

1. Grammatical Error Correction

- Well established task
- several promising approaches
  (see Bryant et al., 2022 for a recent survey)
- Swedish:
  - Granska system (Domeik et al., 2000)
  - Nyberg, 2022
  - Östling and Kurfali, 2022
- back-and-forth MT to the learner's L1 can help



The MultiGEC–2025 shared task

The shared task

UNIVERSITY OF GOTHENBURG

SPRÅKBANKENTEXT

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

Svenska

Språkbanken Text is a department within Språkbanken.

News and events | Blog | **Research** | Tools | Data | FAQ | About us | Contact us

Home / Computational SLA / MultiGEC-2025

Computational SLA

MultiGEC-2025

MultiGED-2023

MultiGEC-2025

Shared task on Multilingual Grammatical Error Correction (MultiGEC-2025)

The Computational SLA working group invites you to participate in the shared task on Multilingual Grammatical Error Correction, **MultiGEC**, covering 12 languages: Czech, English, Estonian, German, Greek, Icelandic, Italian, Latvian, Russian, Slovene, Swedish and Ukrainian (see also the call for participation on the ACL portal).

The results will be presented on March 5, 2025, at the NLP4CALL workshop, colocated with the NoDaLiDa conference to be held in Estonia, Tallinn, on 2--5 March 2025.

The publication venue for system descriptions will be the proceedings of the NLP4CALL workshop.

To register for/express interest in the shared task, please fill in this form.

The shared task in numbers

- 12 languages and 18 subcorpora
- 8 organizers and 28 data providers
- 2 tracks
- 3 evaluation metrics
- 1 multilingual baseline

Tracks

Learner essay

My mother became very sad, no food. But my sister better five months later.

With **minimal edits**

*My mother became very sad, **and ate** no food. But my sister **felt better** five months later.*

With **fluency edits**

*My mother **was** very **distressed and refused to eat. Luckily** my sister **recovered** five months later.*

Evaluation

- 2 **reference-based** metrics (better for minimal edits):
  - GLEU
  - $F_{0.5}$
- Scribendi score (**referenceless** and LM-based, better for fluency edits)

Baseline

Desiderata

Multilingual, simple, completely offline, works for both tracks.

Three approaches tested

1. backtranslation
2. zero-shot LLM-based
3. **one-shot LLM-based**

Steps

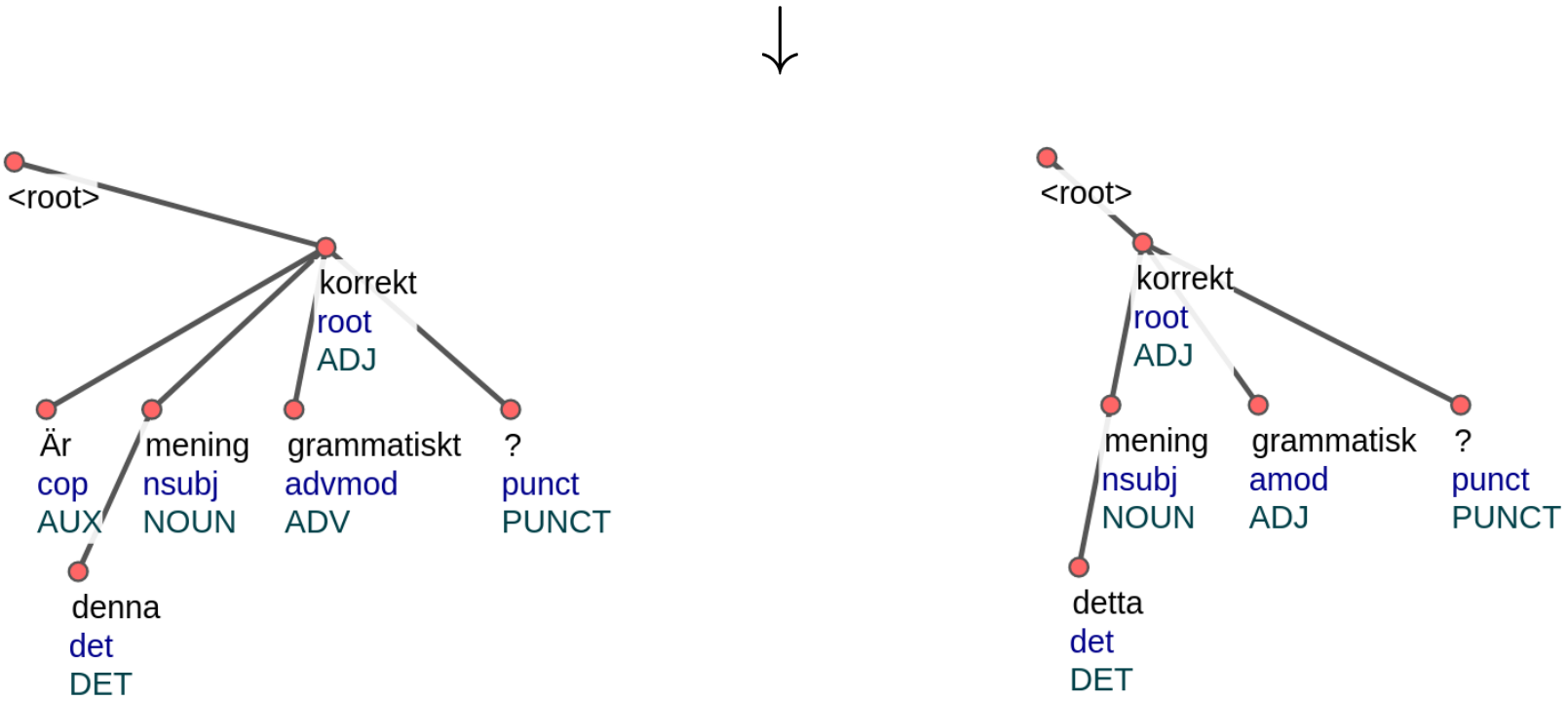
Given a learner sentence:

