CorefUD 0.1 – a pilot experiment on harmonizing coreference datasets for 11 languages

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Daniel Zeman

April 19, 2021
Outline

Coreference in a nutshell

Variability of existing coreference data resources

Our harmonization scheme

Collection CorefUD 0.1

Application Programming Interface for CorefUD data

Case study 1: discontinuous mentions

Case study 2: inducing linear mentions from trees

Conclusions
Coreference in a nutshell
Examples first

(1) **Peter** has eaten all apples **himself**.

ANTECEDENT   ANAPHOR

(2) Don’t eat the apples which are mine!

(3) **This apple** is mine. Don’t eat it!

(4) Mary gave **Peter** an apple. Steve gave **him** another one. **Peter** took them and left.
Coreference in Prague

Long tradition of coreference studies, beginning from early eighties

- 1999 - first tectogrammatic manual, including coreference (btw ord used)
- 2003 - pilot coreference annotation, link-based representation (t-node id)
- 2006 - PDT 2.0 incl. 40k coref links published
- 2006-2011 - extension of textual coreference to full NPs, annotation of bridging
- 2012 - coreference in PCEDT annotated in the (simplified) PDT style
- 2013 - PDT 3.0, coreference of 1st and 2nd person pronouns added

Distinctive features in comparison with most other coreference projects abroad:

- grammatical and textual coreference distinguished
- coreference inseparable from syntax
(1) **Peter** has eaten all apples **himself**.

(2) Don’t eat **the apples** which are mine!

(3) **This apple** is mine. Don’t eat **it**!

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

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

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

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

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

(9) There are **a lot of apples** in the bin. **Each** has a worn. (bound anaphora)

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

(11) **This red apple** is a symbol of happiness. (predication)
Fuzzy boundaries between syntax, coreference and bridging

Figure 1: Types of possible relations between referring expressions, including borderline types.
Variability of existing coreference data resources
Selection criteria

• We are aware of some 50 data resources in total
• Clearly beyond our capacity → sampling was inescapable
• A mixture of selection criteria:
  • data availability (the easier access, the better, personal communication needed in some cases)
  • license (the freer, the better)
  • size (the bigger, the better)
  • diversity of the selected sample (the more diverse, the better)
  • a few examples of parallel datasets desired too
  • at this step only languages whose writing systems are readable to us
17 coreference datasets included in our harmonization study

free licenses

- Czech-PDT (Hajič et al., 2020)
- Czech-PCEDT (Nedoluzhko et al., 2016)
- English-GUM (Zeldes, 2017)
- German-PotsdamCC (Bourgonje and Stede, 2020)
- French-Democrat (Landragin, 2016)
- English-ParCorFull (Lapshinova-Koltunski et al., 2018)
- German-ParCorFull (Lapshinova-Koltunski et al., 2018)

non-free licenses

- Spanish-AnCora (Recasens and Martí, 2010)
- Catalan-AnCora (Recasens and Martí, 2010)
- Polish-PCC (Ogrodniczuk et al., 2013)
- Hungarian-SzegedKoref (Vincze et al., 2018)
- Lithuanian-LCC (Žitkus and Butkienė, 2018)
- Russian-RuCor (Toldova et al., 2014)
- English-OntoNotes (Weischedel et al., 2011)
- English-ARRAU (Uryupina et al., 2020)
- Dutch-COREA (Hendrickx et al., 2008)
- English-PCEDT (Nedoluzhko et al., 2016)
Diversity in existing resources

• Which domain (news, dialogues, stories...)?
• Which relations are annotated?
• What is considered to be a mention?
• Which additional linguistic information resources have (lemmatization, POS tagging, sentence segmentation, tokenization, syntactic trees, document boundaries, etc.)?
two frequent solutions:

- **cluster-based** grouping of mentions
  - coreferential mentions marked (coindexed) by the same cluster identifier
  - slightly prevailing approach

```
  c1 c1 c1 c2 c2 c2 c1
```

- **link-based** grouping of mentions
  - typically just a chain (in the order of linear precedence of mentions)
  - but sometimes tree-shaped (then not isomorphic with the cluster-based solution)

```
  o  o  o  o  o  o  o  o
  o o o o o o o o
```

Coref in nutshell | Resource variability | Our scheme | CorefUD 0.1 | API | Case study 1 | Case study 2 | Conclusions
--- | --- | --- | --- | --- | --- | --- | ---

Diversity in existing resources: representation of coreference
(12)  **Bob, my father-in-law**, got married yesterday.

solutions in datasets:

