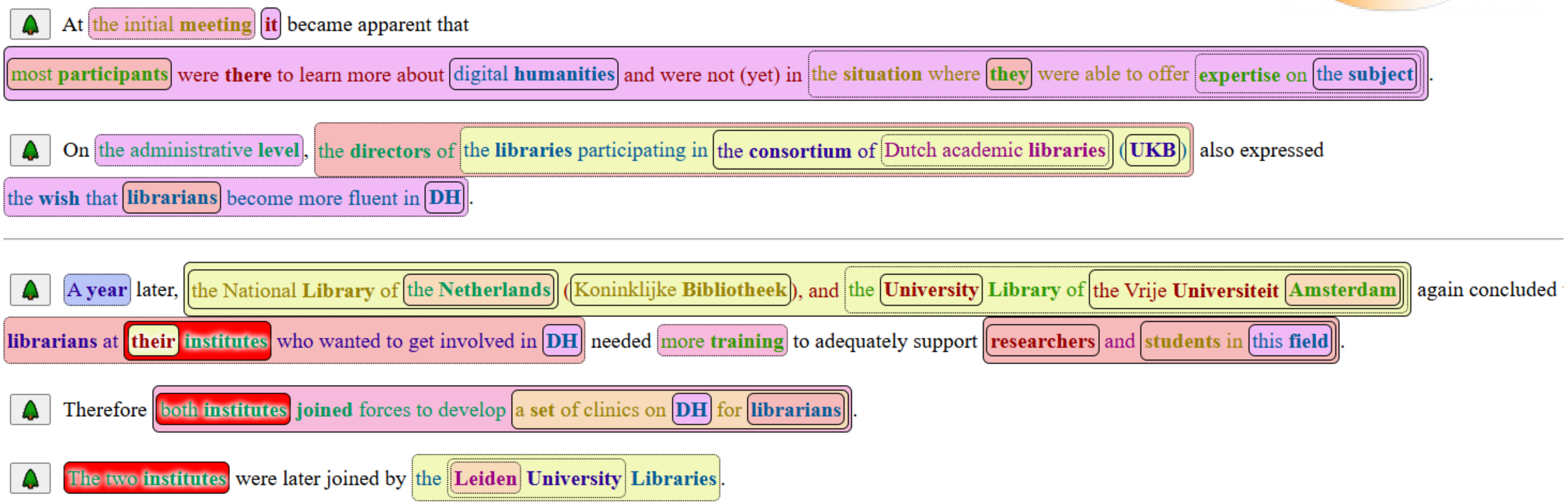
# Multilingual Coreference and Treebanking: Benefits of Interaction

Anja Nedoluzhko

TLT, 5.12.2024

# Prague treebanking tradition and how it evolved



At the initial meeting it became apparent that
most participants were there to learn more about digital humanities and were not (yet) in the situation where they were able to offer expertise on the subject.

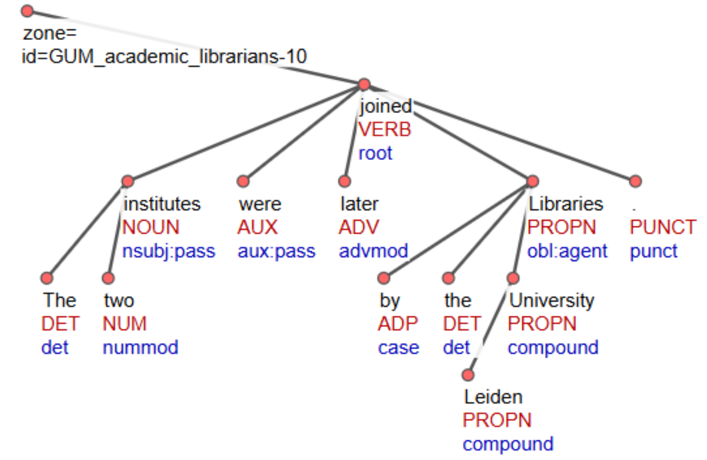
On the administrative level, the directors of the libraries participating in the consortium of Dutch academic libraries (UKB) also expressed
the wish that librarians become more fluent in DH.

A year later, the National Library of the Netherlands (Koninklijke Bibliotheek), and the University Library of the Vrije Universiteit Amsterdam again concluded
librarians at their institutes who wanted to get involved in DH needed more training to adequately support researchers and students in this field.

Therefore both institutes joined forces to develop a set of clinics on DH for librarians.

The two institutes were later joined by the Leiden University Libraries.

The two institutes were later joined by the Leiden University Libraries.



# Outline

Part 1: CorefUD collection

Part 2: CorefUD and UD Treebanking

- (a) Do UD tree structures correspond to mention spans in coreference annotations?
- (b) Are syntactic heads in UD equivalent to the head mentions in coreference annotation?
- (c) How do zeros in coreference relate to empty nodes in syntactic structures?

Part 3: Discussion, conclusions

- benefits of interaction, common challenges, future plans…

# Part 1

# Coreference in Universal Dependencies (CorefUD)

Martin Popel

Michal Novák

Anja Nedoluzhko

Zdeněk Žabokrtský

Daniel Zeman

The Prague CorefUD team

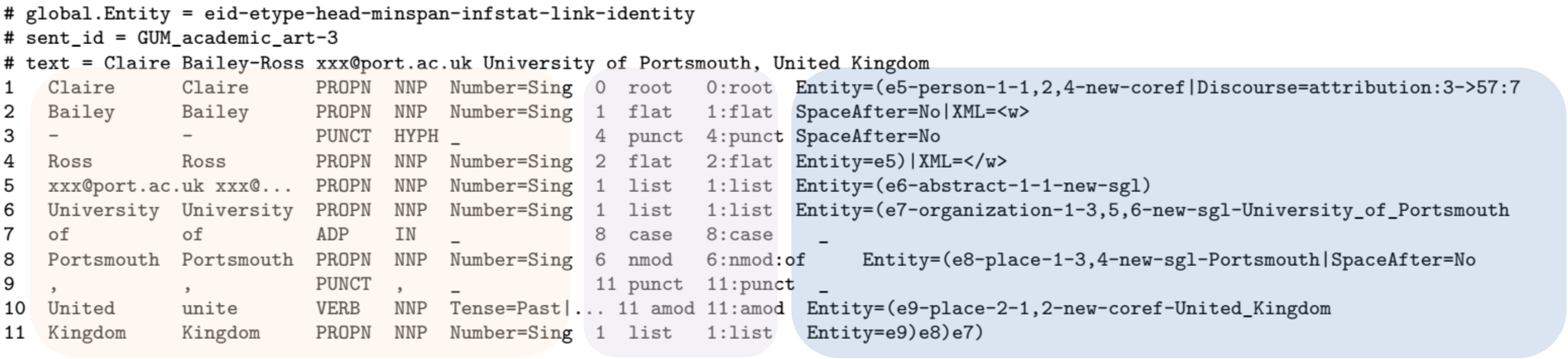
# Examples of coreference

(1) Mary gave **Peter** an apple. Steve gave **him** another one.

ANTECEDENT ANAPHOR

# CorefUD 1.2

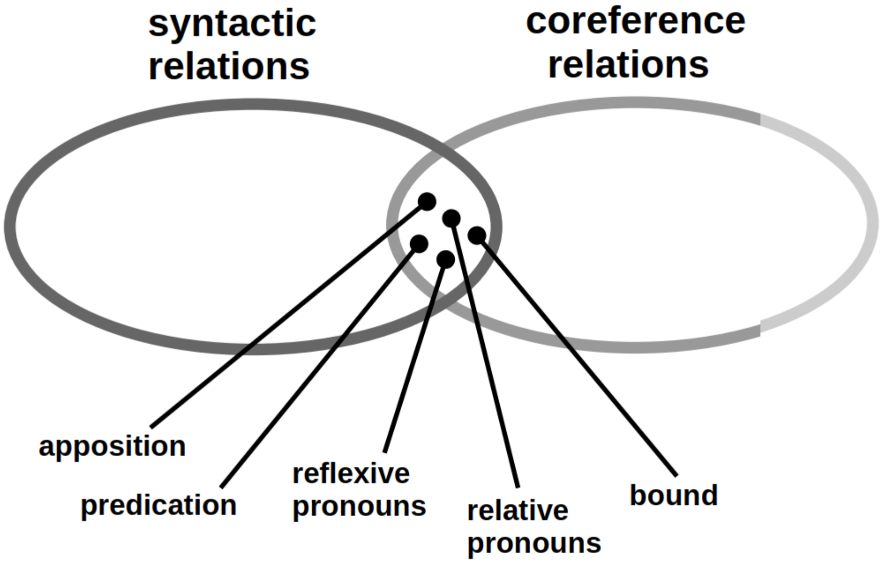
- public edition of CorefUD 1.2 (2024)
- 21 coreference datasets for 15 languages
- harmonized under the same format
- combines annotation of coreference/anaphora (always manual) with annotation of morphology and dependency syntax (manual if available, otherwise automatic)
- the format is valid CoNLL-U; coreference information stored in the MISC column



```
# global.Entity = eid-etype-head-minspan-infstat-link-identity
# sent_id = GUM_academic_art-3
# text = Claire Bailey-Ross xxx@port.ac.uk University of Portsmouth, United Kingdom
1  Claire        Claire      PROPN  NNP  Number=Sing  0  root   0:root   Entity=(e5-person-1-1,2,4-new-coref|Discourse=attribution:3->57:7
2  Bailey        Bailey      PROPN  NNP  Number=Sing  1  flat   1:flat   SpaceAfter=No|XML=<w>
3  -             -           PUNCT  HYPH _            4  punct  4:punct  SpaceAfter=No
4  Ross          Ross        PROPN  NNP  Number=Sing  2  flat   2:flat   Entity=e5)|XML=</w>
5  xxx@port.ac.uk xxx@...    PROPN  NNP  Number=Sing  1  list   1:list   Entity=(e6-abstract-1-1-new-sgl)
6  University    University  PROPN  NNP  Number=Sing  1  list   1:list   Entity=(e7-organization-1-3,5,6-new-sgl-University_of_Portsmouth
7  of            of          ADP    IN   _            8  case   8:case    _
8  Portsmouth    Portsmouth  PROPN  NNP  Number=Sing  6  nmod   6:nmod:of     Entity=(e8-place-1-3,4-new-sgl-Portsmouth|SpaceAfter=No
9  ,             ,           PUNCT  ,    _            11 punct  11:punct  _
10 United        unite       VERB   NNP  Tense=Past|... 11 amod 11:amod  Entity=(e9-place-2-1,2-new-coref-United_Kingdom
11 Kingdom       Kingdom     PROPN  NNP  Number=Sing  1  list   1:list    Entity=e9)e8)e7)
```

# Motivation hypotheses
# coreference + syntax = friends

- mentions often correspond to syntactically meaningful units (noun phrases, subject)
- some coreference relations are expressed primarily by syntactic means (reflexive and relative constructions, apposition, predication with copula)



- reuse of annotation of coordination structures
- syntax is useful for the identification of zero expressions (such as pro-drop) needed for coreference

# CorefUD 1.2

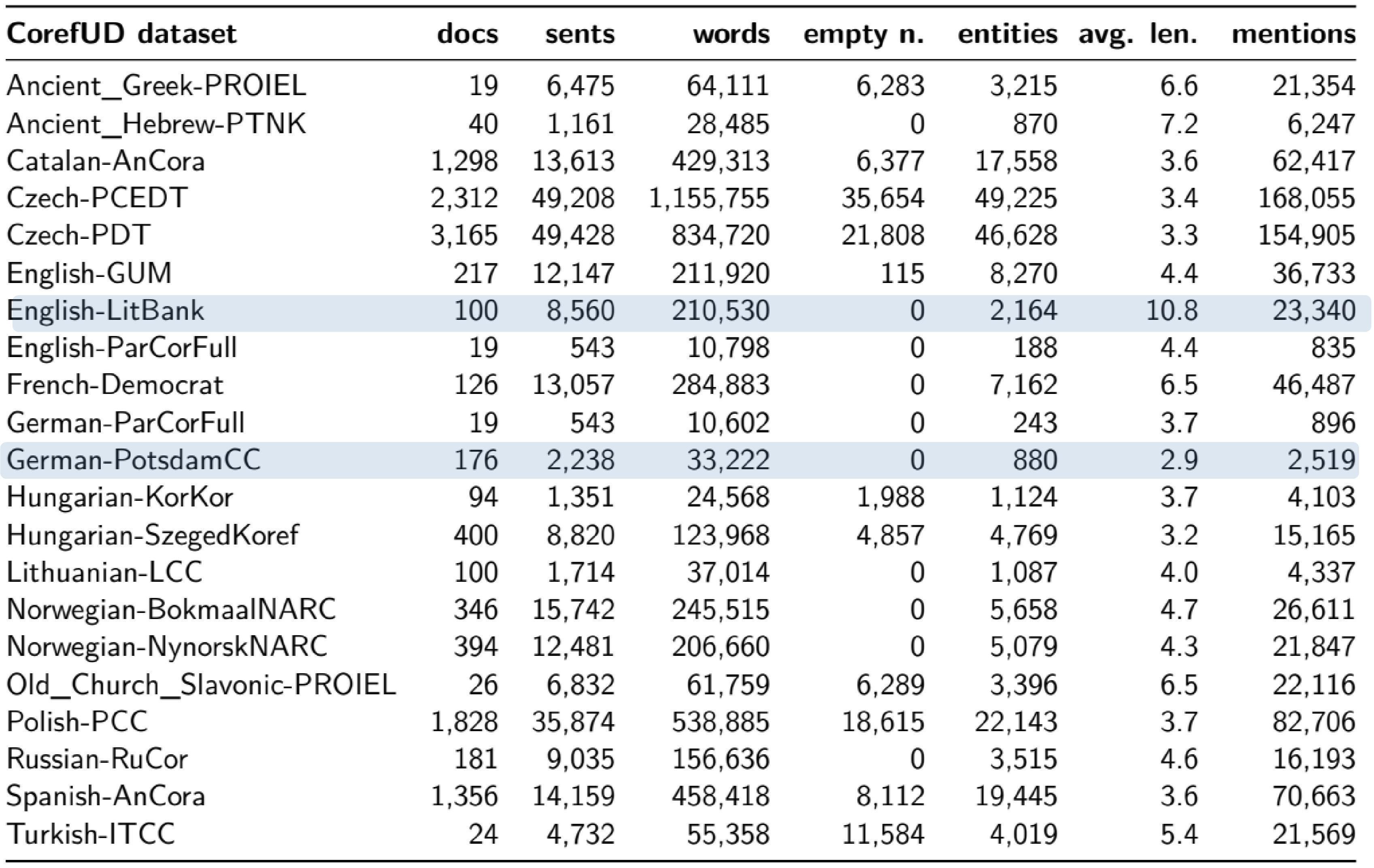
## public datasets:

- Czech-PDT (Hajič et al., 2020)
- Czech-PCEDT (Nedoluzhko et al., 2016)
- English-GUM (Zeldes, 2017)
- English-LitBank (Bamman et al., 2019)
- English-ParCorFull (Lapshinova-Koltunski et al., 2018)
- German-ParCorFull (Lapshinova-Koltunski et al., 2018)
- German-PotsdamCC (Bourgonje and Stede, 2020)
- Norwegian-BokmaalNARC (Mæhlum et al., 2022)
- Norwegian-NynorskNARC (Mæhlum et al., 2022)
- Spanish-AnCora (Recasens and Martí, 2010)
- Catalan-AnCora (Recasens and Martí, 2010)
- French-Democrat (Landragin, 2021)
- Polish-PCC (Ogrodniczuk et al., 2013)
- Lithuanian-LCC (Žitkus and Butkienė, 2018)
- Russian-RuCor (Toldova et al., 2014)
- Hungarian-SzegedKoref (Vincze et al., 2018)
- Hungarian-KorKor (Vadász, 2022)
- Turkish-ITCC (Pamay and Eryiğit, 2018)
- Ancient Greek-PROIEL (Haug and Jøhndal, 2008)
- Old Church Slavonic-PROIEL (Haug and Jøhndal, 2008)
- Ancient Hebrew-PTNK (Swanson et al., 2024)