1. obtain correction hypothesis
2. **annotate learner sentence and correction in UD**
3. extract error patterns
4. generate feedback comments

L2 parsing

2. UD annotation

⟨“Är denna mening grammatiskt korrekt?”, “detta mening korrekt grammatisk?”⟩



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**)

Generating synthetic errors

Synthetic Error-Augmented Parsing of Swedish as a Second Language: Experiments with Word Order

Arianna Masciolini, Emilie Marie Carreau Francis, Maria Irena Szawerna
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
{arianna.masciolini, emilie.francis, maria.szawerna}@gu.se

In Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024, Torino, Italy, 2024

Using correction hypotheses

Bootstrapping the Annotation of UD Learner Treebanks

Arianna Masciolini
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
arianna.masciolini@gu.se

In Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024, Torino, Italy, 2024

Steps

Given a learner sentence:

1. obtain correction hypothesis
2. annotate learner sentence and correction in UD
3. **extract error patterns**
4. generate feedback comments

Errors, patterns and queries

Two related problems

1. how to ***retrieve*** specific (error) patterns from L1-L2 and, in general, parallel treebanks?
2. how to ***extract*** error patterns from one or more L1-L2 sentence pairs?

Error retrieval

A query engine for L1-L2 parallel dependency treebanks

Arianna Masciolini
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
arianna.masciolini@gu.se

In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), Torshavn, Faroe Islands, 2023

Error retrieval

STUnD: ett Sökverktyg för Tvåspråkiga Universal Dependencies-trädbanker

Arianna Masciolini[1,†], Márton A. Tóth[1,†]

[1] *Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs Universitet, Sverige*

In Proceedings of the Huminfra Conference (HiC 2024), Gothenburg, Sweden, 2024

Error retrieval

Exploring parallel corpora with STUnD: a Search Tool for Universal Dependencies

Arianna Masciolini[1], Herbert Lange[1,*] and Márton András Tóth[1,*]

[1]*Department of Swedish, Multilingualism, Language Technology; University of Gothenburg, Sweden*

In the upcoming Huminfra handbook

STUnD

Treebank 1 Browse... it_thvalico-ud-test.conllu

Treebank 2 (optional) Browse... it_valico-ud-test.conllu

Query

FEATS_ "Gender={Fem->Masc}"

Replacement

additional replacement rule (optional)

Mode (•) text ( ) CoNLL-U ( ) tree

Options [ ] context [ ] diff

Reset Search

76 hits - save: T1 file T2 file parallel file

1	A **la** fine ha spiegato a l' uomo che l' aveva liberata che l' altra persona distesa su il terreno era il suo amore e che non gli aveva chiesto niente .	A la fine ha spigato a **l'** uomo che l' aveva liberata , che l' altra personna distesa su il terreno era il suo amore e che non gli aveva chiesto niente .	1
2	Ieri a il parco , stavo camminando in la parte **più vuota** di il parco , perché avevo bisogno di solitudine , quando ho visto una cosa strana .	Ieri a il parco , ero camminando in la parte **lo più vuoto** di il parco , perchè avevo bisogno di solitudine , quando ho visto una cosa strana .	2
3	Un altro uomo si trovava lì , seduto su **una** panchina di il parco , leggendo un giornale con i suoi occhiali .	Un altra uomo , si trova li , seduto su **il** un panchino di il parco , leggendo un giornale con i suoi occhiali .	3
4	Un altro uomo si trovava lì , seduto **su una panchina di il parco** , leggendo un giornale con i suoi occhiali .	Un altra uomo , si trova li , seduto **su il un panchino di il parco** , leggendo un giornale con i suoi occhiali .	4
5	Ieri a il parco si trovava Giulio , un uomo che tutti i giorni a il pomeriggio leggeva il giornale , lì seduto tranquillamente su una panca bianca **con decorazioni in stile di il '700** , stava in il mezzo di il parco dove si vedevano soltanto gli alberi e i fiori , nessun edificio si poteva vedere da lì , forse Giulio sceglieva questo bel posto per dimenticar si un po' di la invasione moderna e globalistica di la città in cui viveva , ma questo non lo sappiamo con sicurezza quindi lo lasciamo da parte e proseguiamo con la nostra storia di questo abitudinario Giulio , il nostro protagonista da l' aspetto comune ma molto simpatico .	Ieri a il parco si trovava Giulio , un uomo che tutti i giorni a il pomeriggio leggeva il giornale , lì seduto tranquilamente su una banca bianca **con arredamnti stilo di il 700** , stava in il mezzo di il parco dove soltando si vedeva l' alberi e le fiori , nessun edificio si poteva vedere da lì , forse Giulio sceglieva questo bel posto per dimnticar si un pò Di la invasione moderna e globalistica di la città in cui viveva , ma questo non lo sappiamo con sicurezza quindi lo lasciamo da un' altra parte e prosiguiamo con la Nostra storia di questo routinario Giulio , il nostro protagonista di un aspetto comune ma Molto simpatico .	5
	Li aveva cominciati a seguire fino a la fine di il parco , dove aveva visto una mazza da baseball che era portata da un bambino felice ; lui si è avvicinato a il bambino e gli ha rubato **la** sua mazza ; povero bambino piangeva molto , non si sapeva se piangeva una bambina o un bambino ; subito dopo dietro a l' uomo robusto , Giulio , con la mazza e aiutato da una forza terribile , lo ha colpito a tal punto che l' uomo è caduto a piombo su il pavimento , ahi !	Gli aveva cominciato a seguire fino a la fine di il parco dove aveva visto un bat di base ball il quale era portato da un bambino felice , lui si è avvicinato a il bambino e l' ha rubato **il** suo bat , povero bambino pianggiava molto , non si sapeva se pianggiava una bambina o un bambino , subito dopo indietro a il uomo robusto Giulio con il bat e aiutato da una forza terribile gli ha attaccato a tal punto che l' uomo era caduto a piombo su il suolo pavimentato ouuch !	