- ignore the relation
  - can be obtained from syntactic annotation (Czech-PDT, PCEDT)
  - cannot be obtained from syntactic annotation (French-Democrat, Lithuanian-LCC)
- mark it as a special type
  - within coreference annotation (English-Ontonotes)
  - out of coreference (Hungarian-SzegedKoref)
- include in the span of one mention (Polish-PCC, ParCorFull)
- annotate in the same way as identity coreference (Dutch-COREA)
## Diversity in existing resources: relations

<table>
<thead>
<tr>
<th>CorefUD dataset</th>
<th>Coref. grouping</th>
<th>Relations among mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cluster-based</td>
<td>link-based</td>
</tr>
<tr>
<td>Catalan-AnCora</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Czech-PCEDT</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Czech-PDT</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>English-GUM</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>English-ParCorFull</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>French-Democrat</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>German-ParCorFull</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>German-PotsdamCC</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Hungarian-SzegedKoref</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>Lithuanian-LCC</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Polish-PCC</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>Russian-RuCor</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>Spanish-AnCora</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>Dutch-COREA</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>English-ARRAU</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>English-OntoNotes</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>English-PCEDT</td>
<td>x</td>
<td>✓</td>
</tr>
</tbody>
</table>

### Coref in nutshell

- **Resource variability**
- **Our scheme**
- **CorefUD 0.1 API**
- **Case study 1**
- **Case study 2**
- **Conclusions**
Diversity in existing resources: mentions

What is considered to be a mention

- formal representation of mentions
  - linear
    - typically a single token identifier or an interval (from-to)
    - possibly discontinuous mentions (in some projects)
    - possibly with a distinguished head token (in some projects)
  - dependency-based
    - mention represented by its head token
    - complete span of the mention defined rather implicitly
  - constituency-based
    - mention represented by a syntactic phrase (such as NP)

- grammatical types of mentions
  - pronouns (different types), full NPs (specific, generic, etc.), VPs, pronominal adverbs
  - zeros (e.g. zero subjects), nominal ellipses
<table>
<thead>
<tr>
<th>Original corpus</th>
<th>Mention representation</th>
<th>Reconstructed zeros</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear span</td>
<td>Syn/sem. head</td>
</tr>
<tr>
<td>Catalan-AnCora</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Czech-PCEDT</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Czech-PDT</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>English-GUM</td>
<td>✓</td>
<td>(✓)</td>
</tr>
<tr>
<td>English-ParCorFull</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>French-Democrat</td>
<td>✓</td>
<td>(✓)</td>
</tr>
<tr>
<td>German-ParCorFull</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>German-PotsdamCC</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Hungarian-SzegedKoref</td>
<td>✓</td>
<td>(✓)</td>
</tr>
<tr>
<td>Lithuanian-LCC</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Polish-PCC</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Russian-RuCor</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Spanish-AnCora</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Dutch-COREA</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>English-ARRAU</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>English-OntoNotes</td>
<td>✓</td>
<td>(✓)</td>
</tr>
<tr>
<td>English-PCEDT</td>
<td>×</td>
<td>✓</td>
</tr>
</tbody>
</table>
Differences in realization of coreference across languages

(13) V roce 1985 přešla na bezkofeinovou recepturu, kterou používá pro svojí novou kolu. It switched to a caffeine-free formula using its new Coke in 1985.

∅ přešla na bezkofeinovou recepturu, kterou používá pro svojí kolu.

it switched to a caffeine-free formula [which] [it uses] [for] [self] Coke.

(14) Obyvatelé města si razili cestu ulicemi zasypanými sklem. Residents picked their way through glass-strewn streets.

Obyvatelé města si razili cestu ulicemi zasypanými sklem. Residents [of the city] [to themselves] picked their way

(Novák and Nedoluzhko, 2015)
Figure 2: Correspondences between Czech and English potentially coreferential expressions

(Novák and Nedoluzhko, 2015)
Previous harmonization efforts

- **wider perspective**: any multilingual corpus
  - *AnCora* – Spanish and Catalan (Recasens and Martí, 2010), *OntoNotes 5.0* – English, Chinese and Arabic (Weischedel et al., 2011), *PCEDT 2.0* – Czech and English (Nedoluzhko et al., 2016), *PAWS* – Czech, English, Polish and Russian (Nedoluzhko et al., 2018), *ParCor* – English and German (Guillou et al., 2014), or *ParCorFull* – English and German (Lapshinova-Koltunski et al., 2018)