## non-public datasets:

- English-OntoNotes (Weischedel et al., 2011)
- English-ARRAU (Uryupina et al., 2020)
- Dutch-COREA (Hendrickx et al., 2008)
- English-PCEDT (Nedoluzhko et al., 2016)

# CorefUD 1.2
statistics in the public datasets



| CorefUD dataset | docs | sents | words | empty n. | entities | avg. len. | mentions |
|---|---|---|---|---|---|---|---|
| Ancient_Greek-PROIEL | 19 | 6,475 | 64,111 | 6,283 | 3,215 | 6.6 | 21,354 |
| Ancient_Hebrew-PTNK | 40 | 1,161 | 28,485 | 0 | 870 | 7.2 | 6,247 |
| Catalan-AnCora | 1,298 | 13,613 | 429,313 | 6,377 | 17,558 | 3.6 | 62,417 |
| Czech-PCEDT | 2,312 | 49,208 | 1,155,755 | 35,654 | 49,225 | 3.4 | 168,055 |
| Czech-PDT | 3,165 | 49,428 | 834,720 | 21,808 | 46,628 | 3.3 | 154,905 |
| English-GUM | 217 | 12,147 | 211,920 | 115 | 8,270 | 4.4 | 36,733 |
| English-LitBank | 100 | 8,560 | 210,530 | 0 | 2,164 | 10.8 | 23,340 |
| English-ParCorFull | 19 | 543 | 10,798 | 0 | 188 | 4.4 | 835 |
| French-Democrat | 126 | 13,057 | 284,883 | 0 | 7,162 | 6.5 | 46,487 |
| German-ParCorFull | 19 | 543 | 10,602 | 0 | 243 | 3.7 | 896 |
| German-PotsdamCC | 176 | 2,238 | 33,222 | 0 | 880 | 2.9 | 2,519 |
| Hungarian-KorKor | 94 | 1,351 | 24,568 | 1,988 | 1,124 | 3.7 | 4,103 |
| Hungarian-SzegedKoref | 400 | 8,820 | 123,968 | 4,857 | 4,769 | 3.2 | 15,165 |
| Lithuanian-LCC | 100 | 1,714 | 37,014 | 0 | 1,087 | 4.0 | 4,337 |
| Norwegian-BokmaalNARC | 346 | 15,742 | 245,515 | 0 | 5,658 | 4.7 | 26,611 |
| Norwegian-NynorskNARC | 394 | 12,481 | 206,660 | 0 | 5,079 | 4.3 | 21,847 |
| Old_Church_Slavonic-PROIEL | 26 | 6,832 | 61,759 | 6,289 | 3,396 | 6.5 | 22,116 |
| Polish-PCC | 1,828 | 35,874 | 538,885 | 18,615 | 22,143 | 3.7 | 82,706 |
| Russian-RuCor | 181 | 9,035 | 156,636 | 0 | 3,515 | 4.6 | 16,193 |
| Spanish-AnCora | 1,356 | 14,159 | 458,418 | 8,112 | 19,445 | 3.6 | 70,663 |
| Turkish-ITCC | 24 | 4,732 | 55,358 | 11,584 | 4,019 | 5.4 | 21,569 |

Novák et al. 2024

# Examples of coreference

(1) Mary gave **Peter** an apple. Steve gave **him** another one.

ANTECEDENT ANAPHOR

(2)Mary gave Peter **an apple**. Steve gave him **another one**. Peter took **them** and left.

(*split antecedent*)

(3) I didn’t like **this apple**. I bit **it** off several times and threw **it** out of the window.

(*near-identity*)

(4) I finished **my apple** and threw **the stub** out the window. (*bridging*)

(5) **I ate Peter’s apple**. He will never forgive me for **that**. (*discourse deixis*)

(6) **My apple**, **the red one**, is really good. (*apposition*)

(7) **This red apple** is **a symbol of happiness**. (*predication*)

# Diversity in existing resources: relations (in CorefUD 1.0)

| CorefUD dataset | Coref. grouping | | Relations among mentions | | | | | |
|---|---|---|---|---|---|---|---|---|
| | cluster-based | link-based | singletons | appos. | pred. | split antec. | disc. deixis | bridg. |
| Catalan-AnCora | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ | × |
| Czech-PCEDT | × | ✓ | (✓) | (✓) | (✓) | ✓ | ✓ | × |
| Czech-PDT | × | ✓ | (✓) | (✓) | (✓) | ✓ | ✓ | ✓ |
| English-GUM | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| English-ParCorFull | ✓ | × | × | ✓ | (✓) | ✓ | ✓ | × |
| French-Democrat | ✓ | × | ✓ | × | × | × | × | × |
| German-ParCorFull | ✓ | × | × | ✓ | (✓) | ✓ | ✓ | × |
| German-PotsdamCC | × | ✓ | ✓ | ✓ | ✓ ? | × | ✓ | × |
| Hungarian-SzegedKoref | ✓ | × | × | ✓ | ? | × | ✓ | ✓ |
| Lithuanian-LCC | × | ✓ | × | × | × | ✓ | × | × |
| Polish-PCC | ✓ | × | ✓ | ✓ | ✓ | × | ✓ | ✓ |
| Russian-RuCor | ✓ | × | × | ✓ | ✓ | × | × | × |
| Spanish-AnCora | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ | × |
| Dutch-COREA | × | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ |
| English-ARRAU | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| English-OntoNotes | ✓ | × | × | ✓ | × | × | ✓ | × |
| English-PCEDT | × | ✓ | (✓) | (✓) | (✓) | ✓ | ✓ | × |

# Shared Tasks on multilingual coreference resolution based on the CorefUD dataset

| Shared task | Languages |
|---|---|
| CRAC 2022 | 10 |
| CRAC 2023 | 12 |
| CRAC 2024 | 15 |
| CRAC 2025 *planned* | more than 15 |

https://ufal.mff.cuni.cz/corefud/crac24

***LLMs***

- systems
  - no syntax use (except for heads),
  - no POS tags, ignore all columns except for forms
- baselines
  - only for zeros
- evaluation
  - for head-matching (BUT given by CorefUD)
    - gold mention heads determined from the dependency tree using the Udapi block corefud.MoveHead
  - for zeros alignment (9 DEPS column)

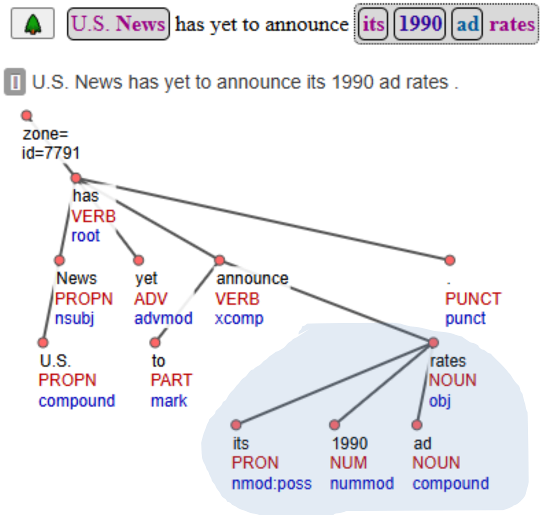
# Part 2

# CorefUD and Treebanks

## UD tree structures and mention spans

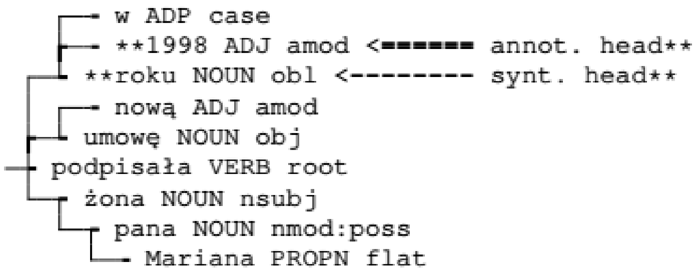
## Syntactic heads in UD and head mentions

## Empty nodes in UD and zero anaphora

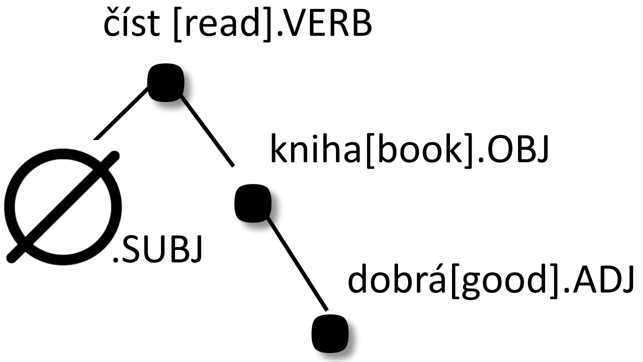


U.S. News has yet to announce its 1990 ad rates .

U.S. News has yet to announce its 1990 ad rates .



```
    ┌── w ADP case
    ├── **1998 ADJ amod <======= annot. head**
  ┌─┴ **roku NOUN obl <-------- synt. head**
  │ ┌── nową ADJ amod
  ├─┴ umowę NOUN obj
──┼ podpisała VERB root
  └─┬ żona NOUN nsubj
    └─┬ pana NOUN nmod:poss
      └── Mariana PROPN flat
```



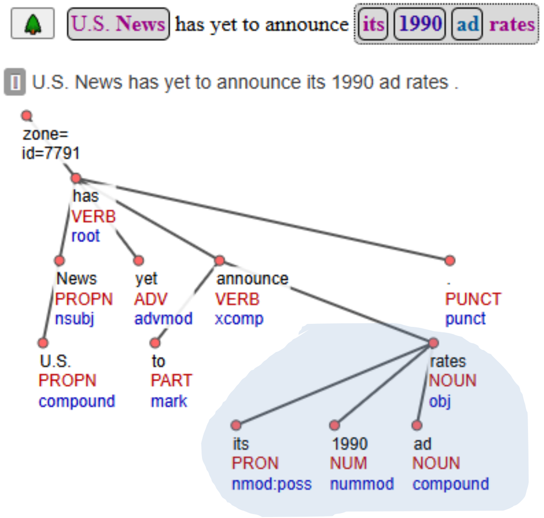
# Part 2

# CorefUD and Treebanks

### UD tree structures and mention spans

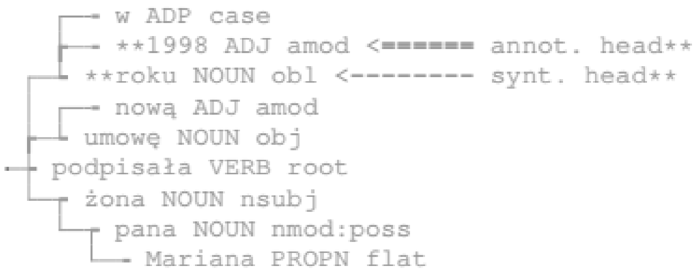
### Syntactic heads in UD and head mentions

### Empty nodes in UD and zero anaphora

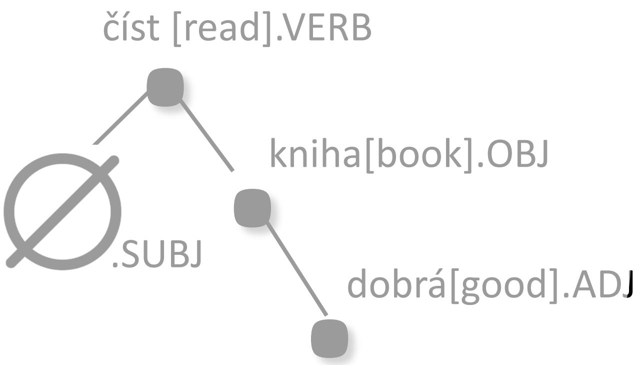


U.S. News has yet to announce its 1990 ad rates .

U.S. News has yet to announce its 1990 ad rates .



```
      ┌── w ADP case
      ├── **1998 ADJ amod <====== annot. head**
    ┌─┴ **roku NOUN obl <-------- synt. head**
    │ ┌── nową ADJ amod
    ├─┴ umowę NOUN obj
  ──┼ podpisała VERB root
    └─┬ żona NOUN nsubj
      └─┬ pana NOUN nmod:poss
        └── Mariana PROPN flat
```

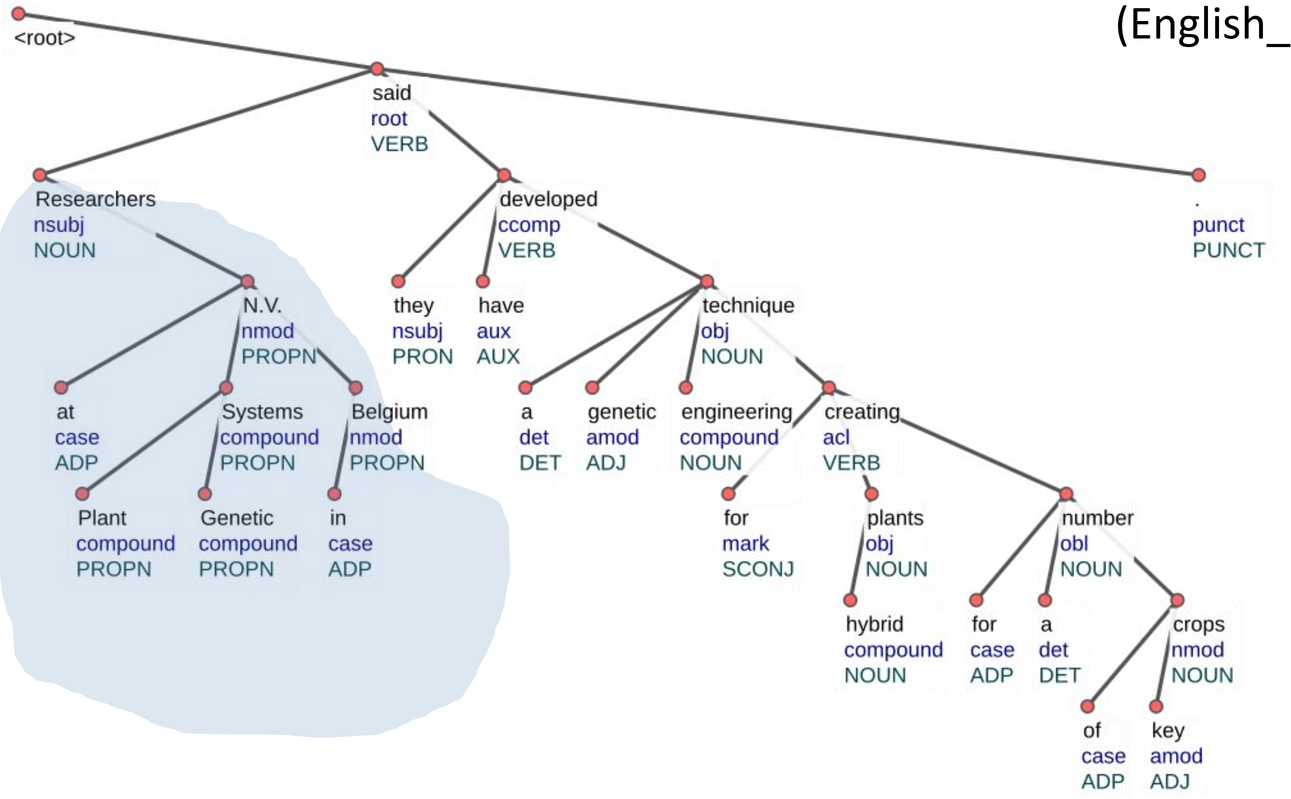


# Do UD tree structures correspond to mention spans in coreference annotations?

**mention** **coreference**

*Researchers at Plant Generic Systems N.V. in Belgium said they have developed a genetic engineering technique for creating hybrid plants for a number of key crops*.