STUnD

Treebank 1 Browse... en_pud-ud-test.conllu

Treebank 2 (optional) Browse... sv_pud-ud-test.conllu

Query

```
TREE_ (FEATS_ "VerbForm={Part->Sup}") [AND [LEMMA "{have->ha}", FEATS_ "Tense=Pres"]]
```

Replacement

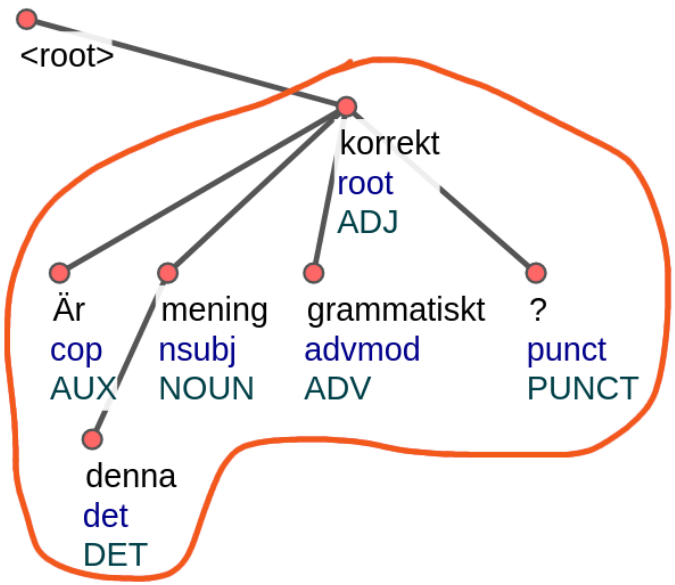
```
CHANGES [FILTER_SUBTREES TRUE (OR [DEPREL_ "aux", DEPREL_ "cop"]), PRUNE TRUE 1]
```

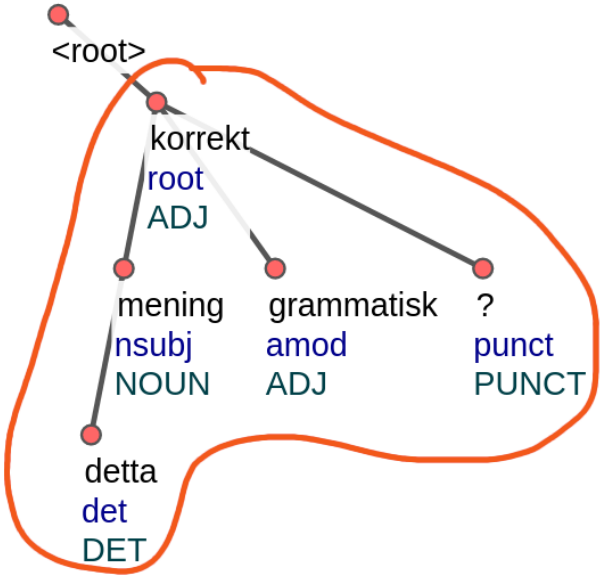
Mode ◉ text ○ CoNLL-U ○ tree ☐ highlight discrepancies search reset

63 hits - save: T1 file T2 file parallel file

That share **has been rising** steadily over the years — only 11 percent of the total vote was cast before Election Day in 1996 , according to the Census Bureau -- and seems likely to jump again this year .	Den andelen **har ökat** stadigt med åren – bara 11 procent av de samlade rösterna lades före valdagen 1996 enligt folkbokföringsbyrån – och det verkar troligt att den kommer öka ordentligt igen .
“ We **’ve requested** other nations to help us populate the zoo with different species of animals , including a pig , ” Saqib said .	” Vi **har bett** andra stater att hjälpa oss befolka djurparken med olika djurarter , inklusive en gris ” , sade Saqib .
Several analysts **have suggested** Huawei is best placed to benefit from Samsung 's setback .	Flera analytiker **har föreslagit** att Huawei har bäst position för att tjäna på Samsungs tillbakagång .
The 10 - week course **has been** " **certified** " by UK spy agency GCHQ .	10-veckorskursen **har** ” **certifierats** ” av brittiska spionmyndigheten GCHQ .
Throughout history , the international hair market **has** always **had** a political dimension , says Tarlo .	Genom historien **har** den internationella hårmarknaden alltid **haft** en politisk dimension , säger Tarlo .
Shenzhen 's traffic police **have opted** for unconventional penalties before .	Shenzhens trafikpolis **har valt** okonventionella straff förut .
Seagal , whose grandmother was from Vladivostok in Russia 's far east , **has made** frequent trips to Russia in recent years and visited Kamchatka and Sakhalin in September .	Seagal , vars farmor kom från Vladivostok i Rysslands östligaste delar **har gjort** flera resor till Ryssland de senaste åren och besökte Kamtjatka och Sakhalin i september .
Researchers **have been investigating** potential for male hormonal contraceptives for around 20 years .	Forskare **har undersökt** potentialen för manliga hormonella preventivmedel i ungefär 20 år .
Ms Pugh **has received** treatment at Papworth and Addenbrooke 's Hospitals in Cambridgeshire .	Ms Pugh **har fått** behandling vid Papworths och Addenbrookes sjukhus i Cambridgeshire .
Students like Rai have been meeting with counsellors at the school to talk about what	Studenter som Rai har träffat kuratorer på skolan för att prata om det som hände , men