- **narrower perspective**: merging multiple existing corpora under the same annotation scheme
  - not many attempts so far
  - **SemEval 2010 Shared task** on Coreference Resolution in Multiple Languages
    - five corpora in six languages: *AnCora* – Spanish and Catalan (Recasens and Martí, 2010), *KNACK-2002* – Dutch (Hoste and De Pauw, 2006), *OntoNotes 2.0* – English (Pradhan et al., 2007), *TüBa-D/Z Treebank* – German (Hinrichs et al., 2005) and *LiveMemories* – Italian (Rodríguez et al., 2010)
      - identity coreference only
  - **Universal Anaphora** (from 2020)
    - initiative led by Massimo Poesio involving many members of the community including ÚFAL
    - CorefUD 0.1 is our contribution to the discussions
Previous common formats

- **CoNLL / CoNLL 2012 / SemEval 2010** (Pradhan et al., 2012, 2011, Recasens et al., 2010)
  - column-based
  - identity coreference only
  - coreference in the last column in open-close notation
  - CoNLL 2011 and 2012 Shared tasks set the standard for its representation and evaluation

- **MMAX / MMAX2** (Müller and Strube, 2001, 2006)
  - XML-based
  - broad variety of linguistic phenomena, including anaphora
  - ARRAU, Polish Coreference Corpus, COREA, Potsdam Commentary Corpus, ParCorFull
  - numerous variations of the format: different number of XML files, different way of capturing sentence boundaries, diverse set of mention attributes, different ways of how mentions are grouped to clusters etc.

- **Prague Markup Language** (Pajas and Štěpánek, 2006)
  - XML-based
  - broad variety of linguistic phenomena, including anaphora
  - PDT, PCEDT
  - used rarely outside ÚFAL
### Example of CoNLL 2012 format

```plaintext
#begin document (bc/cctv/00/cctv_0005; part 003
bc/cctv/00/cctv_0005 3 0 Yes UH (TOP(S(INJ)=s)) -1 -1 Wang_shilin = (ARGM-DIS=) * -
b/cctv/00/cctv_0005 3 1 , * * * * - Wang_shilin = * * -
b/cctv/00/cctv_0005 3 2 ] PRP (NP=) -1 -1 Wang_shilin = (ARGO=) ? (12)
b/cctv/00/cctv_0005 3 3 noticed VBD (VP=) notice 01 1 Wang_shilin = (V=) * -
b/cctv/00/cctv_0005 3 4 that IN (SBAR=) * Wang_shilin = (ARG1=) * -
b/cctv/00/cctv_0005 3 5 many JJ (S(NP(NP=))) * Wang_shilin = * (ARGO=)
b/cctv/00/cctv_0005 3 6 friends NNS * Wang_shilin = * *
b/cctv/00/cctv_0005 3 7 , * * * * - Wang_shilin = * *
b/cctv/00/cctv_0005 3 8 around IN (PP=) -1 -1 Wang_shilin = * * -
b/cctv/00/cctv_0005 3 9 me PRP (NP=) -1 -1 Wang_shilin = * *
bc/cctv/00/cctv_0005 3 10 received VBD (VP=) receive 01 1 Wang_shilin = * (V=) -
b/cctv/00/cctv_0005 3 11 it PRP (NP=))) * Wang_shilin = * (ARG1#) (119)
b/cctv/00/cctv_0005 3 12 . * Wang_shilin = * *

#end document
```

Source: Thomas Wolf: How to train a neural coreference model— Neuralcoref 2
Our harmonization scheme
Basic motivation

• Elementary observations:
  • there are already quite a few coreference datasets around
  • but different annotation schemes applied in different coreference resources
  • virtually impossible to perform multilingual experiments in a wider scale

• A better world must exist!
Sources of inspiration

• the success story of **Universal Dependencies**

• our experience with coreference annotation in the **Prague Dependency Treebank**, in which coreference is integrated with (deep) syntax

• initial spin: recent discussions within the **Universal Anaphora** initiative (Massimo Poesio and others)
Our reasons for convergence towards UD

Why to make a harmonized coreference scheme UD-centric?

• Not only **pragmatic reasons:**
  • UD is a very **popular brand** nowadays, **snowballing** effect, across some 100 languages,
  • numerous technical issues (e.g. tokenization) already somehow **standardized** in UD,
  • existing tools,

• but also **theoretical reasons:**
  • **mentions** often correspond to **syntactically meaningful units** (noun phrase, subject, …)
  • some coreference relations **manifested** primarily by **syntactic means** (reflexive and relative constructions, apposition, predication with copula …)
  • **zero** expressions (such as pro-drop) needed for coreference, syntax useful for their identification
  • reuse of annotation of **coordination** structures
  • verbs of control
Lesson taken from UD history