(English_ARRAU)



*Hypothesis:*

mentions are not just unconstrained subsequences of tokens, but mostly correspond to syntactically meaningful units

# CorefUD 1.2

## public datasets:

- Czech-PDT (Hajič et al., 2020)
- Czech-PCEDT (Nedoluzhko et al., 2016)
- **English-GUM** (Zeldes, 2017)
- English-LitBank (Bamman et al., 2019)
- English-ParCorFull (Lapshinova-Koltunski et al., 2018)
- German-ParCorFull (Lapshinova-Koltunski et al., 2018)
- **German-PotsdamCC** (Bourgonje and Stede, 2020)
- Norwegian-BokmaalNARC (Mæhlum et al., 2022)
- Norwegian-NynorskNARC (Mæhlum et al., 2022)
- Spanish-AnCora (Recasens and Martí, 2010)
- Catalan-AnCora (Recasens and Martí, 2010)
- **French-Democrat** (Landragin, 2021)
- **Polish-PCC** (Ogrodniczuk et al., 2013)
- **Lithuanian-LCC** (Žitkus and Butkienė, 2018)
- **Russian-RuCor** (Toldova et al., 2014)
- **Hungarian-SzegedKoref** (Vincze et al., 2018)
- Hungarian-KorKor (Vadász, 2022)
- Turkish-ITCC (Pamay and Eryiğit, 2018)

l,

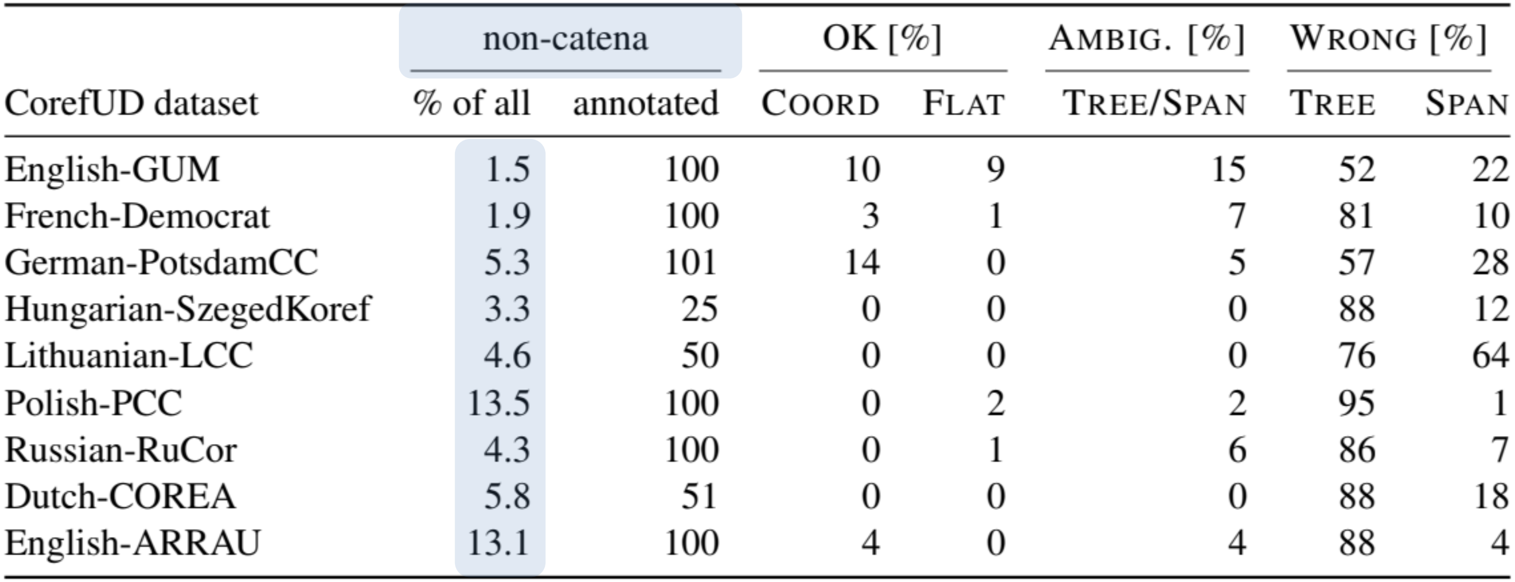
- mentions are not linked with dependency trees
- corpora included into CorefUD_0.2

Ancient Hebrew-PTNK (Swanson et al., 2024)

## non-public datasets:

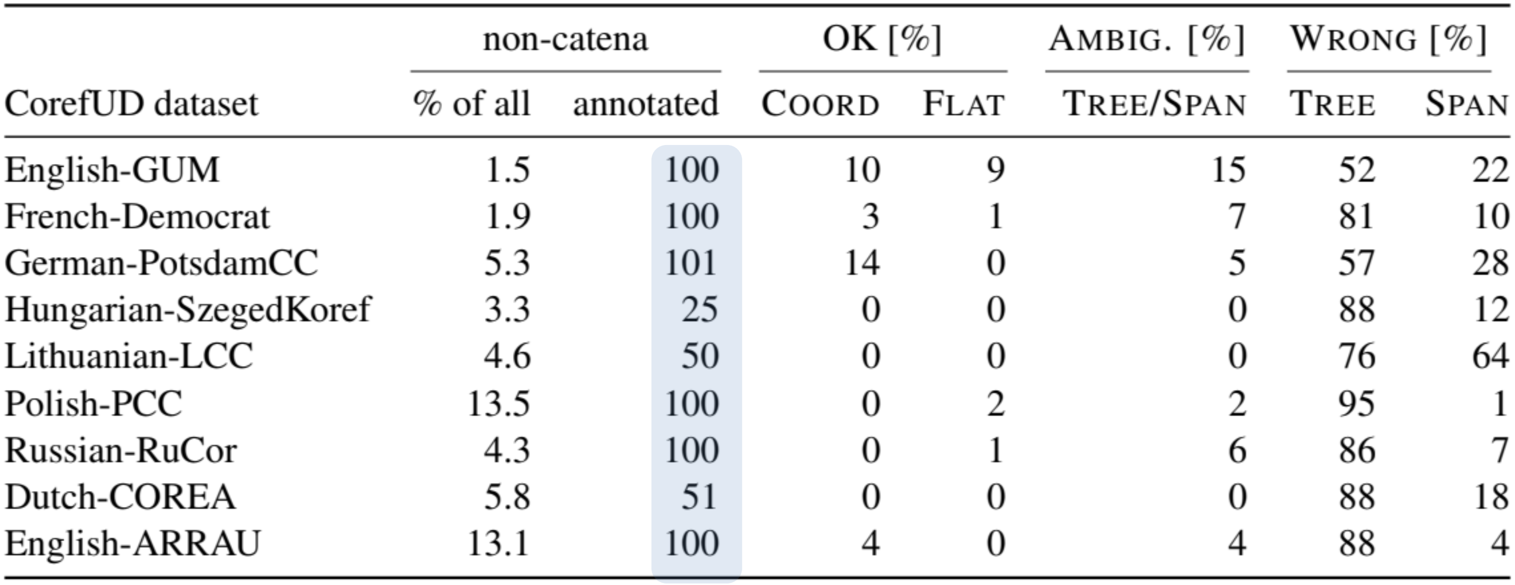
- English-OntoNotes (Weischedel et al., 2011)
- **English-ARRAU** (Uryupina et al., 2020)
- **Dutch-COREA** (Hendrickx et al., 2008)
- English-PCEDT (Nedoluzhko et al., 2016)

# Do UD tree structures correspond to mention spans in coreference annotations?



| | non-catena | | OK [%] | | Ambig. [%] | Wrong [%] | |
|---|---|---|---|---|---|---|---|
| CorefUD dataset | % of all | annotated | Coord | Flat | Tree/Span | Tree | Span |
| English-GUM | 1.5 | 100 | 10 | 9 | 15 | 52 | 22 |
| French-Democrat | 1.9 | 100 | 3 | 1 | 7 | 81 | 10 |
| German-PotsdamCC | 5.3 | 101 | 14 | 0 | 5 | 57 | 28 |
| Hungarian-SzegedKoref | 3.3 | 25 | 0 | 0 | 0 | 88 | 12 |
| Lithuanian-LCC | 4.6 | 50 | 0 | 0 | 0 | 76 | 64 |
| Polish-PCC | 13.5 | 100 | 0 | 2 | 2 | 95 | 1 |
| Russian-RuCor | 4.3 | 100 | 0 | 1 | 6 | 86 | 7 |
| Dutch-COREA | 5.8 | 51 | 0 | 0 | 0 | 88 | 18 |
| English-ARRAU | 13.1 | 100 | 4 | 0 | 4 | 88 | 4 |

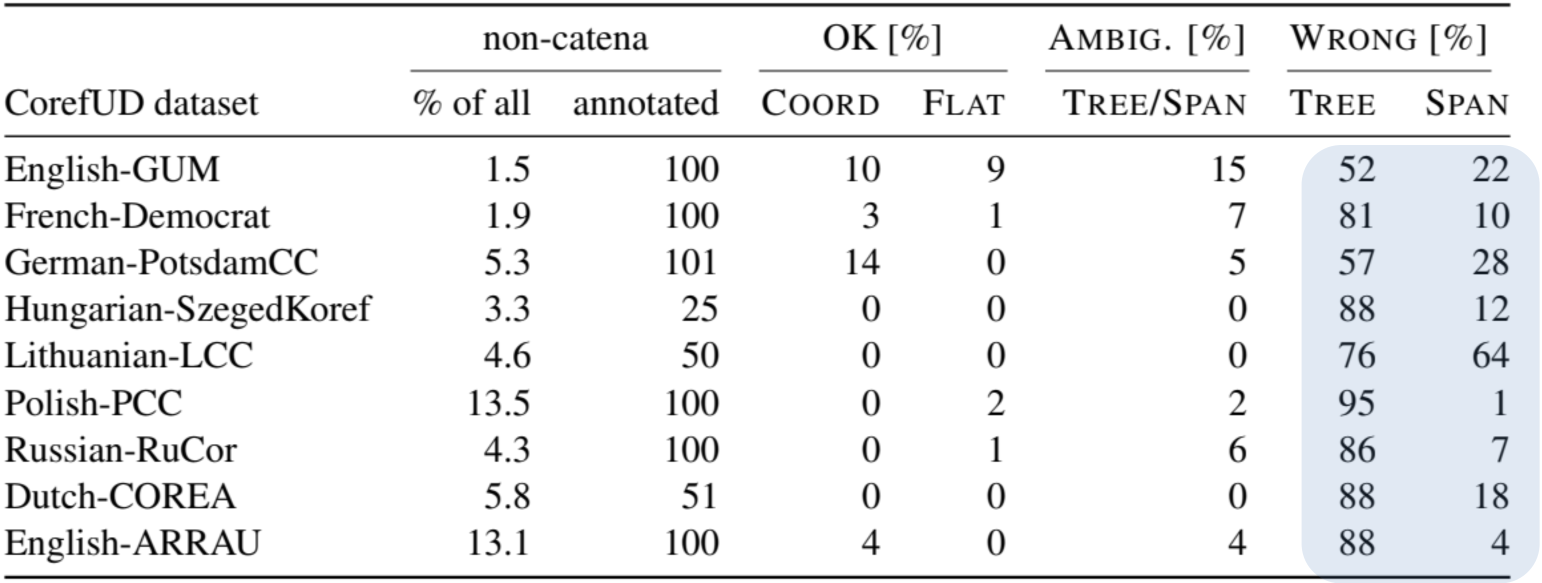
# Do UD tree structures correspond to mention spans in coreference annotations?



| | non-catena | | OK [%] | | AMBIG. [%] | WRONG [%] | |
|---|---|---|---|---|---|---|---|
| CorefUD dataset | % of all | annotated | COORD | FLAT | TREE/SPAN | TREE | SPAN |
| English-GUM | 1.5 | 100 | 10 | 9 | 15 | 52 | 22 |
| French-Democrat | 1.9 | 100 | 3 | 1 | 7 | 81 | 10 |
| German-PotsdamCC | 5.3 | 101 | 14 | 0 | 5 | 57 | 28 |
| Hungarian-SzegedKoref | 3.3 | 25 | 0 | 0 | 0 | 88 | 12 |
| Lithuanian-LCC | 4.6 | 50 | 0 | 0 | 0 | 76 | 64 |
| Polish-PCC | 13.5 | 100 | 0 | 2 | 2 | 95 | 1 |
| Russian-RuCor | 4.3 | 100 | 0 | 1 | 6 | 86 | 7 |
| Dutch-COREA | 5.8 | 51 | 0 | 0 | 0 | 88 | 18 |
| English-ARRAU | 13.1 | 100 | 4 | 0 | 4 | 88 | 4 |

We analyzed a sample of mentions that are not catenae (Popel et al. 2021)

# Do UD tree structures correspond to mention spans in coreference annotations?

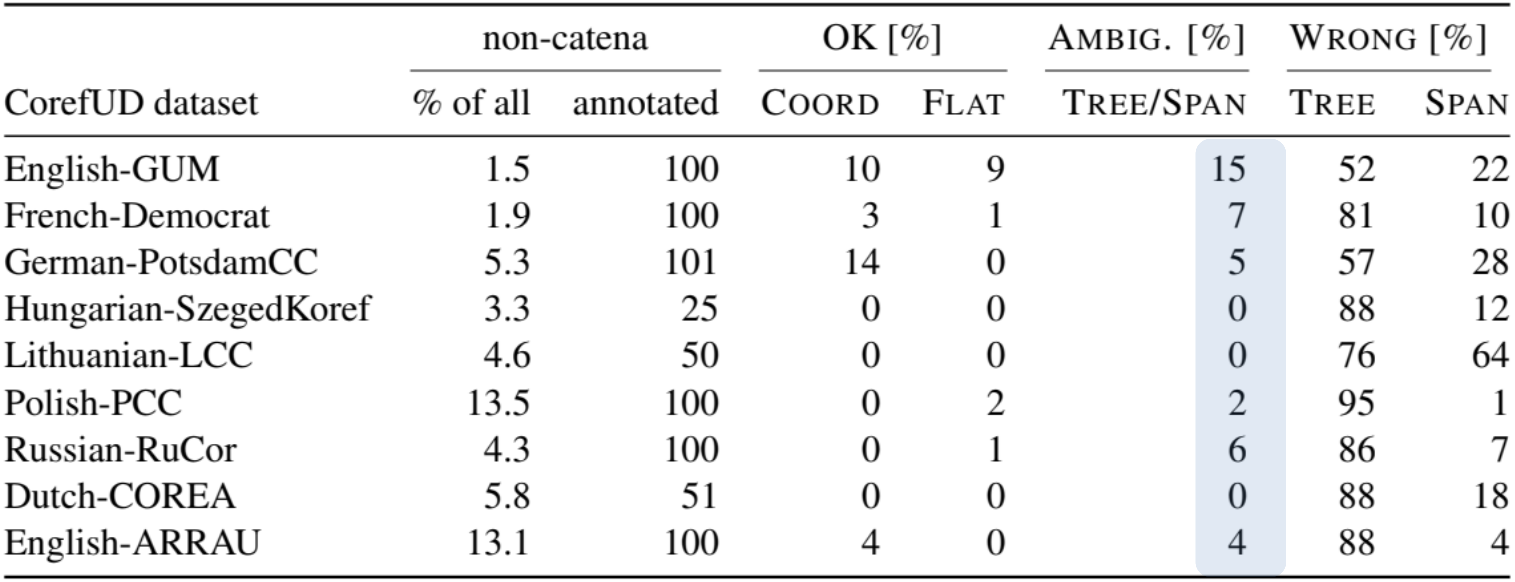


| | non-catena | | OK [%] | | AMBIG. [%] | WRONG [%] | |
|---|---|---|---|---|---|---|---|
| CorefUD dataset | % of all | annotated | COORD | FLAT | TREE/SPAN | TREE | SPAN |
| English-GUM | 1.5 | 100 | 10 | 9 | 15 | 52 | 22 |
| French-Democrat | 1.9 | 100 | 3 | 1 | 7 | 81 | 10 |
| German-PotsdamCC | 5.3 | 101 | 14 | 0 | 5 | 57 | 28 |
| Hungarian-SzegedKoref | 3.3 | 25 | 0 | 0 | 0 | 88 | 12 |
| Lithuanian-LCC | 4.6 | 50 | 0 | 0 | 0 | 76 | 64 |
| Polish-PCC | 13.5 | 100 | 0 | 2 | 2 | 95 | 1 |
| Russian-RuCor | 4.3 | 100 | 0 | 1 | 6 | 86 | 7 |
| Dutch-COREA | 5.8 | 51 | 0 | 0 | 0 | 88 | 18 |
| English-ARRAU | 13.1 | 100 | 4 | 0 | 4 | 88 | 4 |