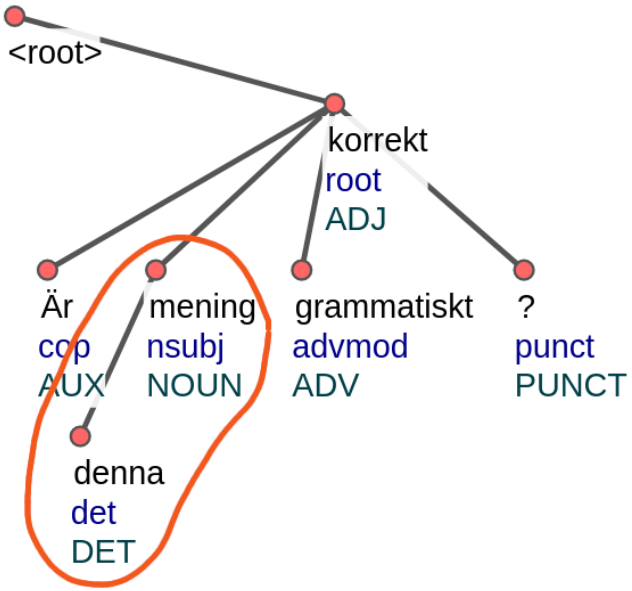
3.1 Error-correction pairs

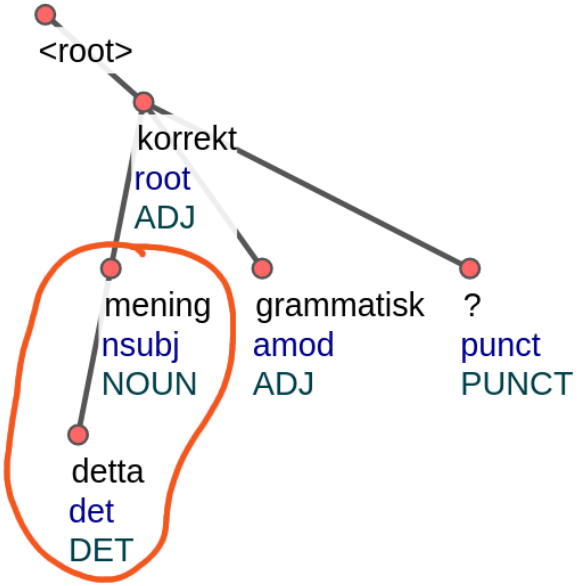




L1: “Är denna mening grammatiskt korrekt?” — L2: “detta mening korrekt grammatisk?”

