A query engine for L1-L2 parallel dependency treebanks

NoDaLiDa 2023

Arianna Masciolini

Språkbanken Text, University of Gothenburg

23.05.2023

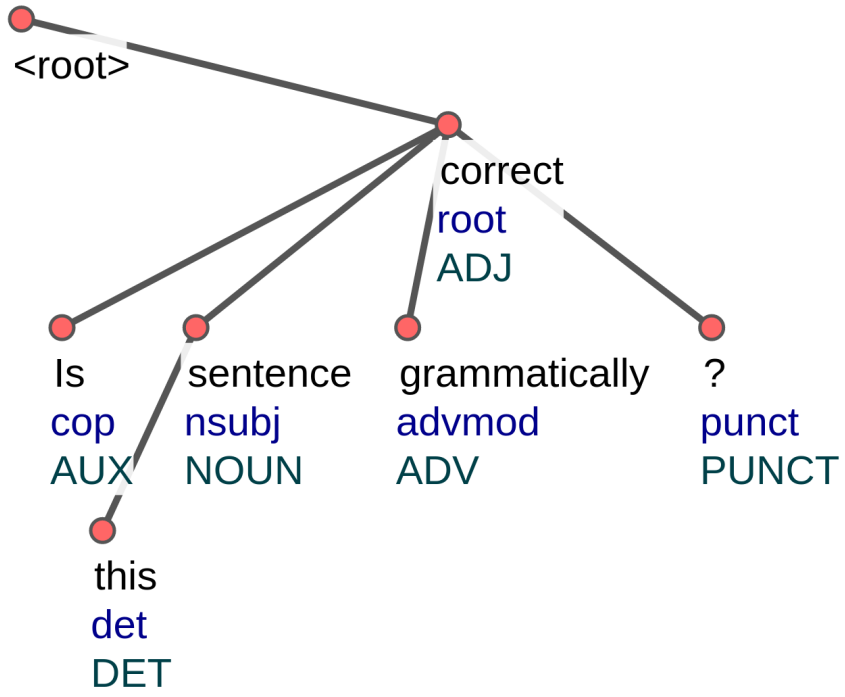
L1-L2 treebanks

- learner sentences ‖ correction hypotheses
- no error labelling, just morphosyntactical annotation
- main design goal: **interoperability**

Handcrafted L1-L2 treebanks

name	language	n. sentences
TLE/ESL	English	5124
CFL	Chinese	451
VALICO-UD	Italian	398

Universal Dependencies 101



“Is this sentence grammatically correct?”

Universal Dependencies 101

```
# text = Is this sentence grammatically correct?

1   Is  be  AUX VBZ Mood=Ind|Number=Sing|Person=3|...   5   cop _   _
2   this    this    DET DT  Number=Sing|PronType=Dem    3   det _   _
3   sentence    sentence    NOUN    NN  Number=Sing 5   nsubj   _   _
4   grammatically   grammatically   ADV RB  _   5   advmod  _   _
5   correct correct ADJ JJ  Degree=Pos  0   root    _   _
6   ?   ?   PUNCT   .   _   5   punct   _   _
```

ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC

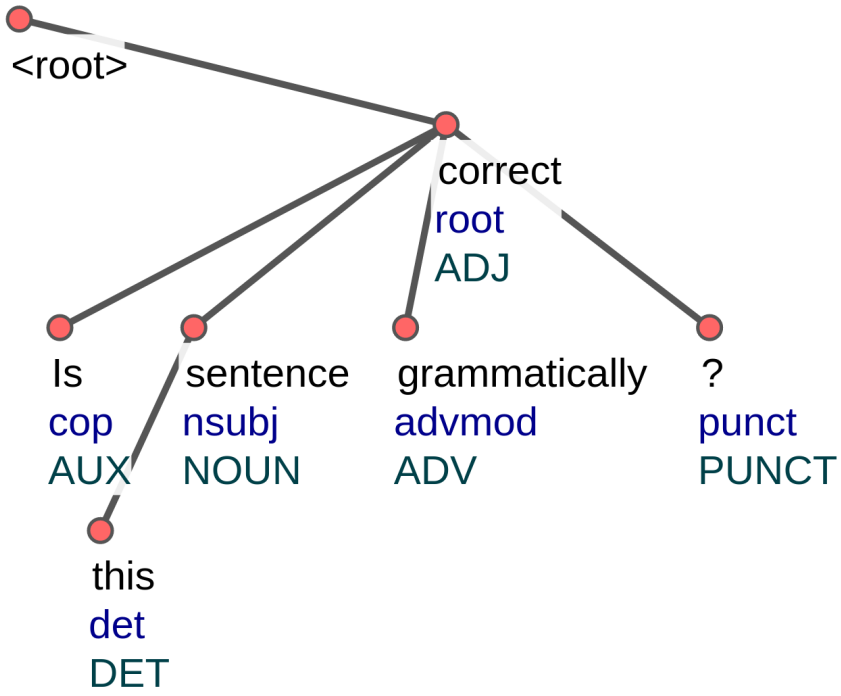
Universal Dependencies 101

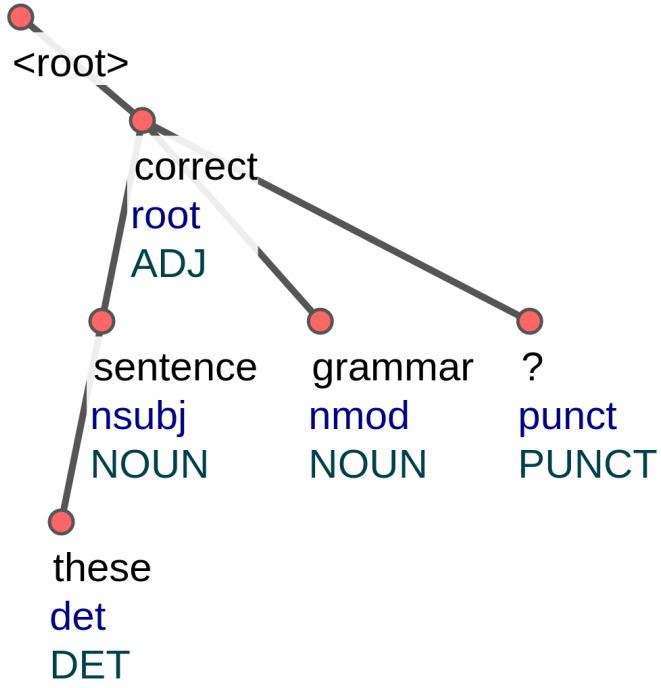
```
# text = Is this sentence grammatically correct?

1   Is  be  AUX VBZ Mood=Ind|Number=Sing|Person=3|...   5   cop _   _
2   this    this    DET DT  Number=Sing|PronType=Dem    3   det _   _
3   sentence    sentence    NOUN    NN  Number=Sing 5   nsubj   _   _
4   grammatically   grammatically   ADV RB  _   5   advmod  _   _
5   correct correct ADJ JJ  Degree=Pos  0   root    _   _
6   ?   ?   PUNCT   .   _   5   punct   _   _
```

ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC

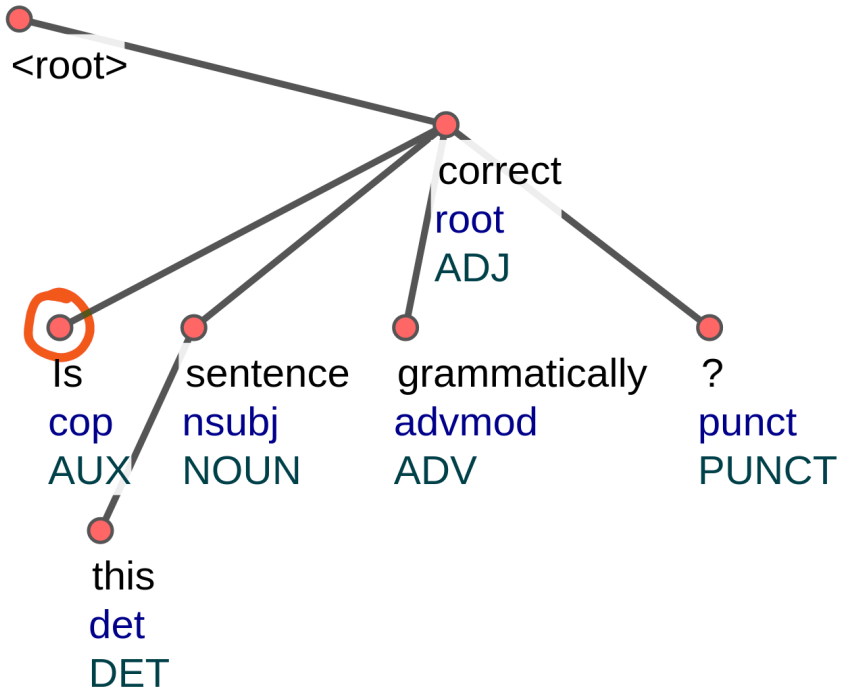
Example

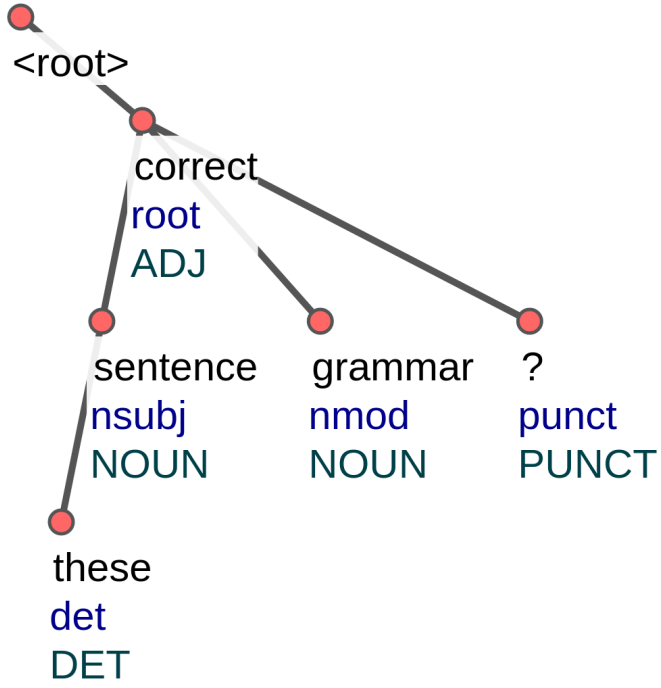




⟨“Is this sentence grammatically correct?” , “these sentence correct grammar?”⟩

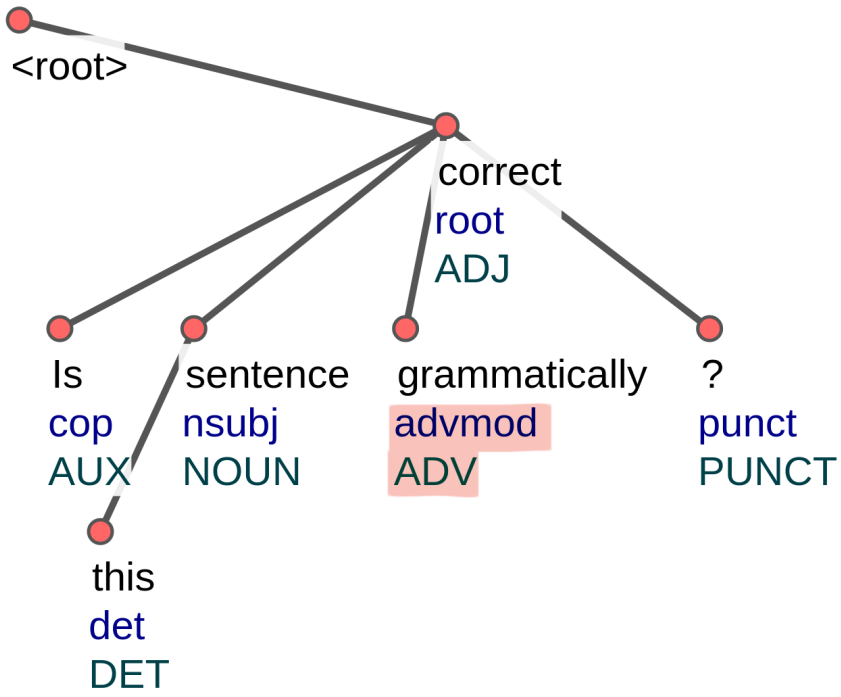
Example

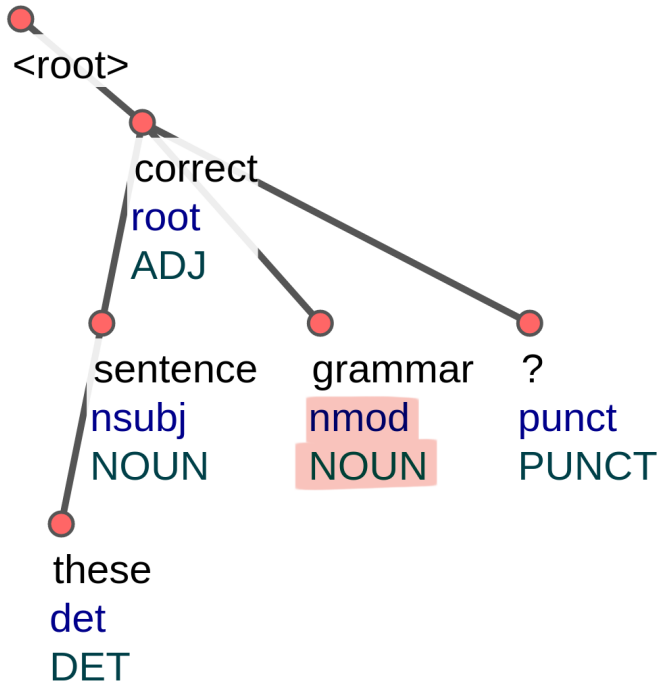




⟨"**Is** this sentence grammatically correct?","_ these sentence correct grammar?"⟩

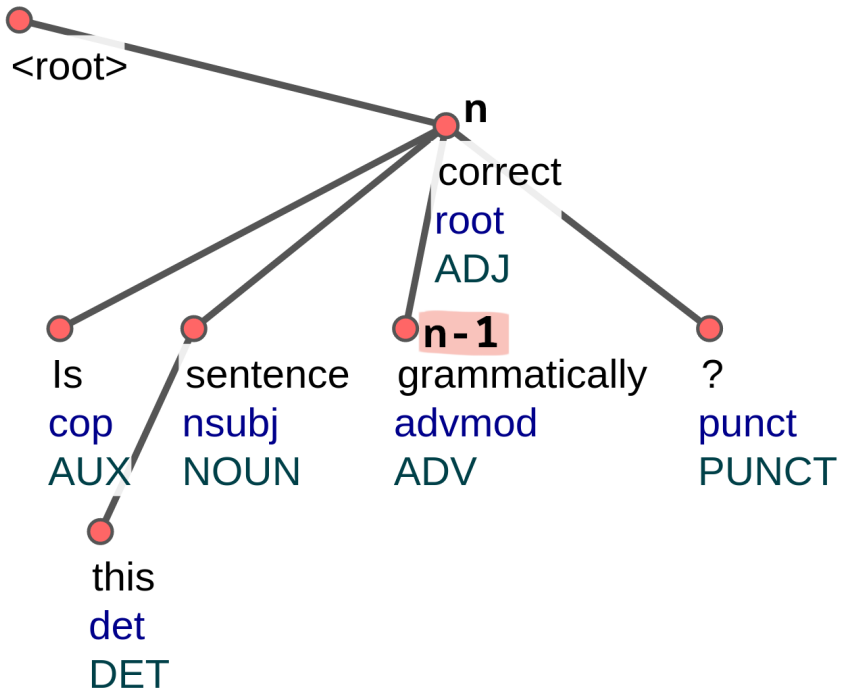
Example

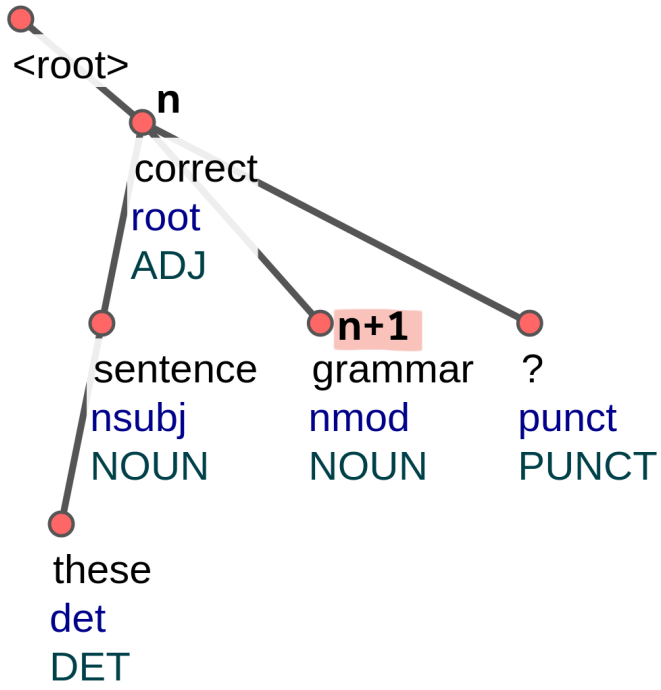




⟨"Is this sentence *grammatically* correct?" , "these sentence correct grammar?"⟩