We analyzed a sample of mentions that are not catenae (Popel et al. 2021)

- WRONGTREE – parsing error
- WRONGSPAN – errors in coreference annotation

# Do UD tree structures correspond to mention spans in coreference annotations?

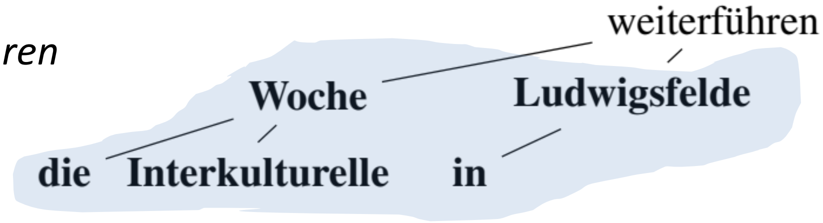


| | non-catena | | OK [%] | | AMBIG. [%] | WRONG [%] | |
|---|---|---|---|---|---|---|---|
| CorefUD dataset | % of all | annotated | COORD | FLAT | TREE/SPAN | TREE | SPAN |
| English-GUM | 1.5 | 100 | 10 | 9 | 15 | 52 | 22 |
| French-Democrat | 1.9 | 100 | 3 | 1 | 7 | 81 | 10 |
| German-PotsdamCC | 5.3 | 101 | 14 | 0 | 5 | 57 | 28 |
| Hungarian-SzegedKoref | 3.3 | 25 | 0 | 0 | 0 | 88 | 12 |
| Lithuanian-LCC | 4.6 | 50 | 0 | 0 | 0 | 76 | 64 |
| Polish-PCC | 13.5 | 100 | 0 | 2 | 2 | 95 | 1 |
| Russian-RuCor | 4.3 | 100 | 0 | 1 | 6 | 86 | 7 |
| Dutch-COREA | 5.8 | 51 | 0 | 0 | 0 | 88 | 18 |
| English-ARRAU | 13.1 | 100 | 4 | 0 | 4 | 88 | 4 |

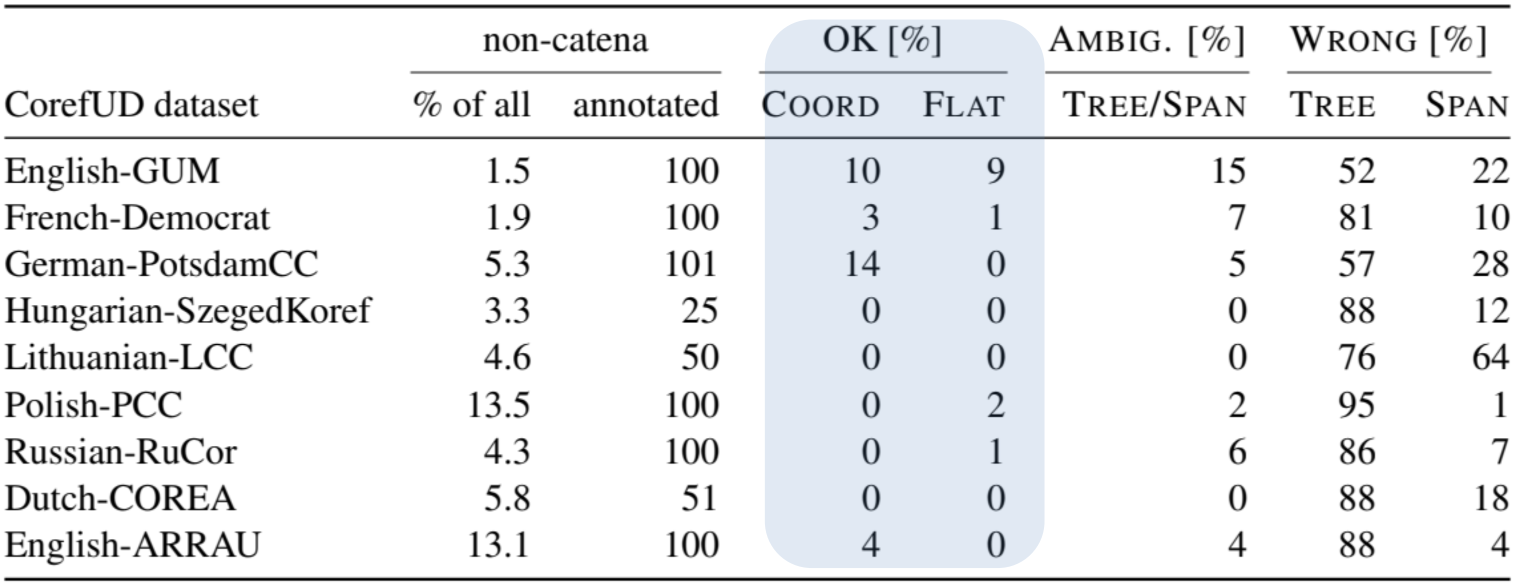
We analyzed a sample of mentions that are not catenae (Popel et al. 2021)

- WRONGTREE – parsing error
- WRONGSPAN – errors in coreference annotation
- AMBIGUOUS – both could be OK

*[die Interkulturelle Woche in Ludwigsfelde] weiterführen*
‘to continue the Intercultural Week in Ludwigsfelde’



# Do UD tree structures correspond to mention spans in coreference annotations?

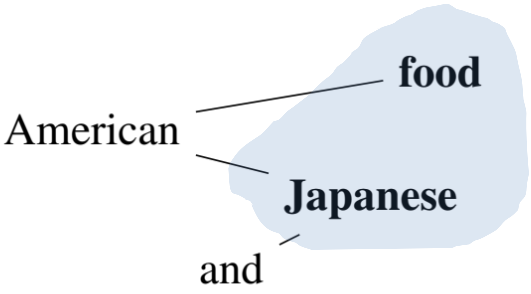


| | non-catena | | OK [%] | | AMBIG. [%] | WRONG [%] | |
|---|---|---|---|---|---|---|---|
| CorefUD dataset | % of all | annotated | COORD | FLAT | TREE/SPAN | TREE | SPAN |
| English-GUM | 1.5 | 100 | 10 | 9 | 15 | 52 | 22 |
| French-Democrat | 1.9 | 100 | 3 | 1 | 7 | 81 | 10 |
| German-PotsdamCC | 5.3 | 101 | 14 | 0 | 5 | 57 | 28 |
| Hungarian-SzegedKoref | 3.3 | 25 | 0 | 0 | 0 | 88 | 12 |
| Lithuanian-LCC | 4.6 | 50 | 0 | 0 | 0 | 76 | 64 |
| Polish-PCC | 13.5 | 100 | 0 | 2 | 2 | 95 | 1 |
| Russian-RuCor | 4.3 | 100 | 0 | 1 | 6 | 86 | 7 |
| Dutch-COREA | 5.8 | 51 | 0 | 0 | 0 | 88 | 18 |
| English-ARRAU | 13.1 | 100 | 4 | 0 | 4 | 88 | 4 |

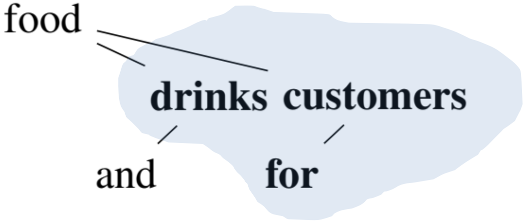
We analyzed a sample of mentions that are not catenae (Popel et al. 2021)

- WRONGTREE – parsing error
- WRONGSPAN – errors in coreference annotation
- AMBIGUOUS – both could be OK
- OK – caused by the UD syntactic guidelines
  - COORDination
  - FLAT structures
    - names, foreign phrases, dates, complex numerals…

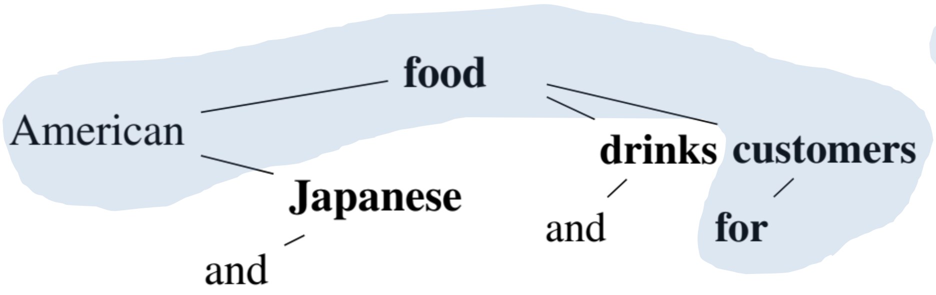
# Coordination in UD tree structures vs. coordination in mention spans



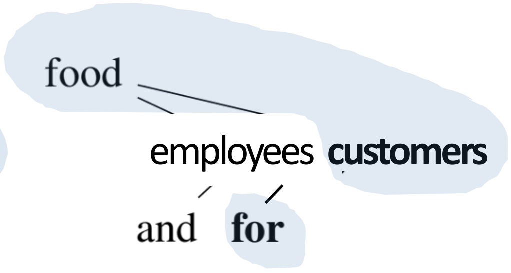
*American and [Japanese food]*



*food and [drinks for customers]*

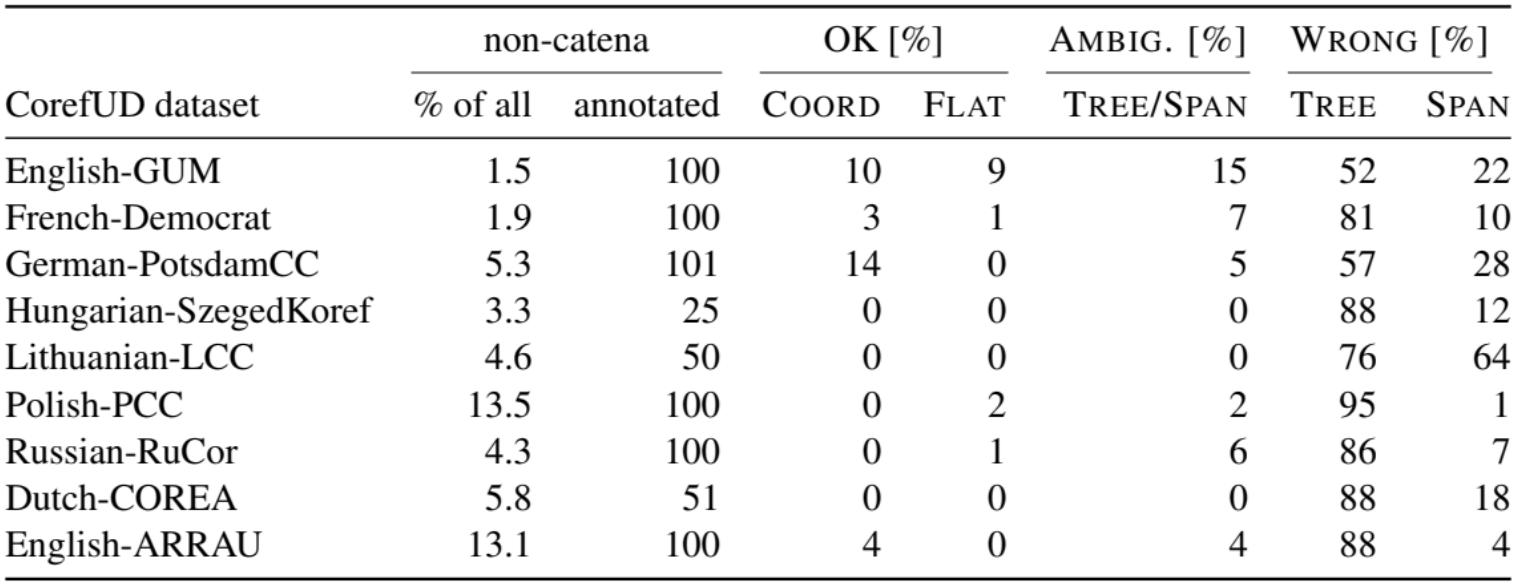


***American** and Japanese **food** and drinks **for customers***



***food for** employees or **customers***

# Do UD tree structures correspond to mention spans in coreference annotations?



| | non-catena | | OK [%] | | AMBIG. [%] | WRONG [%] | |
|---|---|---|---|---|---|---|---|
| CorefUD dataset | % of all | annotated | COORD | FLAT | TREE/SPAN | TREE | SPAN |
| English-GUM | 1.5 | 100 | 10 | 9 | 15 | 52 | 22 |
| French-Democrat | 1.9 | 100 | 3 | 1 | 7 | 81 | 10 |
| German-PotsdamCC | 5.3 | 101 | 14 | 0 | 5 | 57 | 28 |
| Hungarian-SzegedKoref | 3.3 | 25 | 0 | 0 | 0 | 88 | 12 |
| Lithuanian-LCC | 4.6 | 50 | 0 | 0 | 0 | 76 | 64 |
| Polish-PCC | 13.5 | 100 | 0 | 2 | 2 | 95 | 1 |
| Russian-RuCor | 4.3 | 100 | 0 | 1 | 6 | 86 | 7 |
| Dutch-COREA | 5.8 | 51 | 0 | 0 | 0 | 88 | 18 |
| English-ARRAU | 13.1 | 100 | 4 | 0 | 4 | 88 | 4 |

*Hypothesis:*

mentions are not just unconstrained subsequences of tokens, but mostly correspond to syntactically meaningful units

*Conclusion*:

mention spans usually correspond to subtrees of dependency trees quite straightforwardly; and if they do not, it can be explained by errors in dependency trees or by mention span errors in most cases

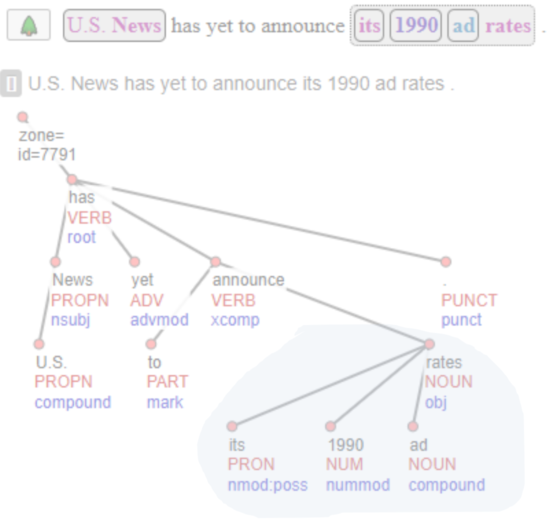
# Part 2

# CorefUD and Treebanks

## UD tree structures and mention spans

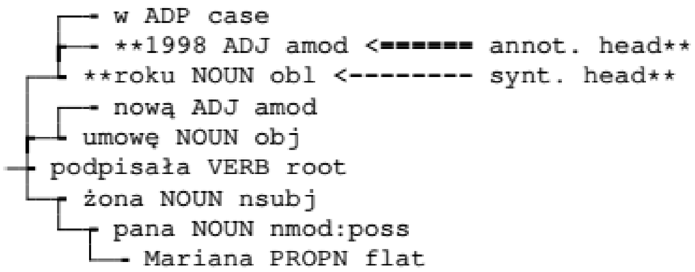
## Syntactic heads in UD and head mentions

## Empty nodes in UD and zero anaphora

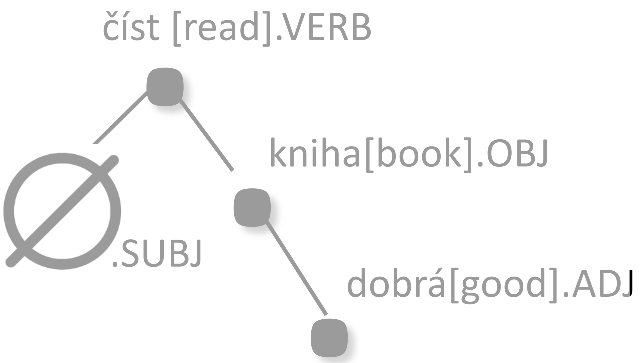


U.S. News has yet to announce its 1990 ad rates .