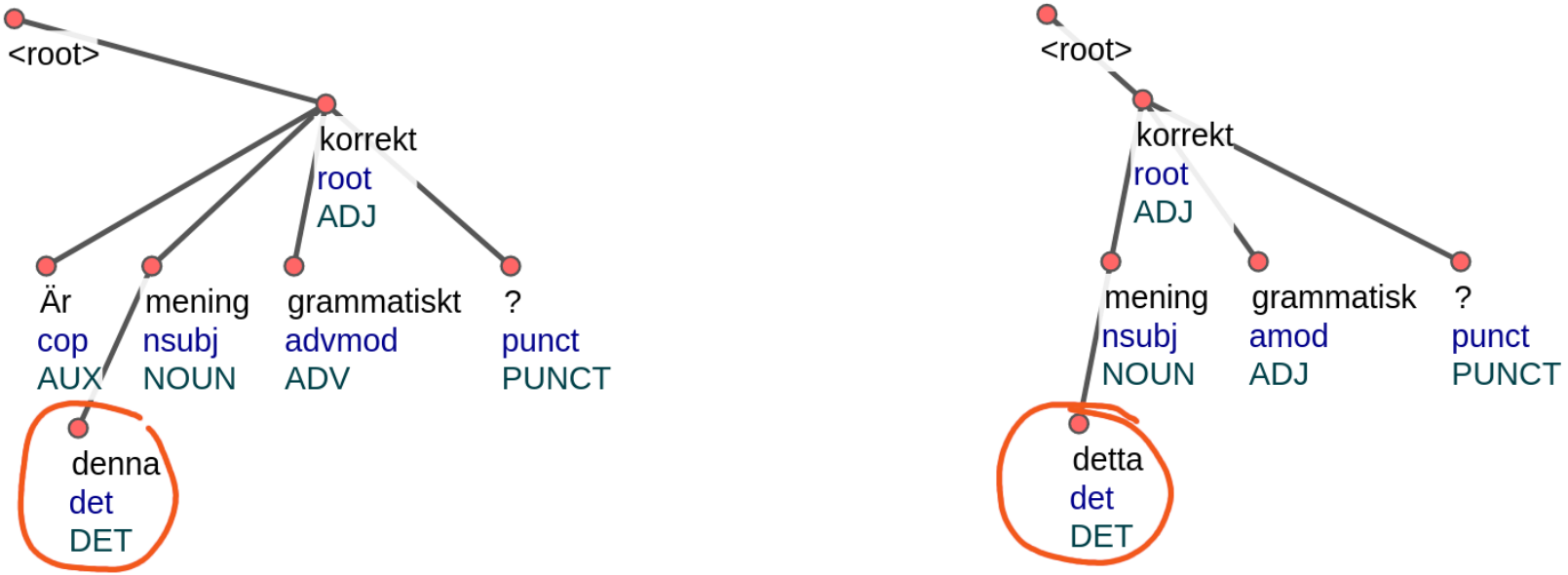
3.1 Error-correction pairs





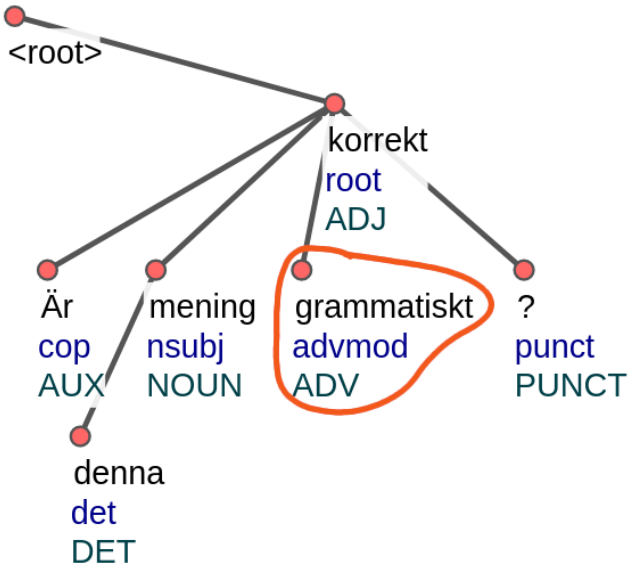
L1: “denna mening” — L2: “detta mening”

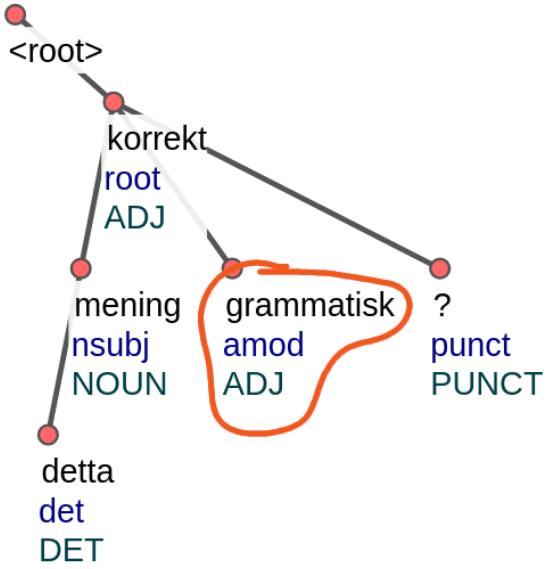
3.1 Error-correction pairs



L1: “denna” — L2: “detta”

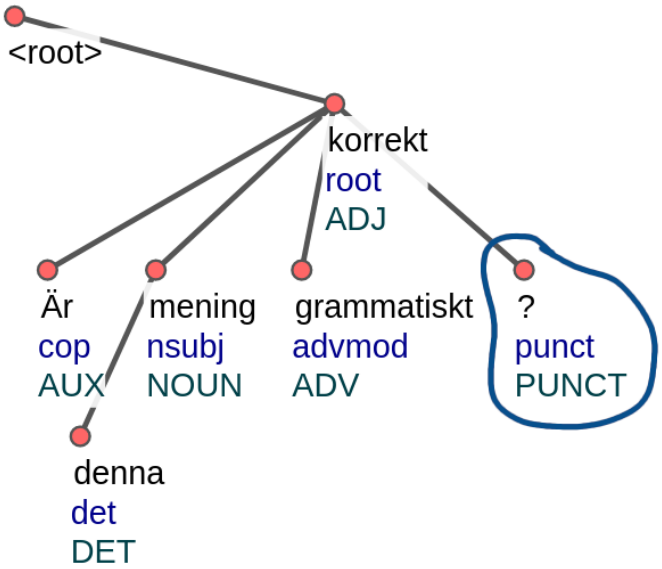
3.1 Error-correction pairs

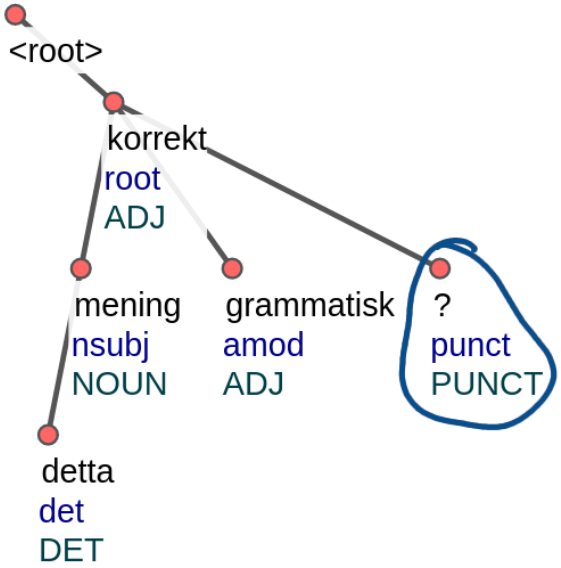




L1: “grammatiskt” — L2: “grammatisk”