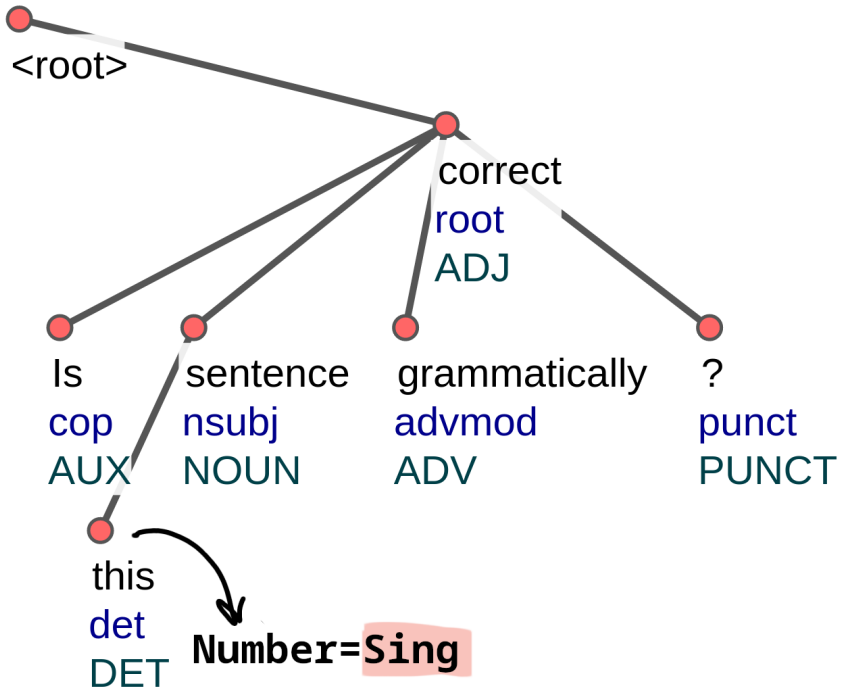
Example

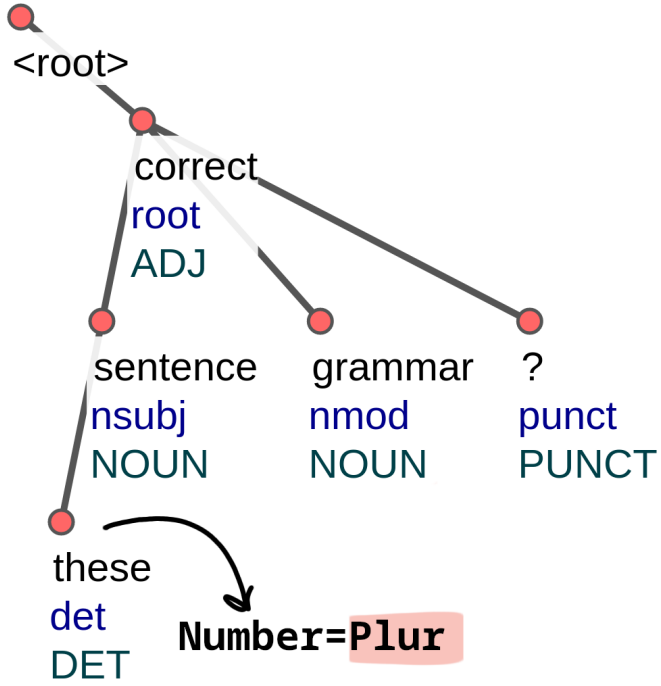




⟨"Is this sentence *grammatically correct*?" , "these sentence correct grammar?"⟩

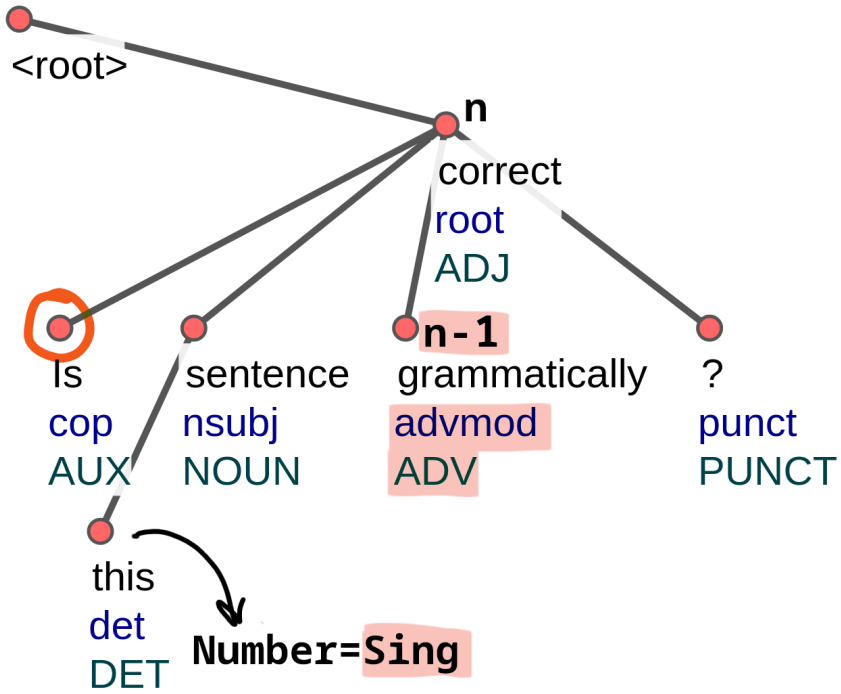
Example

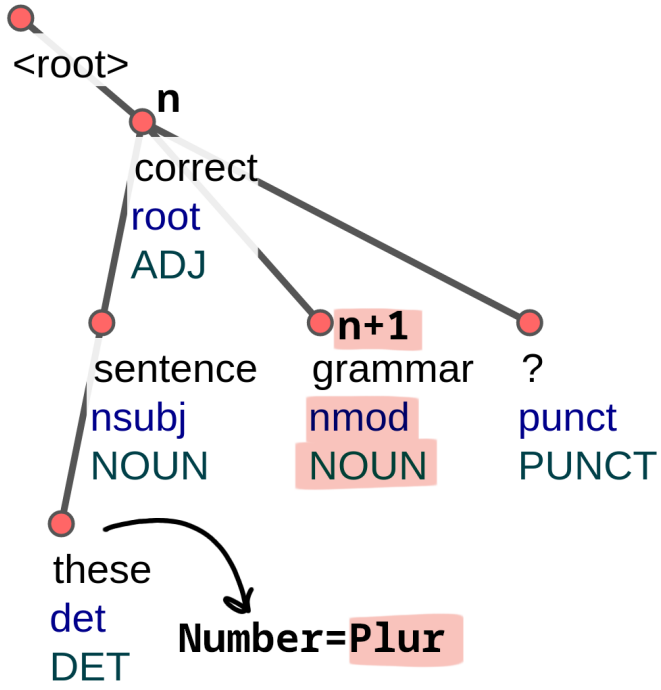




⟨"Is *this* sentence grammatically correct?" , "these sentence correct grammar?"⟩

Example





⟨"**Is** *this* sentence *grammatically correct*?","_ these sentence correct grammar?"⟩

Error retrieval

A major function of a learner corpus is to facilitate retrieval of sentences with specific errors. [. . .] In view of the limitations of error tags described above, we propose the use of L1-L2 parallel treebank for learner error retrieval. A ***search query on such a treebank, consisting of a pair of parse tree patterns with alignments****, can be viewed as a dynamically defined error category.*[1]

[1] Lee et al., 2017. *L1-L2 parallel dependency treebank as learner corpus*

The ESL treebank query engine

Query DET NOUN Corpus ESL

Native Language Any

Error Agreement determiner

☐ Highlight errors

☐ Show corrections

Search

Instructions

Search for *sequences* of words, universal/PTB POS tags and relation labels. Regular expressions are supported for searching words.