U.S. News has yet to announce its 1990 ad rates .



```
w ADP case
**1998 ADJ amod <====== annot. head**
**roku NOUN obl <-------- synt. head**
nową ADJ amod
umowę NOUN obj
podpisała VERB root
żona NOUN nsubj
pana NOUN nmod:poss
Mariana PROPN flat
```



# Are syntactic heads in UD equivalent to the head mentions in coreference annotation?

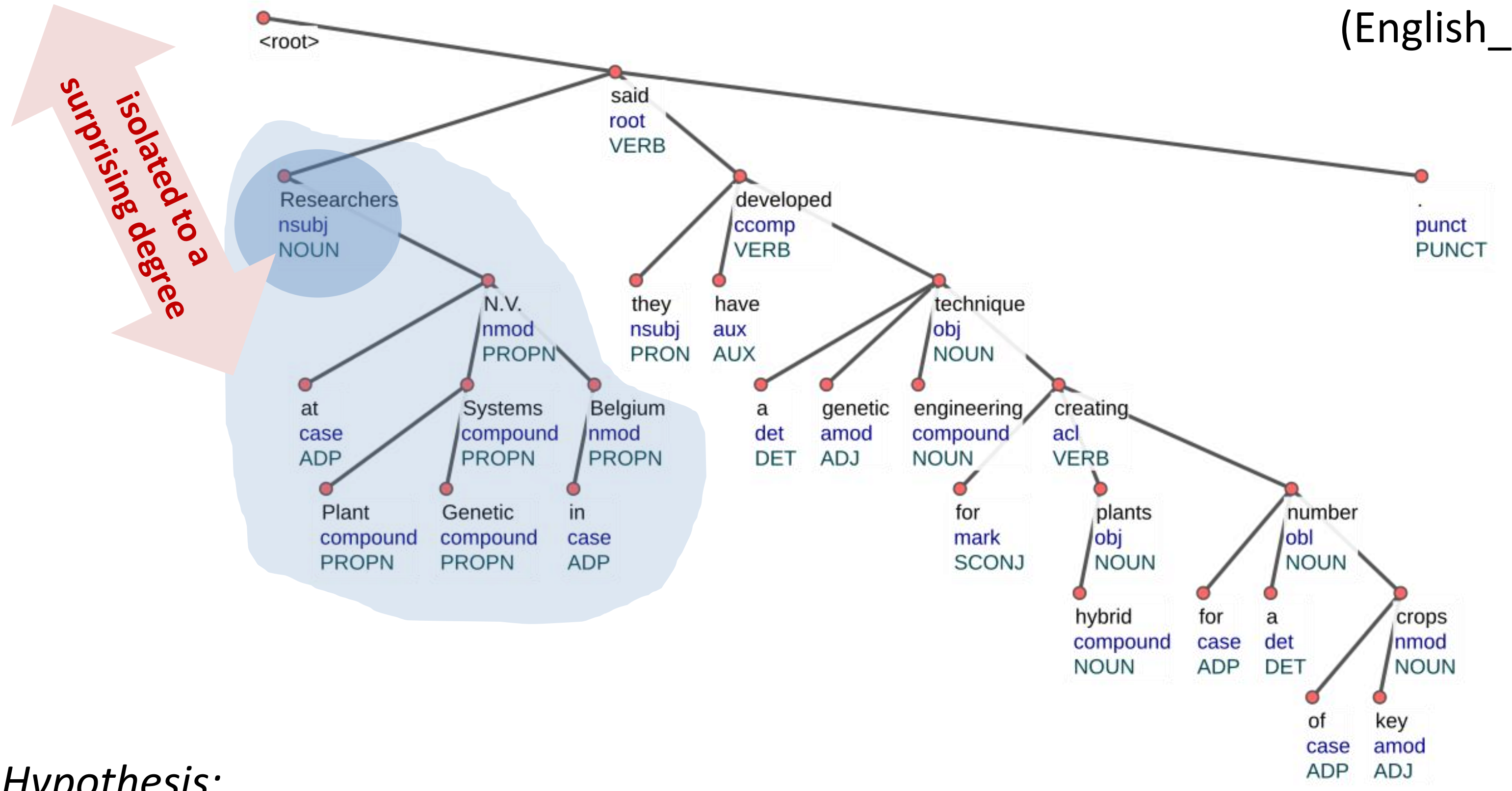
**mention head (min span)**

**mention**

**coreference**

*Researchers at Plant Generic Systems N.V. in Belgium said they have developed a genetic engineering technique for creating hybrid plants for a number of key crops*.

(English_ARRAU)



*Hypothesis:*

mention heads in annotations and syntactic heads in syntactic trees are closely related and compatible

# CorefUD 1.2

## public datasets:

- Czech-PDT (Hajič et al., 2020)
- Czech-PCEDT (Nedoluzhko et al., 2016)
- English-GUM (Zeldes, 2017)
- English-LitBank (Bamman et al., 2019)
- 2018)
- , 2018)
- 0)
- 22)
- wegian-NynorskNARC (Mæhlum et al., 2022)
- Spanish-AnCora (Recasens and Martí, 2010)
- Catalan-AnCora (Recasens and Martí, 2010)

- mention heads are marked explicitly
- coreference annotation created without using hand-annotated syntactic structures

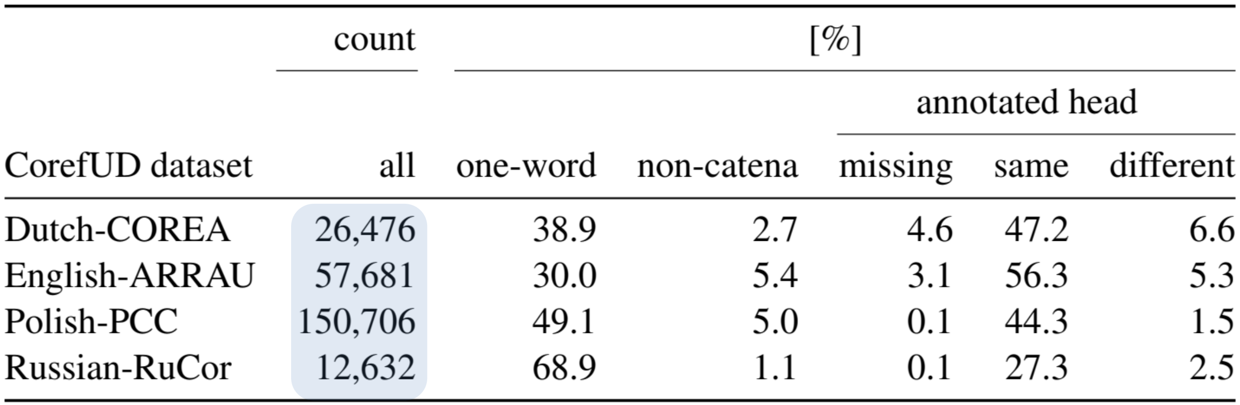
- French-Democrat (Landragin, 2021)
- **Polish-PCC (Ogrodniczuk et al., 2013)**
- Lithuanian-LCC (Žitkus and Butkienė, 2018)
- **Russian-RuCor (Toldova et al., 2014)**
- Hungarian-SzegedKoref (Vincze et al., 2018)
- Hungarian-KorKor (Vadász, 2022)
- Turkish-ITCC (Pamay and Eryiğit, 2018)
- Ancient Greek-PROIEL (Haug and Jøhndal, 2008)
- Old Church Slavonic-PROIEL (Haug and Jøhndal, 2008)
- Ancient Hebrew-PTNK (Swanson et al., 2024)

## non-public datasets:

- English-OntoNotes (Weischedel et al., 2011)
- **English-ARRAU (Uryupina et al., 2020)**
- **Dutch-COREA (Hendrickx et al., 2008)**
- English-PCEDT (Nedoluzhko et al., 2016)

# Are syntactic heads in UD equivalent to the head mentions in coreference annotation?

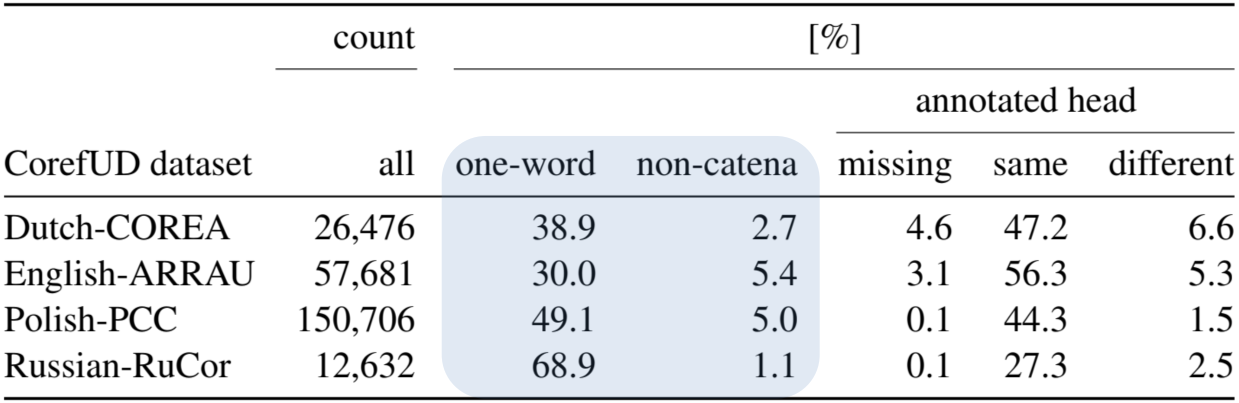
we analyzed mentions, where mention heads differ from dependency heads



| | count | [%] | | | | |
|---|---|---|---|---|---|---|
| | | | | annotated head | | |
| CorefUD dataset | all | one-word | non-catena | missing | same | different |
| Dutch-COREA | 26,476 | 38.9 | 2.7 | 4.6 | 47.2 | 6.6 |
| English-ARRAU | 57,681 | 30.0 | 5.4 | 3.1 | 56.3 | 5.3 |
| Polish-PCC | 150,706 | 49.1 | 5.0 | 0.1 | 44.3 | 1.5 |
| Russian-RuCor | 12,632 | 68.9 | 1.1 | 0.1 | 27.3 | 2.5 |

# Are syntactic heads in UD equivalent to the head mentions in coreference annotation?

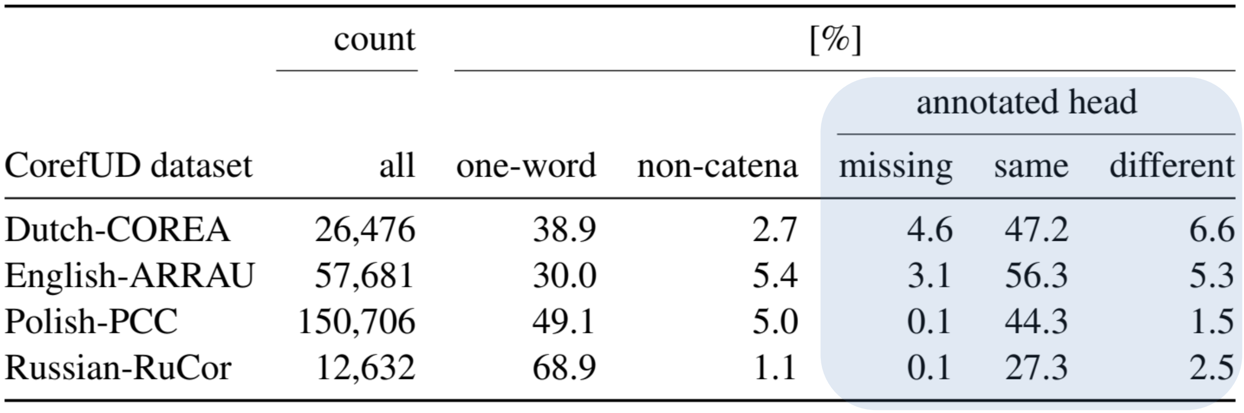
we analyzed mentions, where mention heads differ from dependency heads



| | count | [%] | | | | |
|---|---|---|---|---|---|---|
| | | | | annotated head | | |
| CorefUD dataset | all | one-word | non-catena | missing | same | different |
| Dutch-COREA | 26,476 | 38.9 | 2.7 | 4.6 | 47.2 | 6.6 |
| English-ARRAU | 57,681 | 30.0 | 5.4 | 3.1 | 56.3 | 5.3 |
| Polish-PCC | 150,706 | 49.1 | 5.0 | 0.1 | 44.3 | 1.5 |
| Russian-RuCor | 12,632 | 68.9 | 1.1 | 0.1 | 27.3 | 2.5 |

# Are syntactic heads in UD equivalent to the head mentions in coreference annotation?

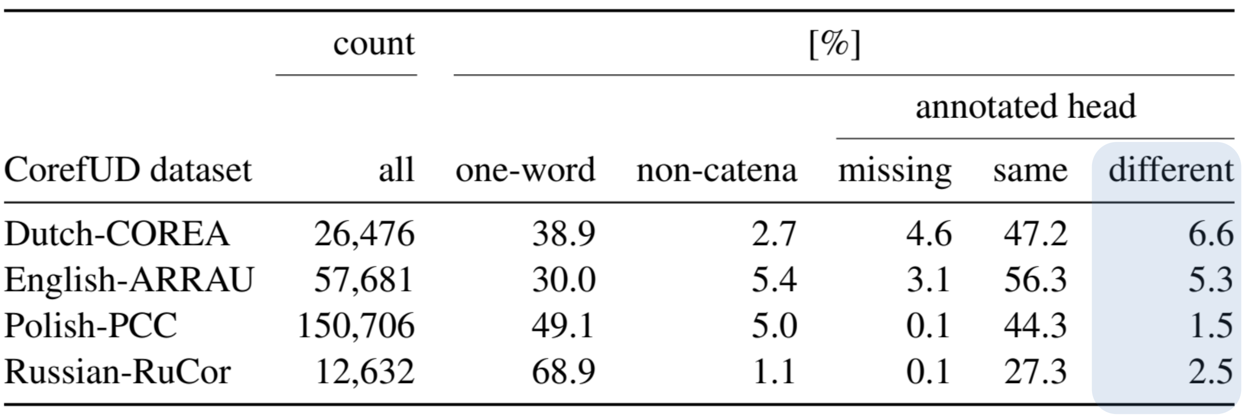
we analyzed mentions, where mention heads differ from dependency heads



| | count | [%] | | | | |
|---|---|---|---|---|---|---|
| | | | | annotated head | | |
| CorefUD dataset | all | one-word | non-catena | missing | same | different |
| Dutch-COREA | 26,476 | 38.9 | 2.7 | 4.6 | 47.2 | 6.6 |
| English-ARRAU | 57,681 | 30.0 | 5.4 | 3.1 | 56.3 | 5.3 |
| Polish-PCC | 150,706 | 49.1 | 5.0 | 0.1 | 44.3 | 1.5 |
| Russian-RuCor | 12,632 | 68.9 | 1.1 | 0.1 | 27.3 | 2.5 |