3.1 Error-correction pairs



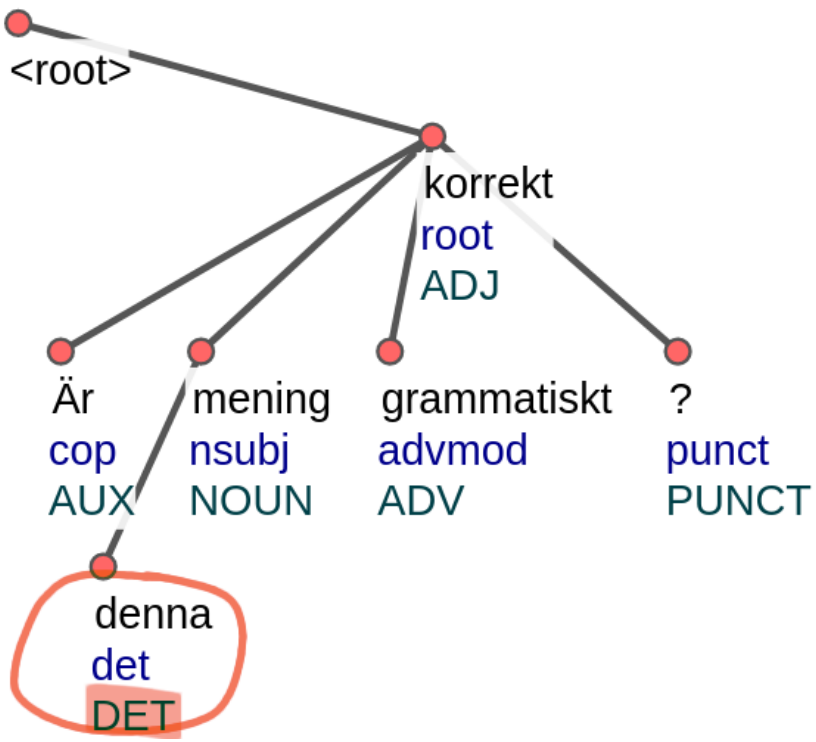


L1: “?” — L2: “?”

UD patterns in gf-ud

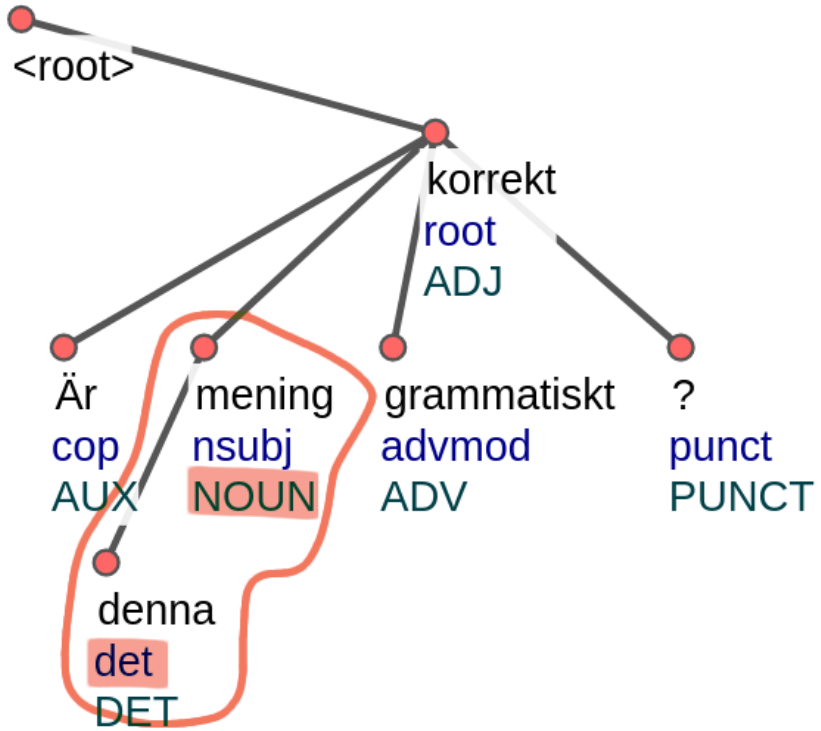
pattern type	**example**
single-token patterns	`POS "DET"`
tree patterns	`TREE (POS "NOUN") [DEPREL "det"]`
sequence patterns	`SEQUENCE [POS "DET", POS "NOUN"]`
logical operators	`AND [POS "NOUN", DEPREL "nsubj"]`

UD patterns in gf-ud



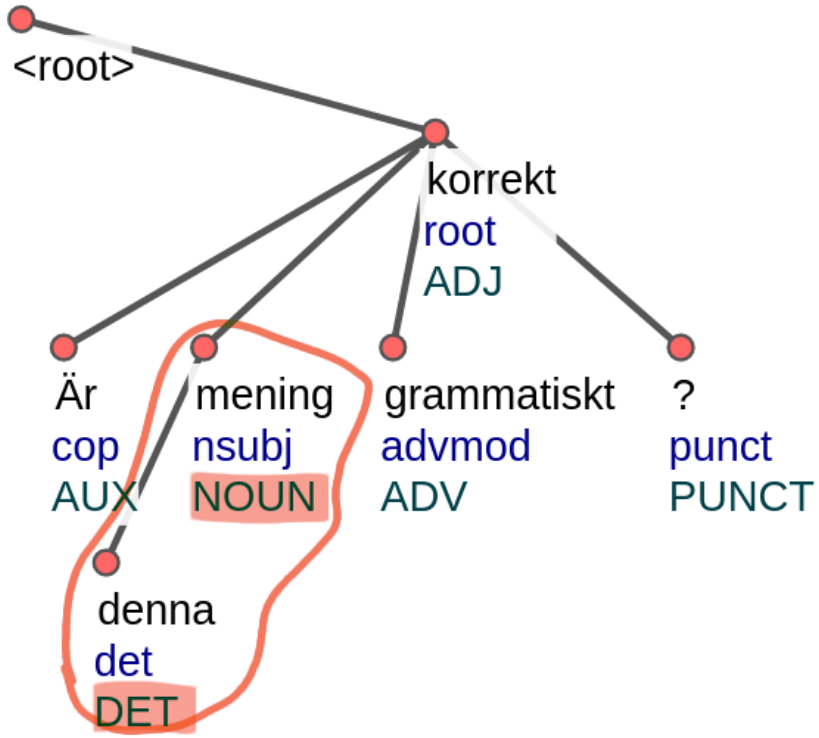
POS “DET”