Examples

- *see it* matches the string "see it"
- *see DET NOUN* matches "see that show", "see the sign", etc.
- *\w+ing something* matches "seeing something", "seeking something", etc.
- *amod NNS* matches adjectival modifier followed by a plural noun, such as "best cakes", "bigger halls", etc.

ESL filters and highliting

Filter query results to sentences with a specific grammatical error and/or specific native language.
An empty query will retrieve all the sentences that correspond to the specified filters.
Highlight grammatical errors and show annotations of sentence corrections using the checkboxes.

Corpus (UD v2.3)

- *ESL* is the Treebank of Learner English
- *English* is the EWT UD corpus

---

http://web.archive.org/web/20220120204838/http://esltreebank.org/

The ESL treebank query engine

Query DET NOUN Corpus ESL

Native Language Any

Error Agreement determiner

☐ Highlight errors

☐ Show corrections

Search

Instructions

Search for *sequences* of words, universal/PTB POS tags and relation labels. Regular expressions are supported for searching words.

Examples

- *see it* matches the string "see it"
- *see DET NOUN* matches "see that show", "see the sign", etc.
- *\w+ing something* matches "seeing something", "seeking something", etc.
- *amod NNS* matches adjectival modifier followed by a plural noun, such as "best cakes", "bigger halls", etc.

ESL filters and highliting

Filter query results to sentences with a specific grammatical error and/or specific native language.
An empty query will retrieve all the sentences that correspond to the specified filters.
Highlight grammatical errors and show annotations of sentence corrections using the checkboxes.

Corpus (UD v2.3)

- *ESL* is the Treebank of Learner English
- *English* is the EWT UD corpus

---

http://web.archive.org/web/20220120204838/http://esltreebank.org/

The ESL treebank query engine

Query DET NOUN Corpus ESL

Native Language Any

Error Agreement determiner

☐ Highlight errors

☐ Show corrections

Search

Instructions

Search for *sequences* of words, universal/PTB POS tags and relation labels. Regular expressions are supported for searching words.

Examples

- *see it* matches the string "see it"
- *see DET NOUN* matches "see that show", "see the sign", etc.
- *\w+ing something* matches "seeing something", "seeking something", etc.
- *amod NNS* matches adjectival modifier followed by a plural noun, such as "best cakes", "bigger halls", etc.

ESL filters and highliting

Filter query results to sentences with a specific grammatical error and/or specific native language.
An empty query will retrieve all the sentences that correspond to the specified filters.
Highlight grammatical errors and show annotations of sentence corrections using the checkboxes.

Corpus (UD v2.3)

- *ESL* is the Treebank of Learner English
- *English* is the EWT UD corpus

---

http://web.archive.org/web/20220120204838/http://esltreebank.org/

The ESL treebank query engine

L2
Query DET NOUN Corpus ESL

Native Language Any

Error Agreement determiner

☐ Highlight errors

☐ Show corrections

Search

Instructions

Search for *sequences* of words, universal/PTB POS tags and relation labels. Regular expressions are supported for searching words.

Examples

- *see it* matches the string "see it"
- *see DET NOUN* matches "see that show", "see the sign", etc.
- *\w+ing something* matches "seeing something", "seeking something", etc.
- *amod NNS* matches adjectival modifier followed by a plural noun, such as "best cakes", "bigger halls", etc.

ESL filters and highliting

Filter query results to sentences with a specific grammatical error and/or specific native language.
An empty query will retrieve all the sentences that correspond to the specified filters.
Highlight grammatical errors and show annotations of sentence corrections using the checkboxes.

Corpus (UD v2.3)

- *ESL* is the Treebank of Learner English
- *English* is the EWT UD corpus

---

http://web.archive.org/web/20220120204838/http://esltreebank.org/

The ESL treebank query engine

Query DET NOUN Corpus ESL

Native Language Any

Error Agreement determiner

☐ Highlight errors

☐ Show corrections

Search

Instructions

Search for *sequences* of words, universal/PTB POS tags and relation labels. Regular expressions are supported for searching words.

Examples

- *see it* matches the string "see it"
- *see DET NOUN* matches "see that show", "see the sign", etc.
- *\w+ing something* matches "seeing something", "seeking something", etc.
- *amod NNS* matches adjectival modifier followed by a plural noun, such as "best cakes", "bigger halls", etc.

ESL filters and highliting

Filter query results to sentences with a specific grammatical error and/or specific native language.
An empty query will retrieve all the sentences that correspond to the specified filters.
Highlight grammatical errors and show annotations of sentence corrections using the checkboxes.

Corpus (UD v2.3)

- *ESL* is the Treebank of Learner English
- *English* is the EWT UD corpus

---

http://web.archive.org/web/20220120204838/http://esltreebank.org/

Desiderata for a new query engine

- **corpus-agnostic**
- **no underlying error taxonomy**
  - error retrieval via **tree *and* sequence queries**
- **parallel L1-L2 matching**

Desiderata for a new query engine

- **corpus-agnostic**
- **no underlying error taxonomy**
  - error retrieval via **tree *and* sequence queries**
- **parallel L1-L2 matching**
- **subsentence extraction** (error highlighting)

Query language

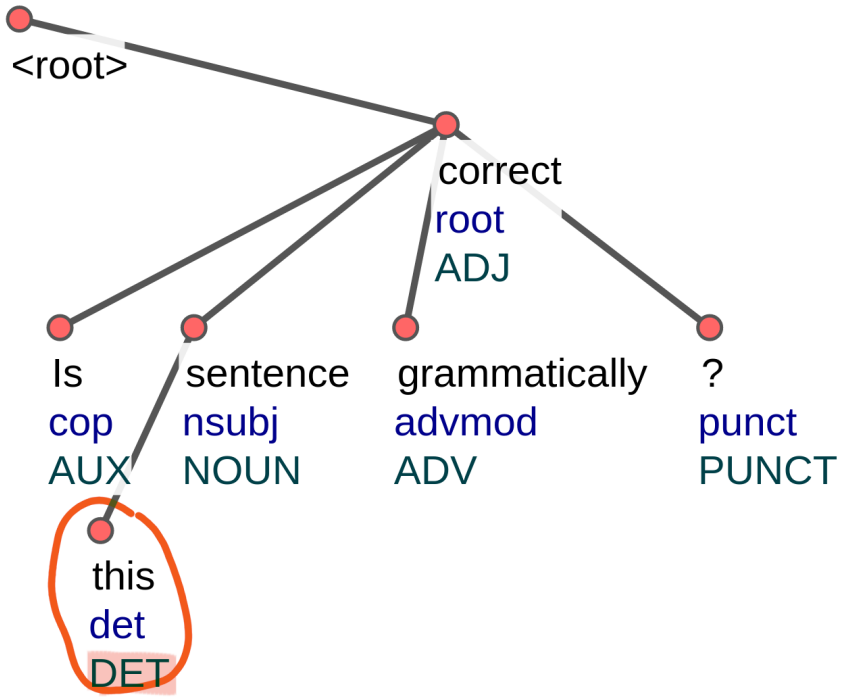
Query languages for UD trees

- several options to choose from
  - PML-TQ, Grew-match, UDAPI...
- decided on `gf-ud`'s embedded query language
  - sufficiently expressive and user-friendly
  - easy to use as a library