- UD’s evolution can be traced back to CoNLL shared task in 2006, and several diverse ‘species’ emerged later (other CoNLLs, Universal Dependency Treebank, HamleDT, …)
- XML was everywhere around at that time, JSON became popular later…
- But, surprisingly, a restricted plain-text format became the winner.
- It seems simplicity is more important than flexibility for this kind of cooperation.
- The lesson taken:
  - File format matters!
  - Even if elaboration of shared guidelines will take ages,
  - it’s crucial to have a simple file format from the beginning.
Our file format decisions

• really strict compliance with the CoNLL-U specification,
• checked mechanically by the CoNLL-U validator
• information about mentions and coreference relations stored in the MISC column
  • other options existed (based on comment lines, or employ enhanced deps, or CoNLL-U Plus)
• all information stored as attribute=value pairs
• all information about a mention stored on the syntactic head’s line
  • this is the main connecting point between coreference and dependency syntax!
• cluster-based representation of coreference groupings
  • file-wide unique identifiers of clusters
Other technical decisions

- UD-style morphological and dependency annotation added
  - even though only automatic in most cases (UDPipe used)
- fully automatized pipelines
  - no added manual annotations
- different tools used for importing the data from the source formats
  - Treex (Perl) for Praguian treebanks
  - ElementTree (Python) for MMAX-based resources
  - OntoNotes API (Java) for Ontonotes
  - Udapi (Python) for already conllu-ized data (GUM)
- Udapi also used in some converters for exporting the data into the CoNLL-U format
Attributes added into MISC column

- **required** for every mention head
  - MentionSpan
  - ClusterId

- **optional** (but allowed only with mention heads)
  - ClusterType
  - SplitAnte
  - Bridging
  - EmptyType
  - MentionMisc
File format example 1: a discontinuous mention (dotted gap corresponding to a rhetorical pause, Polish)

```plaintext
# sent_id = 10060
# text = Konkurencja ze strony . . . ministerstwa
1 Konkurencja konkurencja NOUN ... ClusterId=c32584|...|MentionSpan=1-3,7
2 ze z ADP ...
3 strony strona NOUN ...
4 . . PUNCT ...
5 . . PUNCT ...
6 . . PUNCT ...
7 ministerstwa ministerstwać NOUN ... ClusterId=c32585|MentionSpan=7
```
Wenn sich Günter Grass, Christa Wolf oder Stefan Heym in politischen Angelegenheiten zu Wort melden,
Technici totiž zvládli výměnu zařízení ordinace za víkend.

Translation: However, technicians managed the device replacement ... during the weekend. On Saturday ...
Ezt a lapot mára kellett behozni és rajtam kívül mindenkinél itt volt.

Google-translated: This sheet had to be brought in today and was here for everyone except me.
File format example 5: pieces of non-harmonized information (GUM wikification in MentionMisc)

# sent_id = GUM_academic_art-3
# text = Claire Bailey-Ross claire.bailey-ross@port.ac.uk University of Portsmouth, United Kingdom
# s_type = frag
1 Claire Claire PROPN ...
2 Bailey-Ross Bailey-Ross ...
3 claire.bailey-ross@port.ac.uk claire.bailey-ross@port.ac.uk PROPN
4 University University PROPN
   ClusterId=c7|ClusterType=organization|
   MentionMisc=Wikification:University_of_Portsmouth|MentionSpan=4-9
5 of of ADP
6 Portsmouth Portsmouth PROPN
   ClusterId=c8|ClusterType=place|
   MentionMisc=Wikification:Portsmouth|MentionSpan=6-9
7 , , PUNCT
8 United United PROPN
9 Kingdom Kingdom PROPN ...
   ClusterId=c9|ClusterType=place|
   MentionMisc=Wikification:United_Kingdom|MentionSpan=8-9
### Additional annotations stored in the data