# Are syntactic heads in UD equivalent to the head mentions in coreference annotation?

we analyzed mentions, where mention heads differ from dependency heads

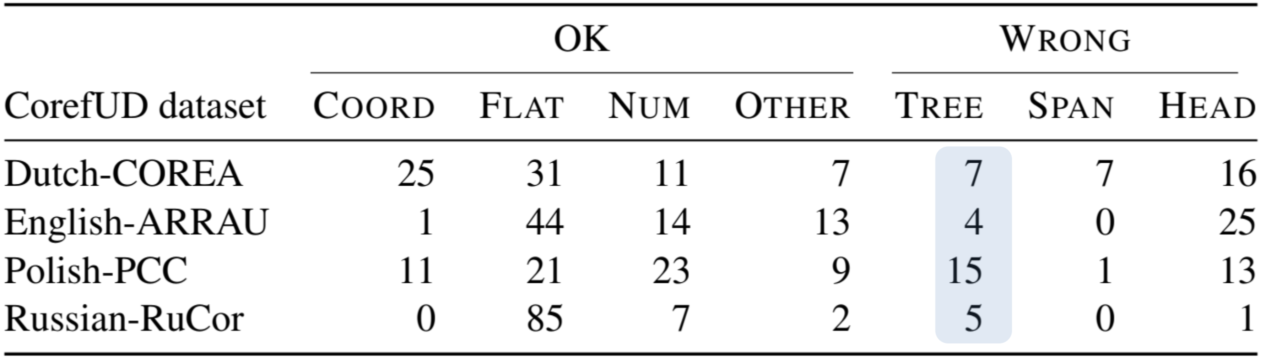


| | count | [%] | | | | |
|---|---|---|---|---|---|---|
| | | | | annotated head | | |
| CorefUD dataset | all | one-word | non-catena | missing | same | different |
| Dutch-COREA | 26,476 | 38.9 | 2.7 | 4.6 | 47.2 | 6.6 |
| English-ARRAU | 57,681 | 30.0 | 5.4 | 3.1 | 56.3 | 5.3 |
| Polish-PCC | 150,706 | 49.1 | 5.0 | 0.1 | 44.3 | 1.5 |
| Russian-RuCor | 12,632 | 68.9 | 1.1 | 0.1 | 27.3 | 2.5 |

sampled 100 examples to be analyzed

- WRONG
  - WRONG-TREE – parsing error
  - WRONG-SPAN – error in coreference annotation
  - WRONG-HEAD – wrongly identified head
- OK
  - OK-COORD – coordination issues
  - OK-FLAT – names, foreign phrases, dates
  - OK-NUM - numerals
  - OK-OTHER

# Are syntactic heads in UD equivalent to the head mentions in coreference annotation?

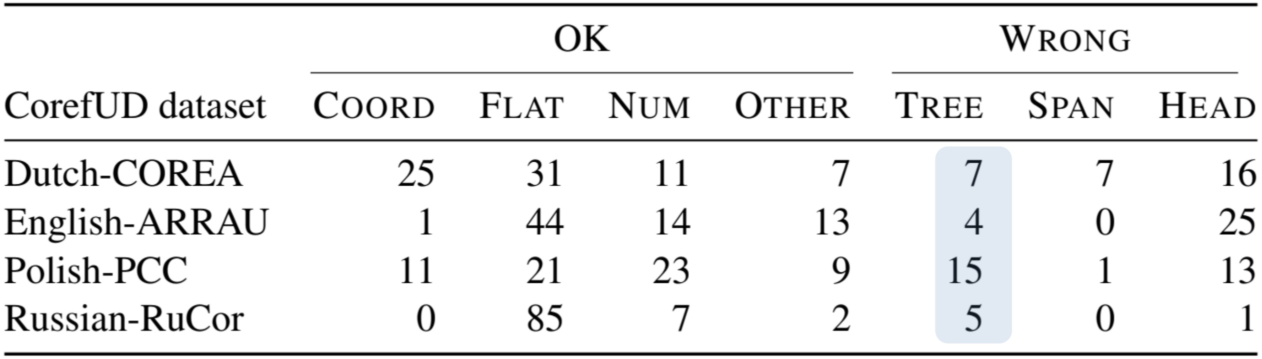


| | OK | | | | WRONG | | |
|---|---|---|---|---|---|---|---|
| CorefUD dataset | COORD | FLAT | NUM | OTHER | TREE | SPAN | HEAD |
| Dutch-COREA | 25 | 31 | 11 | 7 | 7 | 7 | 16 |
| English-ARRAU | 1 | 44 | 14 | 13 | 4 | 0 | 25 |
| Polish-PCC | 11 | 21 | 23 | 9 | 15 | 1 | 13 |
| Russian-RuCor | 0 | 85 | 7 | 2 | 5 | 0 | 1 |

sampled 100 examples to be analyzed

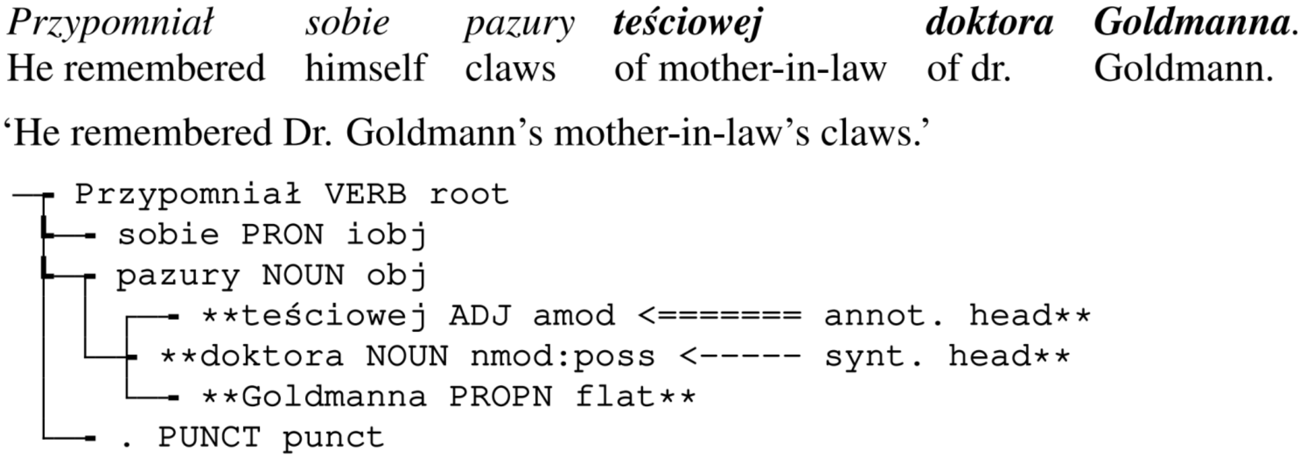
- WRONG
  - WRONG-TREE – parsing error
  - WRONG-SPAN – error in coreference annotation
  - WRONG-HEAD – wrongly identified head
- OK
  - OK-COORD – coordination issues
  - OK-FLAT – names, foreign phrases, dates
  - OK-NUM - numerals
  - OK-OTHER

# Are syntactic heads in UD equivalent to the head mentions in coreference annotation?



| | OK | | | | WRONG | | |
|---|---|---|---|---|---|---|---|
| CorefUD dataset | COORD | FLAT | NUM | OTHER | TREE | SPAN | HEAD |
| Dutch-COREA | 25 | 31 | 11 | 7 | 7 | 7 | 16 |
| English-ARRAU | 1 | 44 | 14 | 13 | 4 | 0 | 25 |
| Polish-PCC | 11 | 21 | 23 | 9 | 15 | 1 | 13 |
| Russian-RuCor | 0 | 85 | 7 | 2 | 5 | 0 | 1 |

- relatively low number of mismatches caused by wrong parsing
  - parsing of NPs is rather simple
- errors in specific constructions

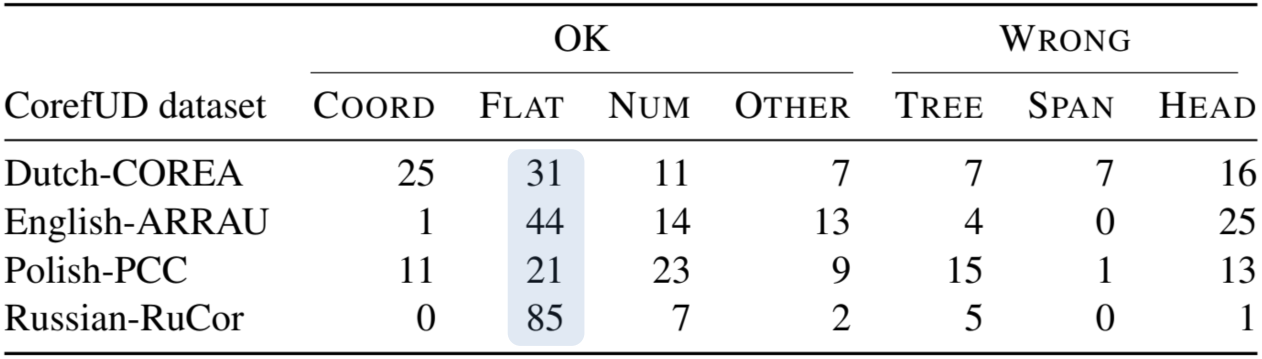


| *Przypomniał* | *sobie* | *pazury* | ***teściowej*** | ***doktora*** | ***Goldmanna***. |
|---|---|---|---|---|---|
| He remembered | himself | claws | of mother-in-law | of dr. | Goldmann. |

‘He remembered Dr. Goldmann’s mother-in-law’s claws.’

```
─┬ Przypomniał VERB root
 ├─ sobie PRON iobj
 ├─┬ pazury NOUN obj
 │ │  ┌─ **teściowej ADJ amod <======= annot. head**
 │ └─┼ **doktora NOUN nmod:poss <----- synt. head**
 │    └─ **Goldmanna PROPN flat**
 └─ . PUNCT punct
```

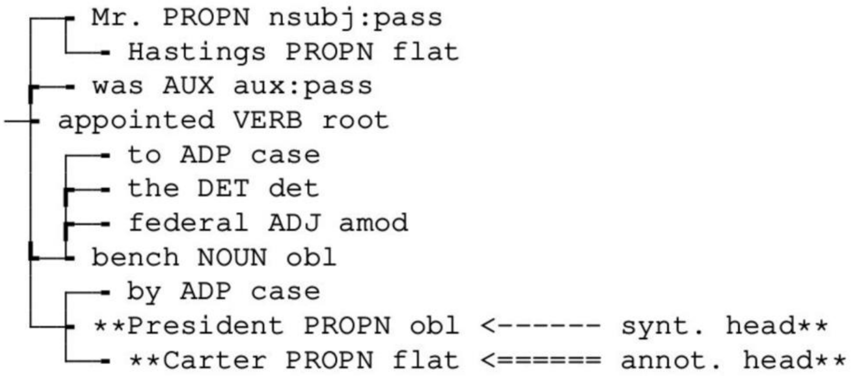
# Are syntactic heads in UD equivalent to the head mentions in coreference annotation?



| | OK | | | | WRONG | | |
|---|---|---|---|---|---|---|---|
| CorefUD dataset | COORD | FLAT | NUM | OTHER | TREE | SPAN | HEAD |
| Dutch-COREA | 25 | 31 | 11 | 7 | 7 | 7 | 16 |
| English-ARRAU | 1 | 44 | 14 | 13 | 4 | 0 | 25 |
| Polish-PCC | 11 | 21 | 23 | 9 | 15 | 1 | 13 |
| Russian-RuCor | 0 | 85 | 7 | 2 | 5 | 0 | 1 |

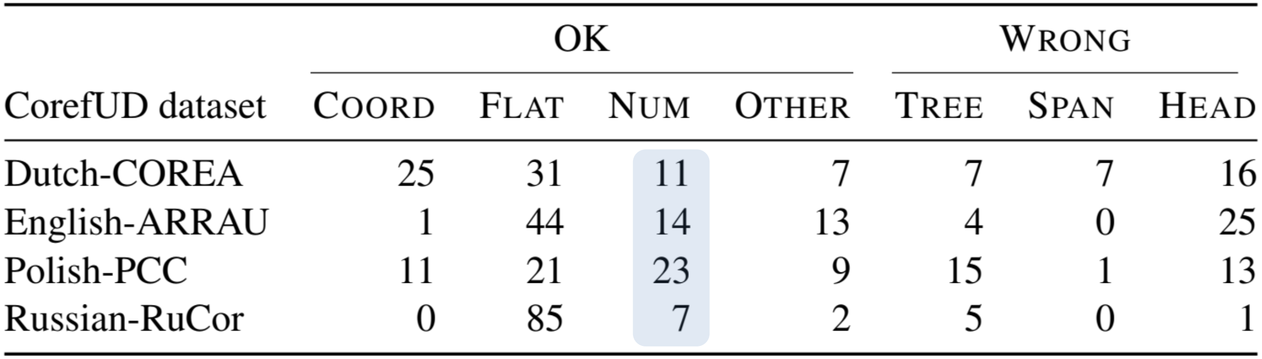
- form a great deal of OK category
- head in “common noun + proper noun” constructions:
  - proper noun in coref. corpora vs. first word according to UD

*Mr. Hastings was appointed to the federal bench by President Carter.*



```
 ┌──┮ Mr. PROPN nsubj:pass
 │  └─╼ Hastings PROPN flat
 ┢──╼ was AUX aux:pass
─┾ appointed VERB root
 │  ┌─╼ to ADP case
 │  ┢─╼ the DET det
 │  ┢─╼ federal ADJ amod
 ┡──┹ bench NOUN obl
 │  ┌─╼ by ADP case
 └──┾ **President PROPN obl <------ synt. head**
    └─╼ **Carter PROPN flat <====== annot. head**
```

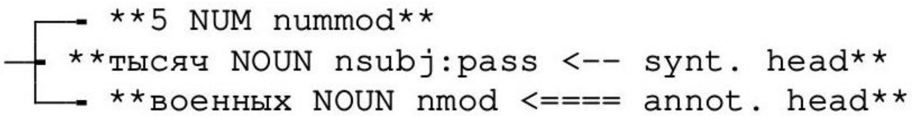
# Are syntactic heads in UD equivalent to the head mentions in coreference annotation?



| | OK | | | | WRONG | | |
|---|---|---|---|---|---|---|---|
| CorefUD dataset | COORD | FLAT | NUM | OTHER | TREE | SPAN | HEAD |
| Dutch-COREA | 25 | 31 | 11 | 7 | 7 | 7 | 16 |
| English-ARRAU | 1 | 44 | 14 | 13 | 4 | 0 | 25 |
| Polish-PCC | 11 | 21 | 23 | 9 | 15 | 1 | 13 |
| Russian-RuCor | 0 | 85 | 7 | 2 | 5 | 0 | 1 |