UD patterns in gf-ud

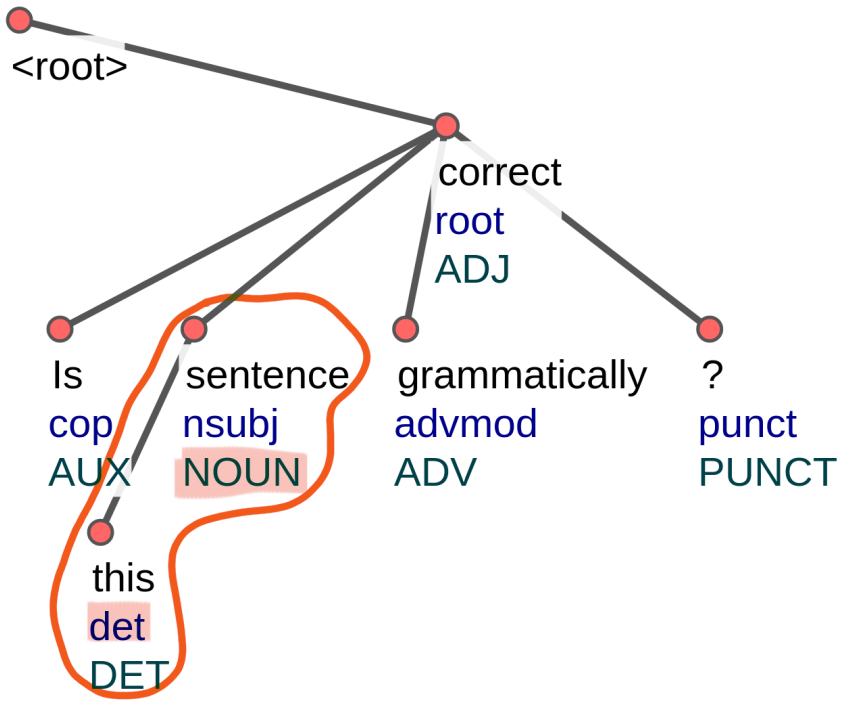
pattern type	example
single-token patterns	`POS "DET"`
tree patterns	`TREE (POS "NOUN") [DEPREL "det"]`
sequence patterns	`SEQUENCE [POS "DET", POS "NOUN"]`
logical operators	`AND [POS "NOUN", DEPREL "nsubj"]`

Single-token patterns



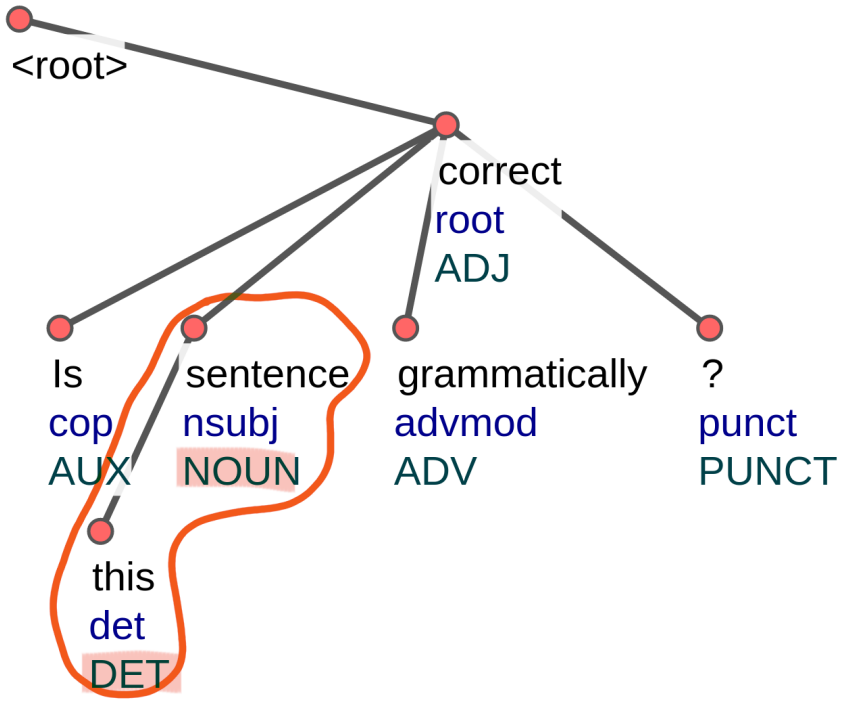

```
POS "DET"
```

Tree patterns



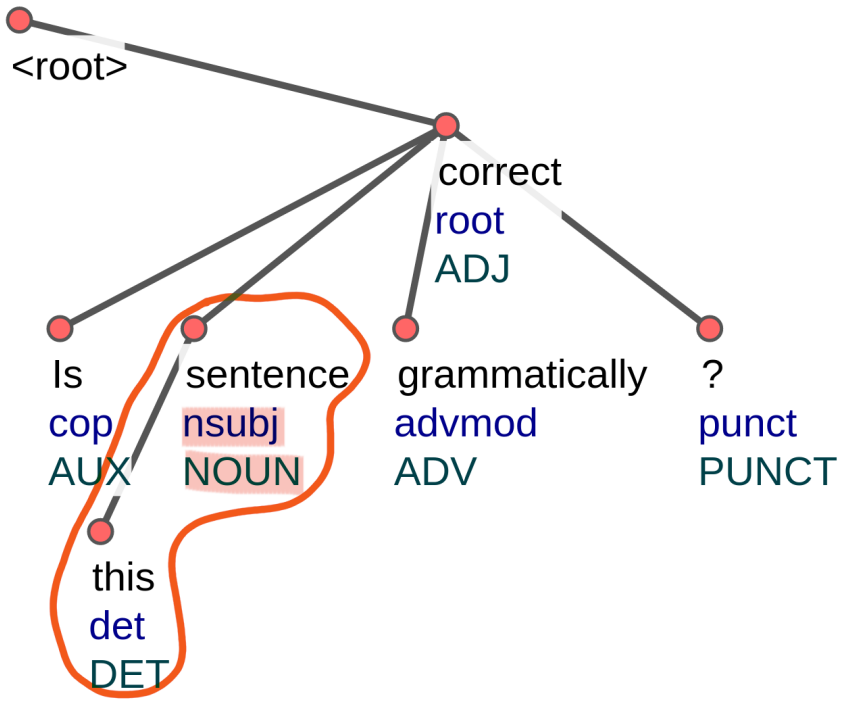

```
TREE (POS "NOUN") [DEPREL "det"]
```

Sequence patterns




```
SEQUENCE [POS "DET", POS "NOUN"]
```

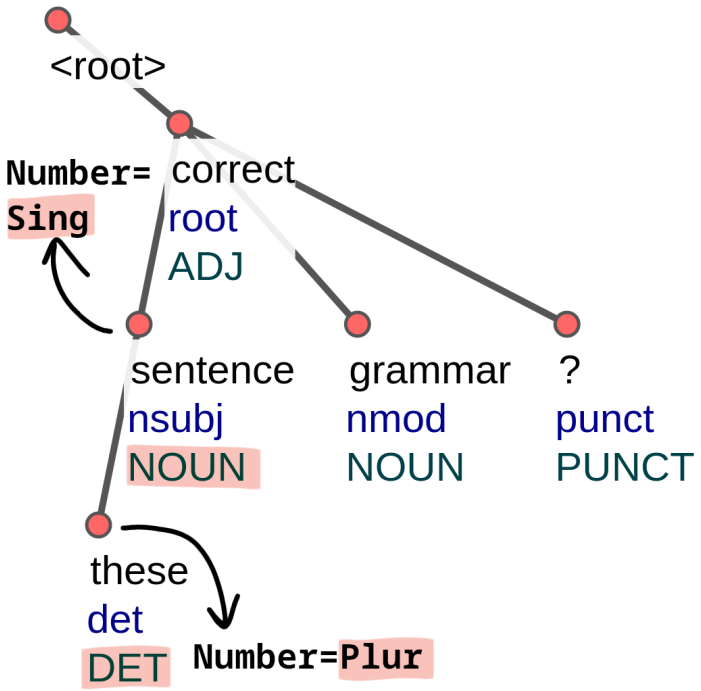
Logical operators




```
AND [POS "NOUN", DEPREL "nsubj"]
```

Error patterns

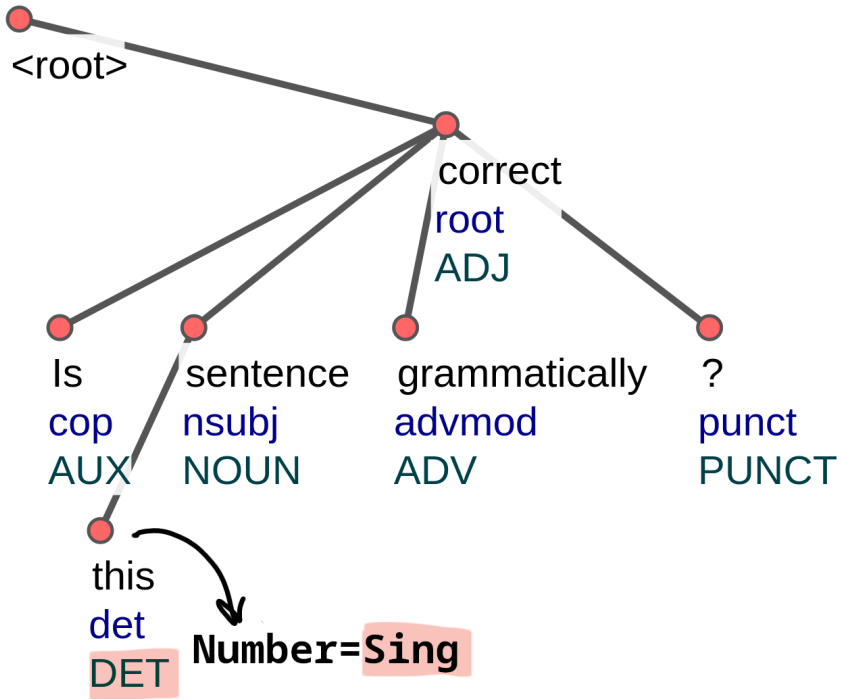
Many errors can be described by a single pattern describing the L2:

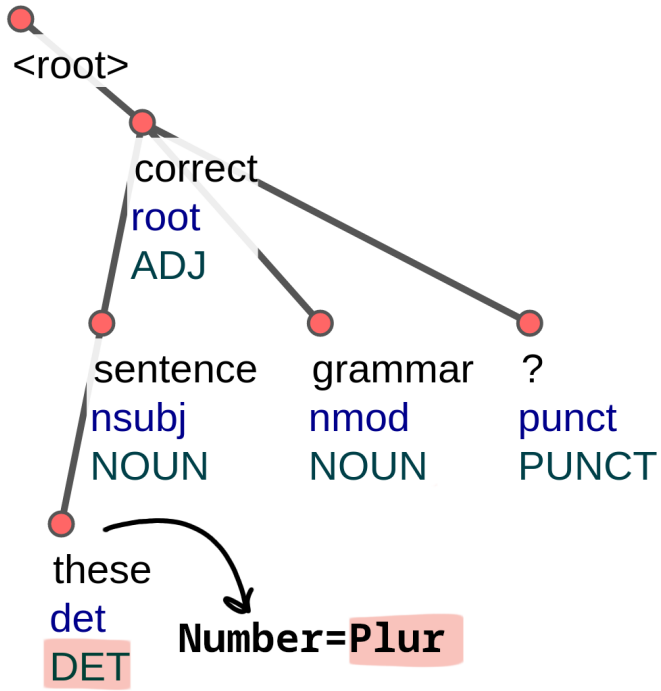



```
TREE (AND [POS "NOUN", FEATS "Number=Sing"]) [AND [POS "DET",
                                FEATS "Number=Plur"]]
```

Error patterns

. . . but often it is useful/necessary to specify them by comparison with the L1 → **L1-L2 patterns**





⟨AND [POS "DET", FEATS "Number=Sing"],AND [POS "DET", FEATS "Number=Plur"]⟩

L1-L2 patterns and extensions

Basic L1-L2 pattern

```
⟨AND [POS "DET", FEATS "Number=Sing"],
AND [POS "DET", FEATS "Number=Plur"]⟩
```

Arrow syntax

```
AND [POS "DET", FEATS "Number={Sing→Plur}"]
```

Variables

```
AND [POS "DET", FEATS "Number={$A →$B}"]
```

Sentence retrieval

A naïve approach

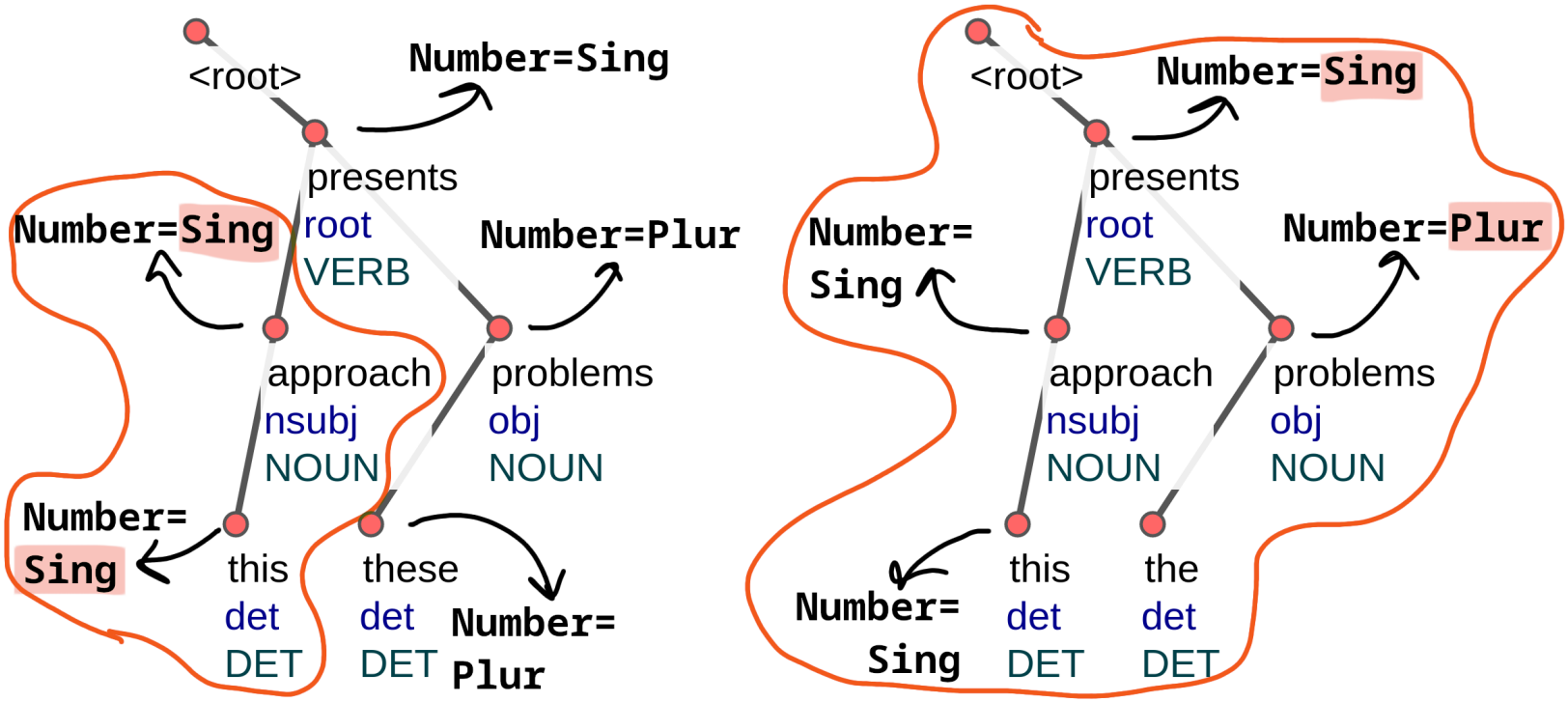
Given treebank and a query, return sentence pairs where

- a subtree of the L1 sentence matches its L1 part
- a subtree of the L2 sentence matches its L2 part

Problem: the matching L1-L2 subtrees may be semantically unrelated with each other $\rightarrow$ false positives

A naïve approach

TREE (FEATS "Number=$A") [FEATS "Number=$A →$B]



⟨"this approach presents *these* problems","this approach presents the problems"⟩

A better approach

Solution: recursively align L1-L2 sentence pairs with `concept-alignment`. Then match the query pattern nonrecursively on the resulting subtree pairs