<table>
<thead>
<tr>
<th>CorefUD dataset</th>
<th>sentence segmentation</th>
<th>tokenization</th>
<th>POS tags</th>
<th>lemmas</th>
<th>syntactic trees</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>orig.</td>
<td>new</td>
<td>orig.</td>
<td>new</td>
<td>orig.</td>
</tr>
<tr>
<td>Catalan-AnCora</td>
<td>✓</td>
<td>UD2.7</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
</tr>
<tr>
<td>Czech-PCEDT</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>convert</td>
<td>✓</td>
</tr>
<tr>
<td>Czech-PDT</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>convert</td>
<td>✓</td>
</tr>
<tr>
<td>English-GUM</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
</tr>
<tr>
<td>English-ParCorFull</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>kept</td>
<td>×</td>
</tr>
<tr>
<td>French-Democrat</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
</tr>
<tr>
<td>German-ParCorFull</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>kept</td>
<td>×</td>
</tr>
<tr>
<td>German-PotsdamCC</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>kept</td>
<td>×</td>
</tr>
<tr>
<td>Hungarian-SzegedKoref</td>
<td>×</td>
<td>rules</td>
<td>✓</td>
<td>kept</td>
<td>×</td>
</tr>
<tr>
<td>Lithuanian-LCC</td>
<td>×</td>
<td>rules</td>
<td>×</td>
<td>rules</td>
<td>×</td>
</tr>
<tr>
<td>Polish-PCC</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
</tr>
<tr>
<td>Russian-RuCor</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
</tr>
<tr>
<td>Spanish-AnCora</td>
<td>✓</td>
<td>UD2.7</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
</tr>
<tr>
<td>Dutch-COREA</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>kept</td>
<td>×</td>
</tr>
<tr>
<td>English-ARRAU</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
</tr>
<tr>
<td>English-OntoNotes</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
</tr>
<tr>
<td>English-PCEDT</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
<td>kept</td>
<td>✓</td>
</tr>
</tbody>
</table>
Collection CorefUD 0.1
Publication of the resulting data

• all datasets harmonized by March 2021 are gathered in a collection called CorefUD 0.1
• due to individual licence limitations, only some datasets can be distributed publicly
• CorefUD 0.1 divided into two parts
  • **public edition**
    • 13 datasets for 10 languages
    • published via LINDAT/CLARIAH-CZ repository
    • distributed with the original licenses
  • **non-public add-on** (UFAL-internal)
    • 4 datasets for 2 languages
• all datasets divided into train/dev/test sections:
  • 8:1:1 (or preserving the original division, if present)
  • test sections not published because of future shared tasks
Two parts of CorefUD 0.1

Public edition on Lindat:
- Catalan-AnCora
- Czech-PCEDT
- Czech-PDT
- English-GUM
- English-ParCorFull
- French-Democrat
- German-ParCorFull

Non-public add-on:
- Dutch-COREA
- English-ARRAU
- English-OntoNotes
- English-PCEDT

Coref in nutshell
Resource variability
Our scheme
CorefUD 0.1
API
Case study 1
Case study 2
Conclusions
Example of extracted statistics: non-singleton mentions

<table>
<thead>
<tr>
<th>CorefUD dataset</th>
<th>total</th>
<th>per 1k</th>
<th>length</th>
<th>mentions distribution of lengths</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>count</td>
<td>words</td>
<td>max</td>
<td>avg.</td>
</tr>
<tr>
<td>Catalan-AnCora</td>
<td>62,417</td>
<td>128</td>
<td>134</td>
<td>4.2</td>
</tr>
<tr>
<td>Czech-PCEDT</td>
<td>178,475</td>
<td>154</td>
<td>79</td>
<td>3.4</td>
</tr>
<tr>
<td>Czech-PDT</td>
<td>169,644</td>
<td>203</td>
<td>99</td>
<td>2.9</td>
</tr>
<tr>
<td>English-GUM</td>
<td>22,896</td>
<td>170</td>
<td>95</td>
<td>2.6</td>
</tr>
<tr>
<td>English-ParCorFull</td>
<td>720</td>
<td>67</td>
<td>37</td>
<td>2.1</td>
</tr>
<tr>
<td>French-Democrat</td>
<td>47,172</td>
<td>166</td>
<td>71</td>
<td>1.7</td>
</tr>
<tr>
<td>German-ParCorFull</td>
<td>900</td>
<td>85</td>
<td>30</td>
<td>2.0</td>
</tr>
<tr>
<td>German-PotsdamCC</td>
<td>2,523</td>
<td>76</td>
<td>34</td>
<td>2.6</td>
</tr>
<tr>
<td>Hungarian-SzegedKoref</td>
<td>15,182</td>
<td>122</td>
<td>36</td>
<td>1.6</td>
</tr>
<tr>
<td>Lithuanian-LCC</td>
<td>4,337</td>
<td>117</td>
<td>19</td>
<td>1.5</td>
</tr>
<tr>
<td>Polish-PCC</td>
<td>82,865</td>
<td>154</td>
<td>108</td>
<td>2.1</td>
</tr>
<tr>
<td>Russian-RuCor</td>
<td>16,254</td>
<td>104</td>
<td>18</td>
<td>1.7</td>
</tr>
<tr>
<td>Spanish-AnCora</td>
<td>70,675</td>
<td>137</td>
<td>90</td>
<td>4.4</td>
</tr>
<tr>
<td>Dutch-COREA</td>
<td>8,663</td>
<td>62</td>
<td>60</td>
<td>2.6</td>
</tr>
<tr>
<td>English-ARRAU</td>
<td>31,906</td>
<td>139</td>
<td>75</td>
<td>2.9</td>
</tr>
<tr>
<td>English-OntoNotes</td>
<td>209,435</td>
<td>128</td>
<td>94</td>
<td>2.5</td>
</tr>
<tr>
<td>English-PCEDT</td>
<td>183,984</td>
<td>157</td>
<td>88</td>
<td>3.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resource variability</th>
<th>Our scheme</th>
<th>CorefUD 0.1</th>
<th>API</th>
<th>Case study 1</th>
<th>Case study 2</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>35/ 57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Coref in nutshell
More stats...