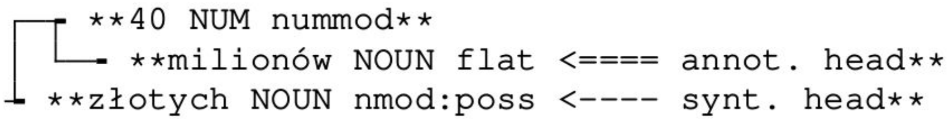
- mention heads inconsistent *across* and also *within* the datasets
- inconsistencies in parse trees (even in the gold ones)
- syntactically governing numerals (e.g. “one of the candidates”) and containers (e.g. “group of tourists”)

| | | |
|---|---|---|
| ***5*** | ***тысяч*** | ***военных*** |
| 5 | thousand.GEN.PL | soldiers |



```
**5 NUM nummod**
**тысяч NOUN nsubj:pass <-- synt. head**
**военных NOUN nmod <==== annot. head**
```

| | | |
|---|---|---|
| ***40*** | ***milionów*** | ***złotych*** |
| 40 | million | złoty |



```
**40 NUM nummod**
**milionów NOUN flat <==== annot. head**
**złotych NOUN nmod:poss <---- synt. head**
```

# Are syntactic heads in UD equivalent to the head mentions in coreference annotation?

*Hypothesis:*

mention heads in annotations and syntactic heads in syntactic trees are closely related and compatible

*Observations and Conclusions*:

- the heads are mostly aligned
- mismatches due to parsing errors are low (4%–15%)
- many mismatches are due to a technical reason
- hence, it can be transformed automatically

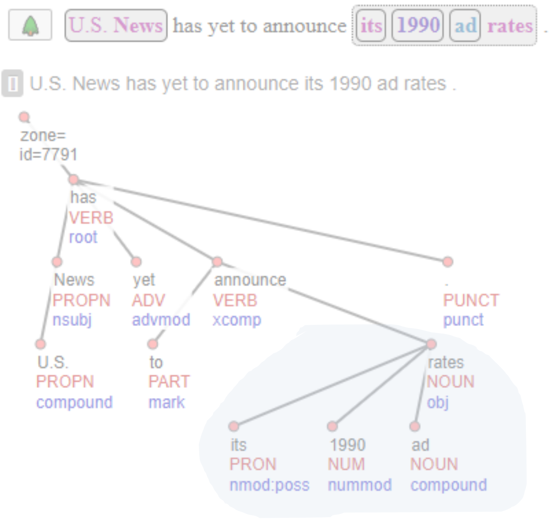
# Part 2

# CorefUD and Treebanks

### UD tree structures and mention spans

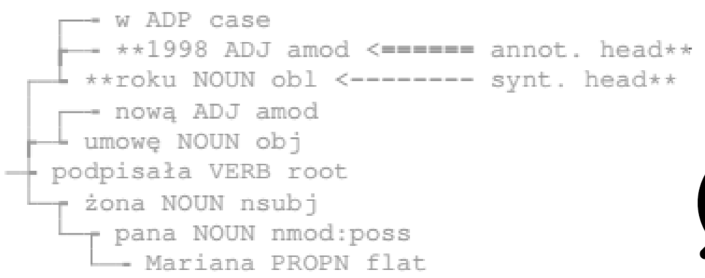
### Syntactic heads in UD and head mentions

### Empty nodes in UD and zero anaphora

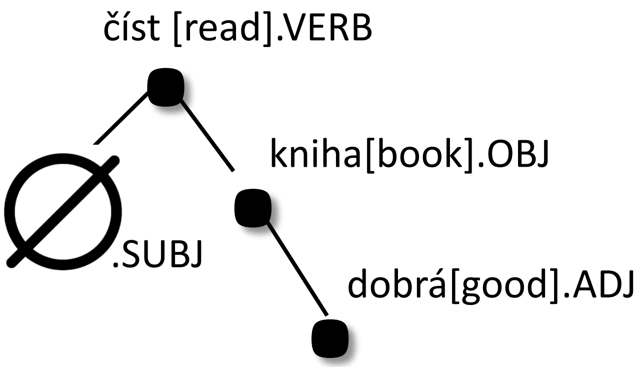


U.S. News has yet to announce its 1990 ad rates .

U.S. News has yet to announce its 1990 ad rates .

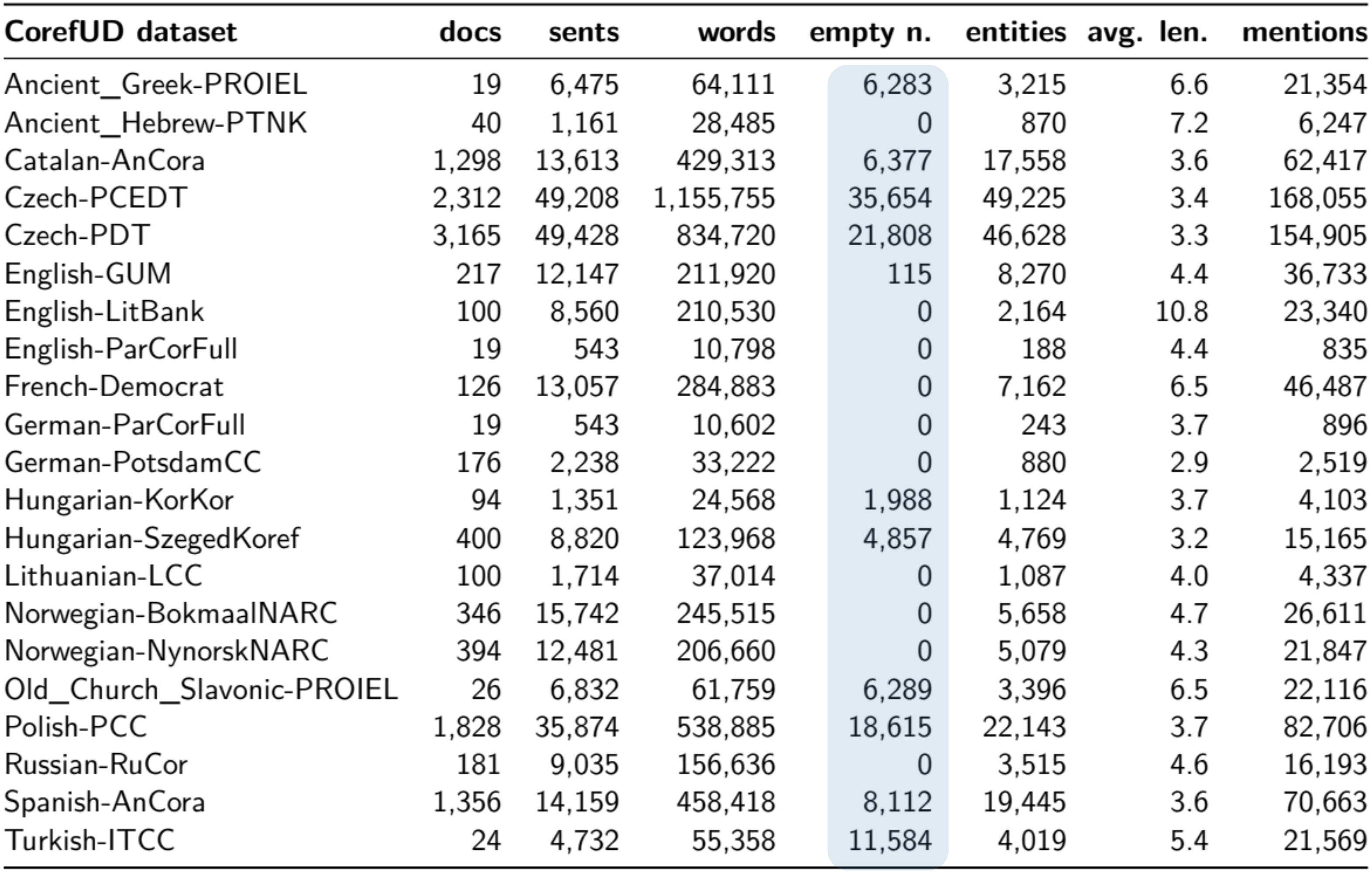


```
    ┌── w ADP case
    ├── **1998 ADJ amod <====== annot. head**
  ┌─┴ **roku NOUN obl <-------- synt. head**
  │ ┌── nową ADJ amod
  ├─┴ umowę NOUN obj
──┤ podpisała VERB root
  └─┬ żona NOUN nsubj
    └─┬ pana NOUN nmod:poss
      └── Mariana PROPN flat
```



# Empty nodes in CorefUD 1.2

- zeros are integral part of some of the datasets
- represented using empty nodes from enhanced UD

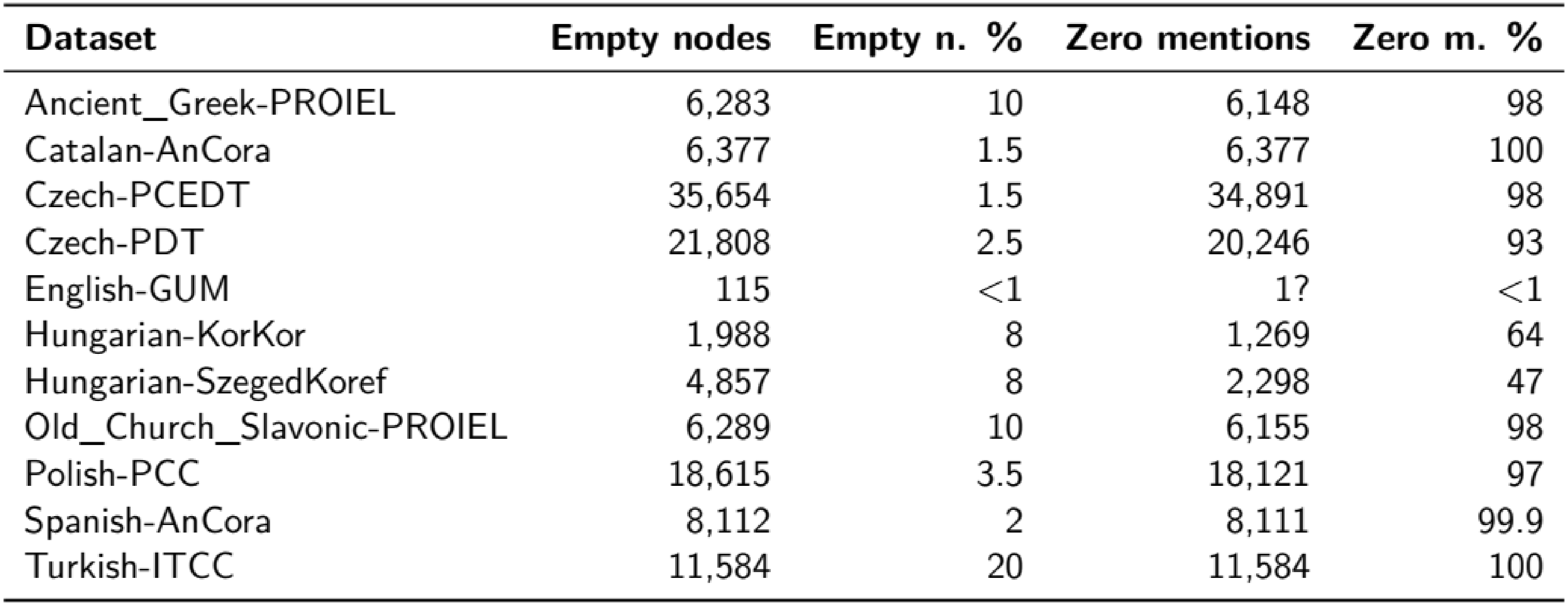


| CorefUD dataset | docs | sents | words | empty n. | entities | avg. len. | mentions |
|---|---:|---:|---:|---:|---:|---:|---:|
| Ancient_Greek-PROIEL | 19 | 6,475 | 64,111 | 6,283 | 3,215 | 6.6 | 21,354 |
| Ancient_Hebrew-PTNK | 40 | 1,161 | 28,485 | 0 | 870 | 7.2 | 6,247 |
| Catalan-AnCora | 1,298 | 13,613 | 429,313 | 6,377 | 17,558 | 3.6 | 62,417 |
| Czech-PCEDT | 2,312 | 49,208 | 1,155,755 | 35,654 | 49,225 | 3.4 | 168,055 |
| Czech-PDT | 3,165 | 49,428 | 834,720 | 21,808 | 46,628 | 3.3 | 154,905 |
| English-GUM | 217 | 12,147 | 211,920 | 115 | 8,270 | 4.4 | 36,733 |
| English-LitBank | 100 | 8,560 | 210,530 | 0 | 2,164 | 10.8 | 23,340 |
| English-ParCorFull | 19 | 543 | 10,798 | 0 | 188 | 4.4 | 835 |
| French-Democrat | 126 | 13,057 | 284,883 | 0 | 7,162 | 6.5 | 46,487 |
| German-ParCorFull | 19 | 543 | 10,602 | 0 | 243 | 3.7 | 896 |
| German-PotsdamCC | 176 | 2,238 | 33,222 | 0 | 880 | 2.9 | 2,519 |
| Hungarian-KorKor | 94 | 1,351 | 24,568 | 1,988 | 1,124 | 3.7 | 4,103 |
| Hungarian-SzegedKoref | 400 | 8,820 | 123,968 | 4,857 | 4,769 | 3.2 | 15,165 |
| Lithuanian-LCC | 100 | 1,714 | 37,014 | 0 | 1,087 | 4.0 | 4,337 |
| Norwegian-BokmaalNARC | 346 | 15,742 | 245,515 | 0 | 5,658 | 4.7 | 26,611 |
| Norwegian-NynorskNARC | 394 | 12,481 | 206,660 | 0 | 5,079 | 4.3 | 21,847 |
| Old_Church_Slavonic-PROIEL | 26 | 6,832 | 61,759 | 6,289 | 3,396 | 6.5 | 22,116 |
| Polish-PCC | 1,828 | 35,874 | 538,885 | 18,615 | 22,143 | 3.7 | 82,706 |
| Russian-RuCor | 181 | 9,035 | 156,636 | 0 | 3,515 | 4.6 | 16,193 |
| Spanish-AnCora | 1,356 | 14,159 | 458,418 | 8,112 | 19,445 | 3.6 | 70,663 |
| Turkish-ITCC | 24 | 4,732 | 55,358 | 11,584 | 4,019 | 5.4 | 21,569 |

# Zeros in the public edition of CorefUD 1.2

- **Czech-PDT (Hajič et al., 2020)**
- **Czech-PCEDT (Nedoluzhko et al., 2016)**
- **English-GUM (Zeldes, 2017)**
- English-LitBank (Bamman et al., 2019)
- English-ParCorFull (Lapshinova-Koltunski et al., 2018)
- German-ParCorFull (Lapshinova-Koltunski et al., 2018)
- German-PotsdamCC (Bourgonje and Stede, 2020)
- Norwegian-BokmaalNARC (Mæhlum et al., 2022)
- Norwegian-NynorskNARC (Mæhlum et al., 2022)
- **Spanish-AnCora (Recasens and Martí, 2010)**
- **Catalan-AnCora (Recasens and Martí, 2010)**
- French-Democrat (Landragin, 2021)
- **Polish-PCC (Ogrodniczuk et al., 2013)**
- Lithuanian-LCC (Žitkus and Butkienė, 2018)
- Russian-RuCor (Toldova et al., 2014)
- **Hungarian-SzegedKoref (Vincze et al., 2018)**
- **Hungarian-KorKor (Vadász, 2022)**
- **Turkish-ITCC (Pamay and Eryiğit, 2018)**
- **Ancient Greek-PROIEL (Haug and Jøhndal, 2008)**
- **Old Church Slavonic-PROIEL (Haug and Jøhndal, 2008)**
- Ancient Hebrew-PTNK (Swanson et al., 2024)