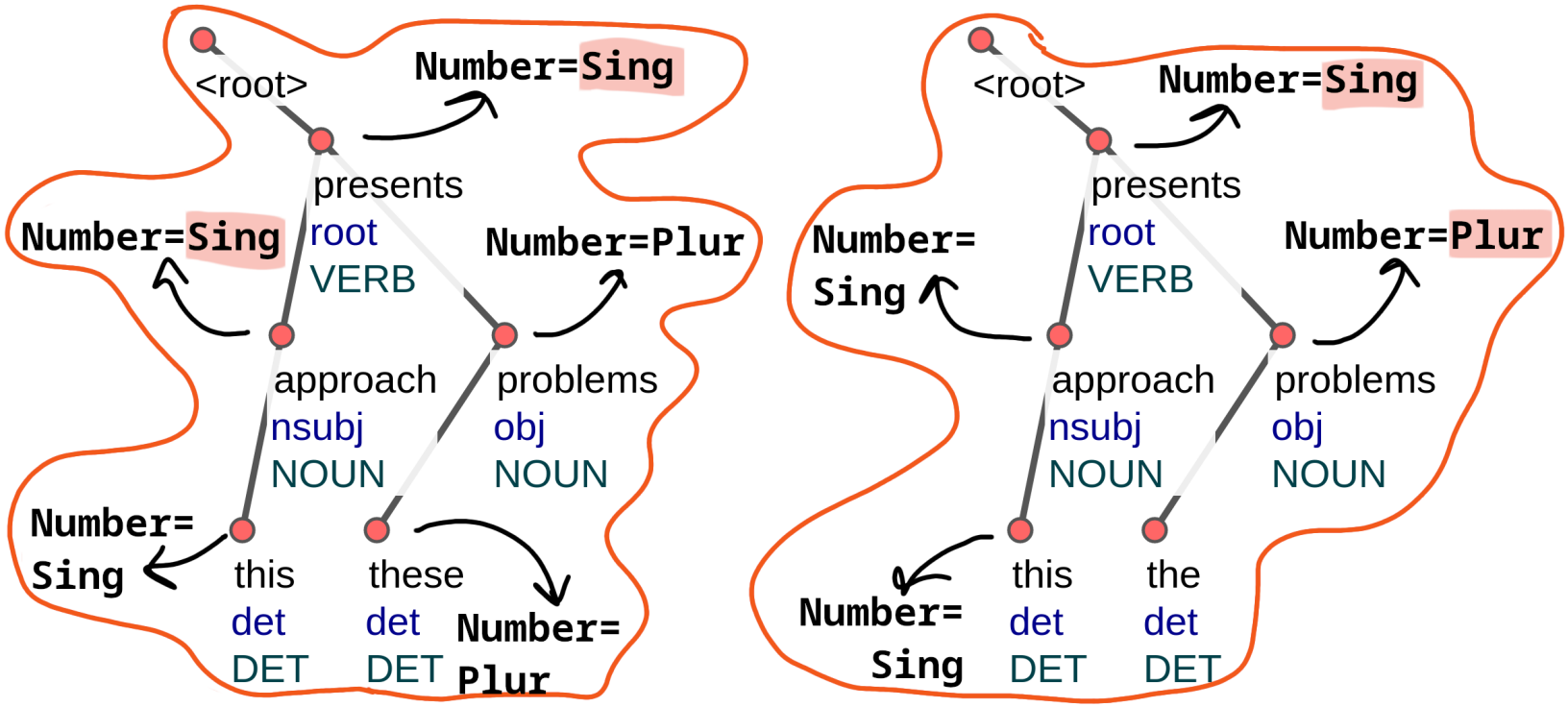
A better approach

Solution: recursively align L1-L2 sentence pairs with `concept-alignment`. Then match the query pattern nonrecursively on the resulting subtree pairs

Problem: dependents involved in the match may be semantically unrelated with each other $\rightarrow$ still some false positives

A better approach

TREE (FEATS "Number=$A") [FEATS "Number=$A →$B]



⟨"this approach presents *these* problems","this approach presents the problems"⟩

Our approach

Solutions: recursively check that dependents are also aligned with each other

Our approach

Solutions: recursively check that dependents are also aligned with each other

Given the query
`TREE (FEATS "Number=$A") [FEATS "Number=$A` $\rightarrow$`$B]`:

- ⟨"this approach presents *these* problems","this approach presents the problems"⟩ matches the pattern, but
- "this approach" is not aligned with "the problems"⟩

Therefore, the sentence does **not** contain a number agreement error.

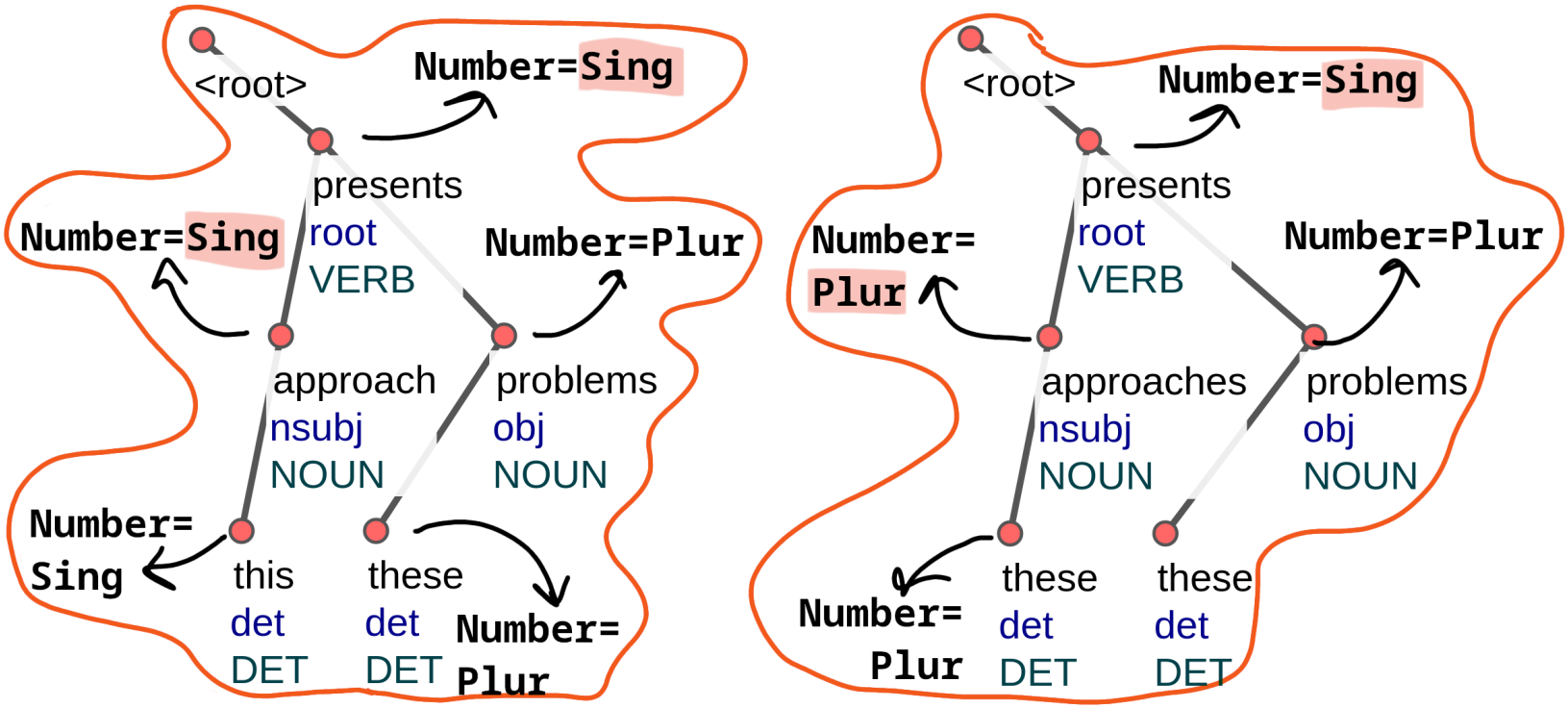
Extracting subsentences

When it comes “highlighting errors”

- easy to highlight the matched subtrees
- additional query-based pruning to deal with very large (sub)trees