• If interested in some more statistics, or in the CorefUD format description, or in the survey of the input resources, there’s a detailed technical report (some 70 pages):

https://ufal.mff.cuni.cz/corefud
Application Programming Interface for CorefUD data
API - coreference object model added to Udapi

- toolkit for
  - querying, statistics
  - visualization (text-based, HTML, LaTeX,...)
  - format conversions (e.g. GUM to CorefUD)
  - manipulation (automatic fixes)
  - wrappers for UDPipe (tagging, parsing)

- OO classes for
  - mention (head, words, span, cluster, bridging, misc)
  - coreference cluster (mentions, cluster_type, split_ante)
  - bridging links (source mention, target cluster, relation)

- fast loading (lazy deserialization) of CoNLL-U
  - MISC deserialized from string to dict only when needed
  - coref objects loaded only when needed

- automatic handling of tedious tasks
  - square-brackets co-indexing
  - mention/cluster ordering
>>> import udapi
>>> doc = udapi.Document("en_parcorfull-corefud-dev.conllu")
>>> doc[0].draw(attributes="ord,form,upos,deprel,misc")

# sent_id = 222
# text = Russia's Putin sacks chief of staff Sergei Ivanov

1 Russia PROPN nmod:poss _
2 's PART case _
3 Putin PROPN nsubj ClusterId=c156|MentionMISC=mention:np,nptype:antecedent|MentionSpan=1-3
4 sacks VERB root _
5 chief NOUN obj ClusterId=c157|MentionMISC=mention:np,nptype:antecedent|MentionSpan=5-9
6 of ADP case _
7 staff NOUN nmod _
8 Sergei PROPN flat _
9 Ivanov PROPN flat _
>>> from collections import Counter

>>> for cluster in doc.coref_clusters.values():
...    print(f" {cluster.cluster_id} has {len(cluster.mentions)} mentions:"
...          )
...    counter = Counter()
...    for mention in cluster.mentions:
...        counter[' '.join([w.form for w in mention.words])] += 1
...    for form, count in counter.most_common():
...        print(f"{count:4}: {form}")

c156 has 20 mentions:
  11: Mr Putin
  2: his
  2: he
  1: Russia 's Putin
  1: Russian President Vladimir Putin
  1: Vladimir Putin
  1: him
  1: President Putin

c157 has 19 mentions:
  7: Mr Ivanov
  3: his
  1: chief of staff Sergei Ivanov
...
Case study 1: discontinuous mentions
Linear vs. tree discontinuity of mentions

- linear discontinuity
  - There are one or more tokens (a gap) in the middle that do not belong to the mention.
- non-treelet (dependency-tree discontinuity)
  - A mention does not correspond to a treelet.
  - treelet = connected subgraph of the dependency tree
Linear vs. tree discontinuity of mentions

- **linear discontinuity**
  - There are one or more tokens (a gap) in the middle that do not belong to the mention.

- **non-treelet (dependency-tree discontinuity)**
  - A mention does not correspond to a treelet.
  - treelet = connected subgraph of the dependency tree
  - treelet $\neq$ subtree = a node and all its descendants

Coref in nutshell

Resource variability  Our scheme  CorefUD 0.1  API  **Case study 1**  Case study 2  Conclusions
Linear vs. tree discontinuity of mentions

- linear discontinuity
  - There are one or more tokens (a gap) in the middle that do not belong to the mention.
- non-treelet (dependency-tree discontinuity)
  - A mention does not correspond to a treelet.
  - treelet = connected subgraph of the dependency tree
  - treelet \neq \text{subtree} = a \text{ node and all its descendants}
  - Shall we identify multiple heads too for such mention?
  - May be caused by imperfect automatic parsing.
Causes of linear discontinuity of mentions

- linguistically justifiable discontinuities
  - non-projective constructions (esp. in freer word-order languages)
  - shared modifiers in coordination constructions
  - parenthetical constructions
- spurious
  - various punctuation
  - empty node inserted into unfortunate position
  - mentions that contain multiple sentences
**Brief statistics on discontinuities**