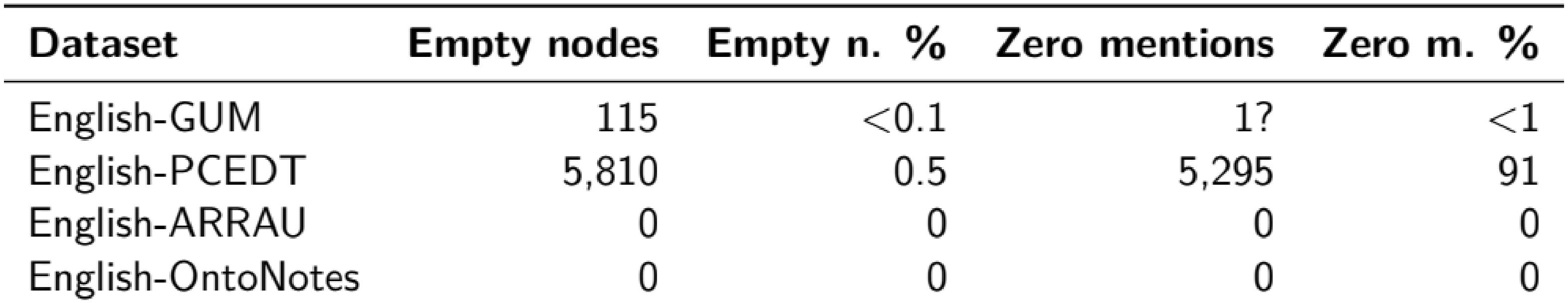
# How do zeros in coreference relate to empty nodes in syntactic structures?



| Dataset | Empty nodes | Empty n. % | Zero mentions | Zero m. % |
|---|---:|---:|---:|---:|
| Ancient_Greek-PROIEL | 6,283 | 10 | 6,148 | 98 |
| Catalan-AnCora | 6,377 | 1.5 | 6,377 | 100 |
| Czech-PCEDT | 35,654 | 1.5 | 34,891 | 98 |
| Czech-PDT | 21,808 | 2.5 | 20,246 | 93 |
| English-GUM | 115 | <1 | 1? | <1 |
| Hungarian-KorKor | 1,988 | 8 | 1,269 | 64 |
| Hungarian-SzegedKoref | 4,857 | 8 | 2,298 | 47 |
| Old_Church_Slavonic-PROIEL | 6,289 | 10 | 6,155 | 98 |
| Polish-PCC | 18,615 | 3.5 | 18,121 | 97 |
| Spanish-AnCora | 8,112 | 2 | 8,111 | 99.9 |
| Turkish-ITCC | 11,584 | 20 | 11,584 | 100 |

- difference across languages
- difference across annotation approaches in what is annotated as zero mentions

# Empty nodes and zero mentions in English datasets in CorefUD 1.2 numbers and examples



| Dataset | Empty nodes | Empty n. % | Zero mentions | Zero m. % |
|---|---:|---:|---:|---:|
| English-GUM | 115 | <0.1 | 1? | <1 |
| English-PCEDT | 5,810 | 0.5 | 5,295 | 91 |
| English-ARRAU | 0 | 0 | 0 | 0 |
| English-OntoNotes | 0 | 0 | 0 | 0 |

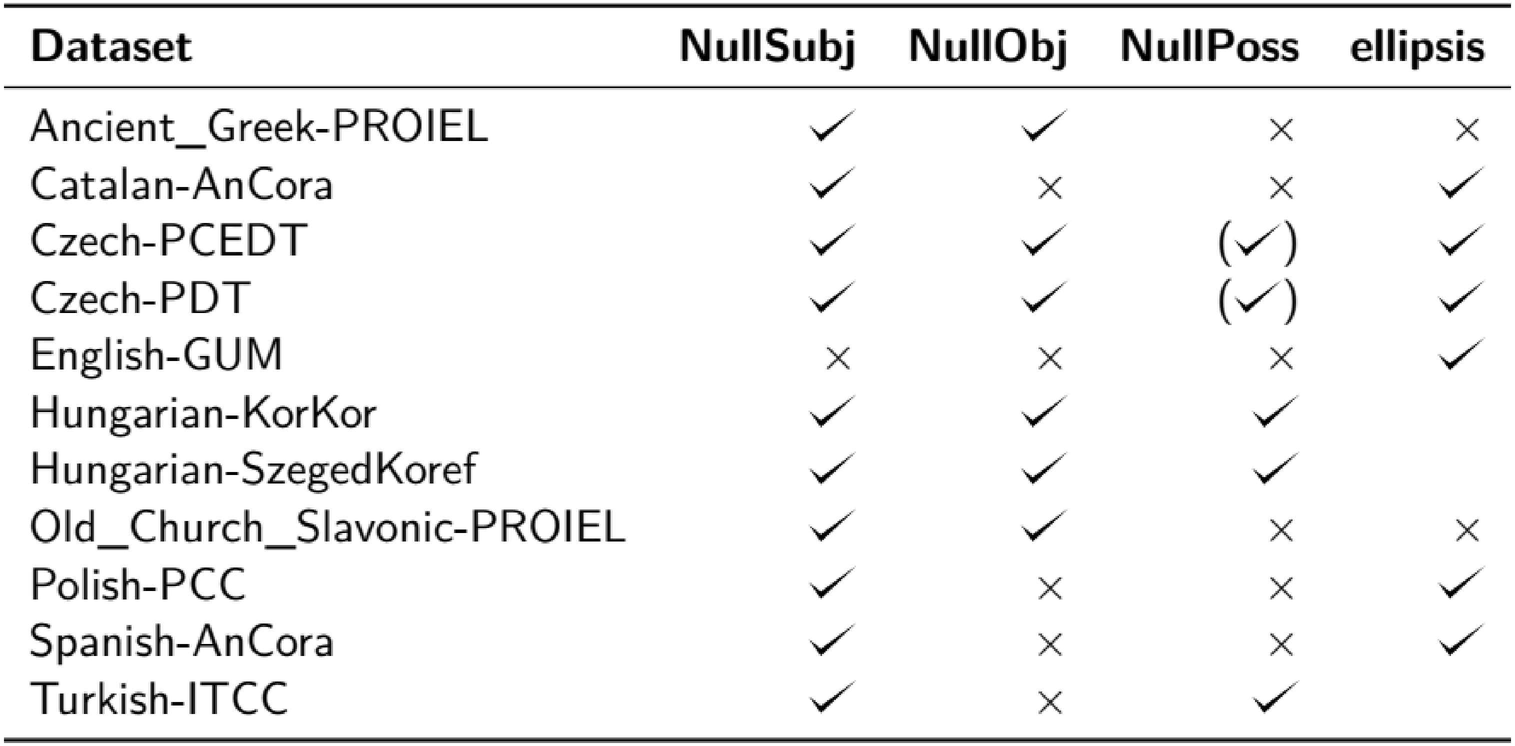
*After the death of his mother in 1846 and [the death of] his father in 1848, he sailed west, arriving in San Francisco possibly in November 1849.* (English_GUM)

*Four of the five surviving workers have asbestos-related diseases, including three [workers] with recently diagnosed cancer.* (English_PCEDT)

# Distribution of zero mentions

Reconstructed zeros for coreference may have different “functions”

- zero subjects
- zero objects
- possessives
- ellipses
- …



| Dataset | NullSubj | NullObj | NullPoss | ellipsis |
|---|---|---|---|---|
| Ancient_Greek-PROIEL | ✓ | ✓ | × | × |
| Catalan-AnCora | ✓ | × | × | ✓ |
| Czech-PCEDT | ✓ | ✓ | (✓) | ✓ |
| Czech-PDT | ✓ | ✓ | (✓) | ✓ |
| English-GUM | × | × | × | ✓ |
| Hungarian-KorKor | ✓ | ✓ | ✓ | |
| Hungarian-SzegedKoref | ✓ | ✓ | ✓ | |
| Old_Church_Slavonic-PROIEL | ✓ | ✓ | × | × |
| Polish-PCC | ✓ | × | × | ✓ |
| Spanish-AnCora | ✓ | × | × | ✓ |
| Turkish-ITCC | ✓ | × | ✓ | |

# Reconstructed zero subjects
# Examples

*Bilmiyorum* $Ø_{subj}$ . ‘I don't know’ (Turkish_ITCC)

*Przemysł atomowy zapowiedział, że* $Ø_{subj}$ *wystąpi o odszkodowania w miliardowej wysokości*. ‘The nuclear industry has announced that [it] will seek billions in damages.’ (Polish_PCC)

*A szeleteket ízlés szerint vajjal vagy olívaolajjal* $Ø_{subj}$ *kenik meg*. ‘[They] spread the slices with butter, sandwich spread, or olive oil according to taste.’ (Hungarian_KorKor)

$Ø_{subj}$ *οὐκ ἦλθον καταλῦσαι ἀλλὰ πληρῶσαι*. ‘[I] have not come to abolish but to fulfill.’ (Ancient_Greek_PROIEL)

*Avšak nyní se již* $Ø_{subj}$ *těší na své zimní zasedání*. ‘But now [they] are looking forward to their winter session . (Czech_PCEDT)

# Reconstructed zero objects
## Examples

*Μὴ νομίσητε ὅτι ἦλθον καταλῦσαι τὸν νόμον ἢ τοὺς προφήτας. οὐκ ἦλθον Ø*$_{obj}$ *καταλῦσαι ἀλλὰ Ø πληρῶσαι.* ‘Do not think that I have come to abolish the law or the prophets. I have not come to abolish but to fulfill.’ `(Ancient_Greek_PROIEL)`

*Az angol John Montaguról, Sandwich negyedik grófjáról nevezték el Ø*$_{subj}$ *Ø*$_{obj}$ ."It was named after the English John Montagu, the fourth Earl of Sandwich.“ `(Hungarian_KorKor)`

*видѣвъ Ø*$_{obj}$ *же воѥвода избѣже вънъ* ‘As [he] saw [this] the commander went out’ `(Old_Church_Slavonic_PROIEL)`

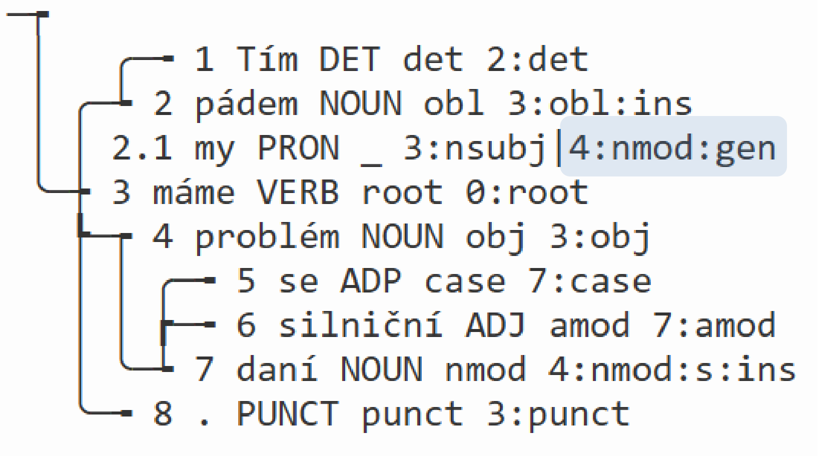
*Ve spojení s tím, že vyrovnávání obchodů není tak pružné, jak bychom si představovali Ø*$_{obj}$*, to vede ke dvěma věcem*. ‘Coupled with the fact that trade balancing is not as flexible as we would imagine [it], this leads to two things.’ `(Czech_PDT)`

# Reconstructed zero possessives
## Examples

*Göğsü* $Ø_{poss}$ *[onun] körük gibi inip kalkıyordu o*. ‘Her chest was rising and falling like a bellows.’ `(Turkish_ITCC)`

*Úgy tűnik, hogy a kérése* $Ø_{poss}$ *nyitott fülekre talált, vagy legalábbis az orosz törvényalkotók hasonlóan gondolkodnak*. ‘It seems that her request has fallen on deaf ears, or at least Russian lawmakers think similarly.’ `(Hungarian_KorKor)`

*Tím pádem* $Ø_{poss}$ *máme problém se silniční daní*. ‘So, we have a problem with road tax.’ `(Czech_PDT)`



```
1 Tím DET det 2:det
2 pádem NOUN obl 3:obl:ins
2.1 my PRON _ 3:nsubj|4:nmod:gen
3 máme VERB root 0:root
4 problém NOUN obj 3:obj
5 se ADP case 7:case
6 silniční ADJ amod 7:amod
7 daní NOUN nmod 4:nmod:s:ins
8 . PUNCT punct 3:punct
```

# Part 3

# Conclusions and Discussion

# Discussion and conclusions

- Initial hypotheses: deep integration of coreference and syntax
- “Positive” findings
  - mention spans usually correspond to subtrees
  - mention heads usually correspond to heads in UD trees
  - more questions with zeros, looking for reasons
- “Negative” findings
  - CRAC shared tasks showed that systems don’t need syntactic information

# Discussion and conclusions

- Observations:
  - there are linguistic phenomena that are relevant – but at the same time difficult to handle – both for dependency syntax and for coreference annotation schemes and it seems natural to use the same annotation conventions for both.
    - coordination and other paratactic relations, zeros, named entities etc.

- syntactic viewpoint increases annotation consistency
  - e.g. in defining mention spans or heads

# Future plans for near future

- extension of CorefUD is always welcome
- detailed improvement of many special issues in CorefUD is also very welcome
- definitely more research is needed on data,
  - zeros
  - other phenomena, e.g. coordination
- CRAC2025 shared task: focus on LLMs

# Acknowledgements

**I thank ChatGPT for inventing (almost) the same topic for my presentation!**

**prompt** (originally in Czech):
What is the best topic for Anna Nedoluzhko to present as a key-note talk at the conference Treebanks and Linguistic Theories?

**ChatGPT said**:
"Coreference and Syntax: Bridging Treebanks and Cross-Linguistic Phenomena“
**Abstract:**
The lecture could present the relationship between dependency syntax and coreference based on data from CorefUD. It would focus on questions such as:

- How syntactic structures influence coreference relations.
- Multilingual challenges in annotating coreference and dependency structures.
- The role of zero elements (null anaphora, omitted arguments) in tree representations and their relationship to coreference chains.

**Reason:**
This topic connects the core areas of TLT—tree structures and their application in linguistic theory—with practical and theoretical insights into coreference.

# Acknowledgements

**Thank you for attention!**

**Questions?**

# References

All information about CorefUD, shared tasks and guidelines may be found here:

**https://ufal.mff.cuni.cz/corefud**

Papers most relevant for the talk:

- Nedoluzhko Anna, Novák Michal, Popel Martin, Žabokrtský Zdeněk, Zeldes Amir, Zeman Daniel: **CorefUD 1.0: Coreference Meets Universal Dependencies**. In: LREC 2022
- Popel Martin, Žabokrtský Zdeněk, Nedoluzhko Anna, Novák Michal, Zeman Daniel: **Do UD Trees Match Mention Spans in Coreference Annotations?**. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*, Copyright © Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-955917-10-0, pp. 3570-3576, 2021
- Nedoluzhko Anna, Novák Michal, Popel Martin, Žabokrtský Zdeněk, Zeman Daniel: **Is one head enough? Mention heads in coreference annotations compared with UD-style heads**. In: Depling, SyntaxFest 2021