UD patterns in gf-ud



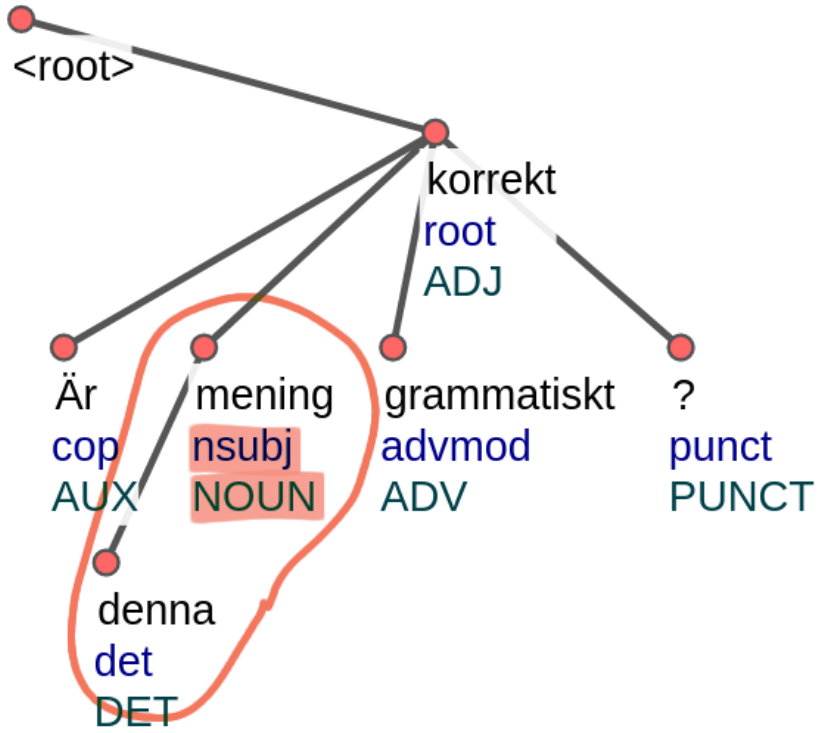
TREE (POS “NOUN”) [DEPREL “det”]

UD patterns in gf-ud



SEQUENCE [POS “DET”, POS “NOUN”]

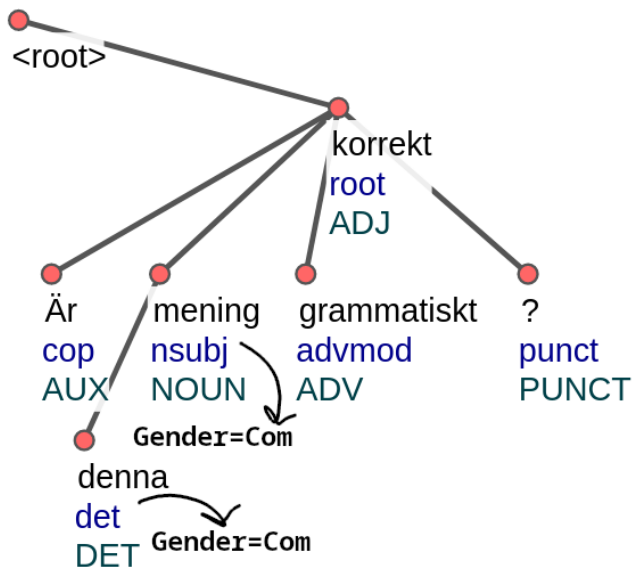
UD patterns in gf-ud

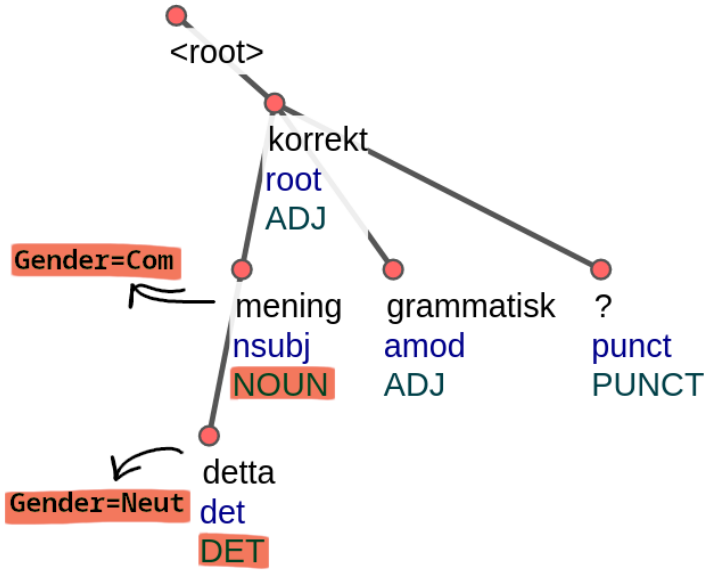


AND [POS “NOUN”, DEPREL “nsubj”]