<table>
<thead>
<tr>
<th>CorefUD dataset</th>
<th>disc. mentions [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>German-PotsdamCC</td>
<td>6.3</td>
</tr>
<tr>
<td>Czech-PCEDT</td>
<td>4.1</td>
</tr>
<tr>
<td>Czech-PDT</td>
<td>3.1</td>
</tr>
<tr>
<td>English-PCEDT</td>
<td>2.8</td>
</tr>
<tr>
<td>English-ARRAU</td>
<td>1.2</td>
</tr>
<tr>
<td>Polish-PCC</td>
<td>1.0</td>
</tr>
<tr>
<td>English-ParCorFull</td>
<td>0.7</td>
</tr>
<tr>
<td>Russian-RuCor</td>
<td>0.5</td>
</tr>
<tr>
<td>Hungarian-SzegedKoref</td>
<td>0.4</td>
</tr>
<tr>
<td>German-ParCorFull</td>
<td>0.3</td>
</tr>
<tr>
<td>Dutch-COREA</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Statistics on discontinuous/non-treelet mentions

<table>
<thead>
<tr>
<th>CorefUD dataset</th>
<th>continuous [%] tree</th>
<th>continuous [%] non-tree</th>
<th>discontinuous [%] tree</th>
<th>discontinuous [%] non-tree</th>
<th>discontinuity cause [%] empty</th>
<th>discontinuity cause [%] coord</th>
<th>discontinuity cause [%] other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalan-AnCora</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Czech-PCEDT</td>
<td>89.1</td>
<td>6.8</td>
<td>1.2</td>
<td>2.9</td>
<td>0.1</td>
<td>1.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Czech-PDT</td>
<td>96.1</td>
<td>0.7</td>
<td>1.1</td>
<td>2.0</td>
<td>0.2</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>English-GUM</td>
<td>98.5</td>
<td>1.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>English-ParCorFull</td>
<td>97.0</td>
<td>2.3</td>
<td>0.4</td>
<td>0.3</td>
<td>0.0</td>
<td>0.7</td>
<td>0.0</td>
</tr>
<tr>
<td>French-Democrat</td>
<td>98.0</td>
<td>2.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>German-ParCorFull</td>
<td>97.9</td>
<td>1.7</td>
<td>0.2</td>
<td>0.1</td>
<td>0.0</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>German-PotsdamCC</td>
<td>90.3</td>
<td>3.4</td>
<td>4.3</td>
<td>2.0</td>
<td>0.0</td>
<td>2.5</td>
<td>3.8</td>
</tr>
<tr>
<td>Hungarian-SzegedKoref</td>
<td>96.4</td>
<td>3.2</td>
<td>0.3</td>
<td>0.0</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Lithuanian-LCC</td>
<td>95.3</td>
<td>4.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Polish-PCC</td>
<td>86.3</td>
<td>12.7</td>
<td>0.2</td>
<td>0.8</td>
<td>0.0</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Russian-RuCor</td>
<td>95.5</td>
<td>4.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Spanish-AnCora</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Dutch-COREA</td>
<td>94.1</td>
<td>5.7</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>English-ARRAU</td>
<td>86.5</td>
<td>12.3</td>
<td>0.4</td>
<td>0.8</td>
<td>0.0</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>English-OntoNotes</td>
<td>94.0</td>
<td>6.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>English-PCEDT</td>
<td>96.0</td>
<td>1.2</td>
<td>1.1</td>
<td>1.7</td>
<td>0.1</td>
<td>1.6</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Coref in nutshell
Resource variability
Our scheme
CorefUD 0.1
API
Case study 1
Case study 2
Conclusions
• ~$100\%^1$ shared modifier in a coordination

(15) information about stock purchases and sales by corporate insiders.

(16) U.S. analysts and money managers

---

$^1$all the following proportions are estimated on <30 randomly selected examples for each language
(17) ...dass Eltern unter Kindertagesstätten wählen können, die unterschiedliche pädagogische Konzepte bieten.

‘...that parents from daycare-centers choose can, that different educational concepts offer.’