Extracting subsentences

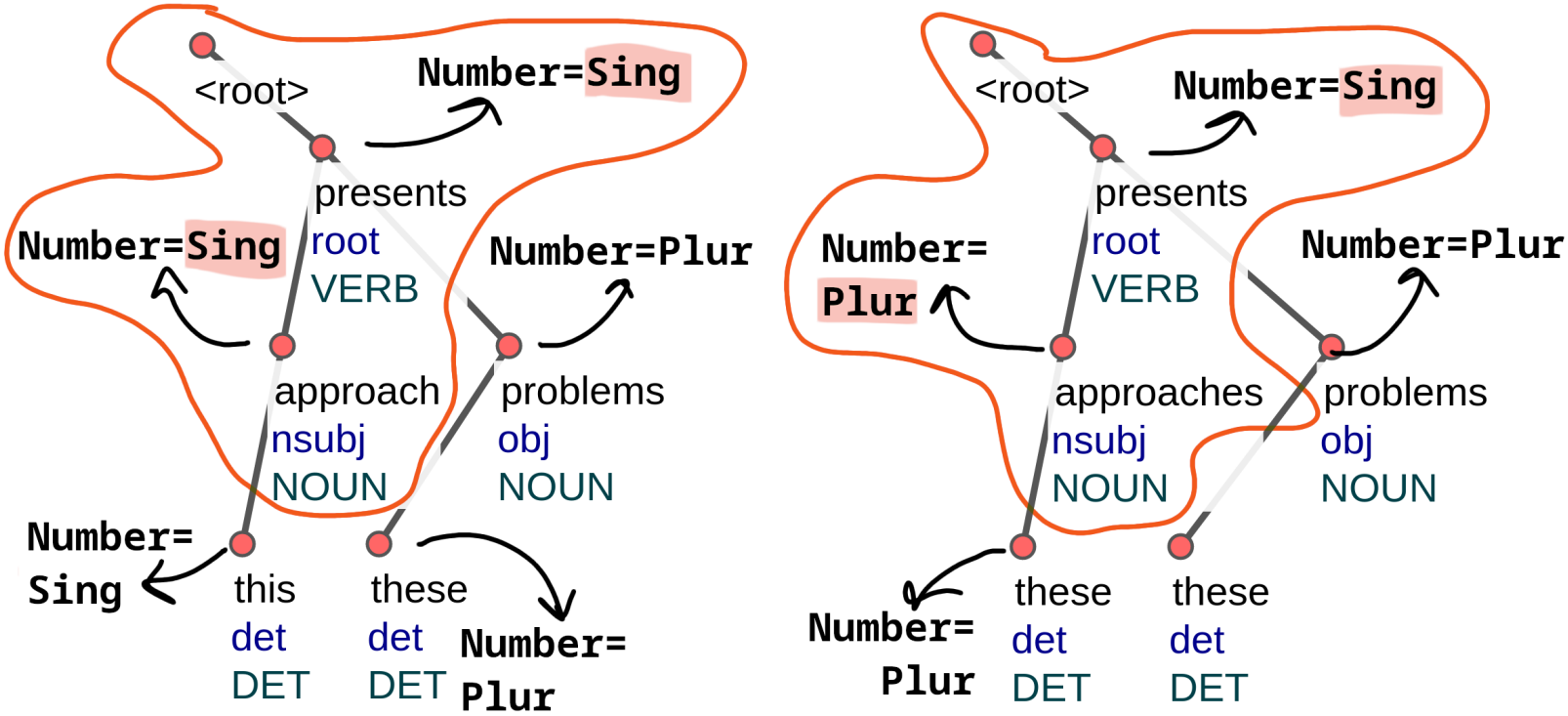
TREE (FEATS "Number=$A") [FEATS "Number=$A →$B]



⟨"this approach presents these problems","*these approaches* presents these problems"⟩

Extracting subsentences

TREE (FEATS "Number=$A") [FEATS "Number=$A →$B]

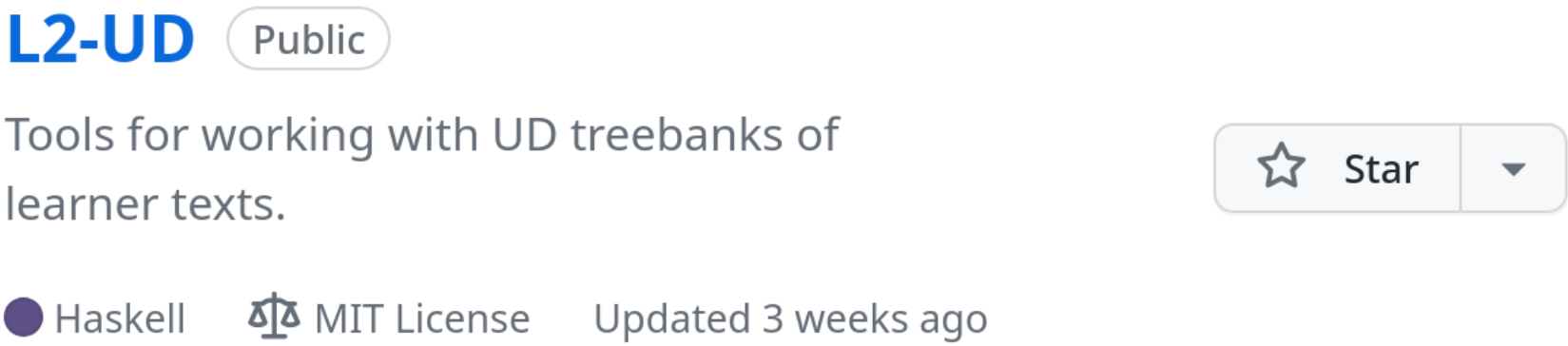


⟨"this **approach** **presents** these problems","*these* ***approaches*** **presents** these problems"⟩

Example output

L1 sentence	L2 sentence
Torino är en stor ort i Italien , och jag är född där och det är **den** bästa **platsen** för mig .	Torino är en stor ort i Italien , och jag är född där och det är **den** bästa **plats** för mig .
Pengar är ingenting utan **de** tre **sakerna** .	Pengar är ingenting utan **de** tre **saker** .
Om en elev inte fokuserar i klassen kan det bli samma sak på **den** individuella **lektionen** också för att hen kan inte förändra sin personlighet .	Om en elev inte fokuserar i klassen kan det bli samma sak på **den** individuella **lektion** också för att hen kan inte förändra sin personlighet .
Den stora **parken** på gården har två barnleksaker , många träd och små vägar för promenader .	**Den** stora **park** på gården har två barnleksaker , många träd och små vägar för promenader .
Jag tycker om **den orten** eftersom jag växte upp där , och jag har studerat där också .	Jag tycker om **den ort** eftersom jag växte där , och jag har studerat där också .

Where is the code?



`github.com/harisont/L2-UD`

Evaluation

Evaluation

- bilingual, both on handcrafted and automatically parsed error annotated data
- sentence-level precision + recall of a single-token, tree and sequence query per corpus
- error patterns typical of the language at hand

Data

- VALICO-UD (Italian, 398 manually validated sentences)
- DaLAJ (Swedish, 2087 automatically parsed sentences)

Results

	VALICO-UD	DaLAJ
single-token	P=43% R=100%	P=77% R=58%
tree	P=100% R=40%	P=75% R=90%
sequence	-	P=89% R=62%

- some error annotation issues were found in the process
- bottlenecks: automatic UD annotation (for DaLAJ) and alignment

To summarize

- new query engine for L1-L2 treebanks
  - corpus- and language-agnostic
  - pattern matching language for UD trees, extended to allow more concise queries
  - parallel queries
  - subsentence extraction
- small-scale bilingual evaluation on manually validated and automatically parsed data
- first tool in a larger toolkit for L2 UD treebanks

Future work

Query engine:

- better variables
- experiments on multilingual parallel treebanks

L2-UD:

- error extraction
- incorrect similar example retrieval

Thank you!

References

- Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. *Universal Dependencies for learner English*. arXiv preprint arXiv:1605.04278, 2016
- 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
- Bruno Guillaume. *Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion*. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Online, 2021

References

- Prasanth Kolachina and Aarnte Ranta. *From abstract syntax to Universal Dependencies*. CSLI Publications volume 13, 2016
- John SY Lee, Herman Leung, and Keying Li. *Towards Universal Dependencies for learner Chinese*. In Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), Gothenburg, Sweden, 2017
- 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, Pisa, Italy, 2017
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. *Universal Dependencies*. Computational Linguistics, 47(2):255–308, 2021

References

- Arianna Masciolini and Aarne Ranta. *Grammar-based concept alignment for domain-specific Machine Translation*. In Proceedings of the Seventh International Workshop on Controlled Natural Language (CNL 2020/21), Amsterdam, Netherlands, 2021
- Aarne Ranta and Prasanth Kolachina. *From Universal Dependencies to abstract syntax*. In Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), Gothenburg, Sweden, 2017
- Petr Pajas and Jan Štěpánek. *System for querying syntactically annotated corpora*. In Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, Suntec, Singapore, 2009

References

- Martin Popel, Zdeněk Žabokrtskỳ, and Martin Vojtek. *UDAPI: Universal API for Universal Dependencies*. In Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), 2017
- Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. DaLAJ-a dataset for linguistic acceptability judgments for Swedish: Format, baseline, sharing. arXiv preprint arXiv:2105.06681, 2021. Description of the DaLAJ v1 corpus (outdated data, up-to-date format).