L1-L2 UD patterns

Many errors can be represented as UD patterns describing the L2

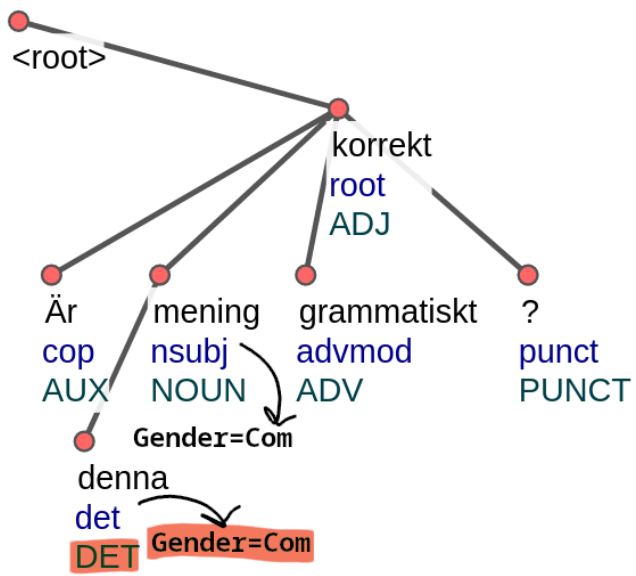


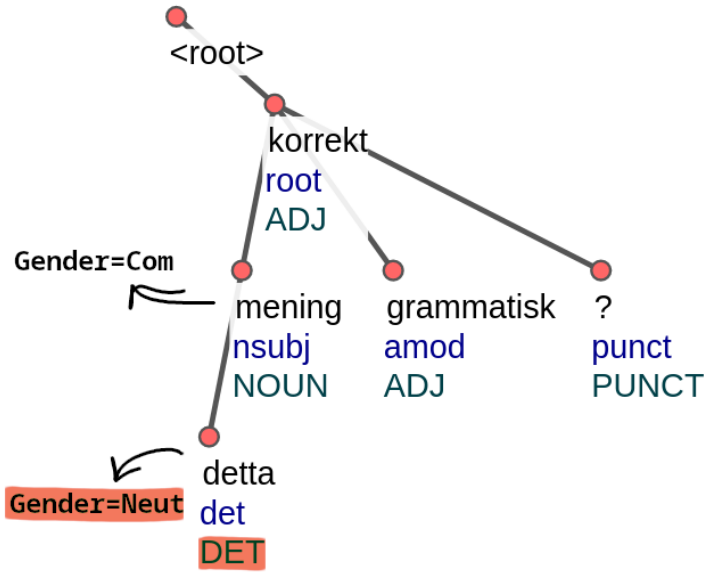


TREE (AND [POS “NOUN”, FEATS_ “Gender=Com”]) [AND [POS “DET”, FEATS_ “Gender=Neutr”]]

L1-L2 UD patterns

Sometimes, it is useful (or even necessary) to compare the L1 and L2 → **L1-L2 patterns** (pairs of UD patterns)





⟨ AND [POS "DET", FEATS_ "Gender=Com"], AND [POS "DET", FEATS_ "Gender=Neutr"] ⟩

Pattern extraction

Towards automatically extracting morphosyntactical error patterns from L1-L2 parallel dependency treebanks

Arianna Masciolini and **Elena Volodina** and **Dana Dannélls**
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
firstname.lastname@gu.se

In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Toronto, Canada, 2023

Steps

Given a learner sentence:

1. obtain correction hypothesis
2. annotate learner sentence and correction in UD
3. extract error patterns
4. **generate feedback comments**

Feedback

Feedback in CALL

"Är detta mening grammatiskt korrekt?"

type	**example**
correct/incorrect	Try again!
correct answer	Är denna mening grammatiskt korrekt?
highlighting	Är detta mening grammatiskt korrekt?
metalinguistic	Pay attention to gender agreement!
example	Detta är en exempelmening → Denna är en exempelmening
error label	`M-Gend`

. . . or any combination of the above!

~~Are feedback comments useful?~~

Some more useful questions:

- *what kind* of feedback is useful?
- in *which cases*?
- *how* should it be used?

. . . a flexible, general-purpose way to automatically generate feedback comments can be a tool to answer these questions!

4. Feedback Comment Generation

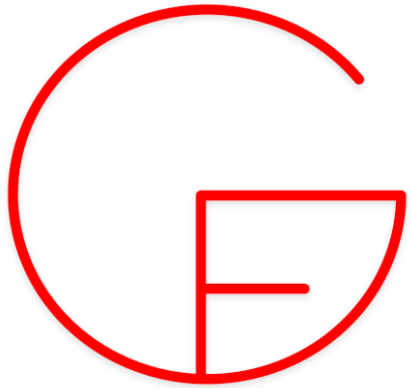
. . . (far) future work! Some (less vague) ideas:

- data2text task
  - error patterns $\rightarrow$ feedback comments, ideally:
    - in multiple languages
    - adjustable to the learner's level

$\downarrow$

idea: **a GF CNL**

Grammatical Framework 101



A generative grammar formalism/programming language for **multilingual grammar engineering**:

- GF grammar = 1 *abstract syntax* + *n concrete syntaxes*
- especially well suited for defining *application grammars*
- interoperable with UD (does that help?)

FCG with GF

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.

FCG with GF

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!

FCG with GF

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!

To sum up

Status update

past	**present**	**future**
- “lazy” L2 parsing experiments - basic pattern extraction - query engine	- SweLL-based L2 Swedish treebank - better pattern extraction - MultiGEC shared task	- parsing model for L2 Swedish - (feedback comment generation)

Some discussion points

- ideas & feedback about the ideas for feedback
- practical things about what I should do with my thesis
- whatever you want

Fika!