(18) der Kampf gegen den Top-Terroristen und seine Helfer

‘the fight against the top terrorist and his helpers’
~50% shared modifier in a coordination

(19) **ostój kolorowych kwiatów i motyli**, **niekiedy bardzo rzadkich gatunków**
mainstay colorful flowers and butterflies, sometimes very rare species

‘a mainstay of **colorful flowers** and butterflies, sometimes **very rare species**’

parenthesis

(20) **...komórek rozrodczych matki lub (rzadziej) ojca**
...of-cells reproductive of-mother or (less-frequently) of-father

‘...of the **mother’s** or (less frequently) **father’s** reproductive cells’

other non-projective constructions

(21) **dar to trudny niekiedy do przyjęcia**
gift it difficult sometimes to accepting

*a gift sometimes difficult to accept*
• shared modifier in coordination

(22) vybrat nejlepší lidi, účinně je řídit a dobře zaplatit  
choose best people, effectively them manage and well pay  

‘choose the best people, manage them effectively and pay them well’

• secondary predication

(23) když 0 má s dodavatelem tepla sepsanou smlouvu  
when he has with supplier of heat written contract  

‘when he has a contract with the heat supplier’

• quantified nominal interrupted by a verb

(24) ze 3500 firem jich dnes zůstala jen polovina  
of 3,500 companies of them today remain only half  

‘of the 3,500 companies, only half remain today’
## Head UPOS distribution [%]

<table>
<thead>
<tr>
<th>CorefUD dataset</th>
<th>NOUN</th>
<th>PRON</th>
<th>PROPN</th>
<th>DET</th>
<th>ADJ</th>
<th>VERB</th>
<th>ADV</th>
<th>NUM</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalan-AnCora</td>
<td>51.1</td>
<td>14.7</td>
<td>24.9</td>
<td>2.5</td>
<td>0.5</td>
<td>1.4</td>
<td>0.0</td>
<td>4.9</td>
<td>0.0</td>
</tr>
<tr>
<td>Czech-PCEDT</td>
<td>43.3</td>
<td>27.5</td>
<td>7.0</td>
<td>13.4</td>
<td>1.1</td>
<td>2.9</td>
<td>1.3</td>
<td>0.7</td>
<td>2.9</td>
</tr>
<tr>
<td>Czech-PDT</td>
<td>47.5</td>
<td>20.0</td>
<td>11.7</td>
<td>9.5</td>
<td>6.0</td>
<td>2.1</td>
<td>1.7</td>
<td>0.9</td>
<td>0.6</td>
</tr>
<tr>
<td>English-GUM</td>
<td>53.9</td>
<td>21.8</td>
<td>17.0</td>
<td>0.0</td>
<td>0.8</td>
<td>1.7</td>
<td>0.3</td>
<td>4.0</td>
<td>0.5</td>
</tr>
<tr>
<td>English-ParCorFull</td>
<td>24.1</td>
<td>46.1</td>
<td>24.2</td>
<td>0.7</td>
<td>0.3</td>
<td>2.3</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>French-Democrat</td>
<td>52.9</td>
<td>27.6</td>
<td>8.2</td>
<td>7.2</td>
<td>0.4</td>
<td>1.7</td>
<td>0.8</td>
<td>0.3</td>
<td>0.8</td>
</tr>
<tr>
<td>German-ParCorFull</td>
<td>27.5</td>
<td>47.0</td>
<td>18.8</td>
<td>1.3</td>
<td>0.3</td>
<td>2.6</td>
<td>1.3</td>
<td>0.2</td>
<td>0.9</td>
</tr>
<tr>
<td>German-PotsdamCC</td>
<td>66.7</td>
<td>15.7</td>
<td>10.1</td>
<td>0.6</td>
<td>1.4</td>
<td>0.5</td>
<td>3.3</td>
<td>0.0</td>
<td>1.7</td>
</tr>
<tr>
<td>Hungarian-SzegedKoref</td>
<td>50.6</td>
<td>13.4</td>
<td>6.2</td>
<td>1.7</td>
<td>2.1</td>
<td>3.6</td>
<td>6.9</td>
<td>0.2</td>
<td>15.4</td>
</tr>
<tr>
<td>Lithuanian-LCC</td>
<td>42.5</td>
<td>13.0</td>
<td>22.9</td>
<td>4.9</td>
<td>0.3</td>
<td>2.7</td>
<td>1.1</td>
<td>0.8</td>
<td>12.0</td>
</tr>
<tr>
<td>Polish-PCC</td>
<td>60.4</td>
<td>8.1</td>
<td>9.2</td>
<td>1.9</td>
<td>3.7</td>
<td>11.9</td>
<td>0.9</td>
<td>0.8</td>
<td>3.2</td>
</tr>
<tr>
<td>Russian-RuCor</td>
<td>39.2</td>
<td>26.4</td>
<td>23.4</td>
<td>8.2</td>
<td>0.9</td>
<td>0.7</td>
<td>0.2</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>Spanish-AnCora</td>
<td>51.4</td>
<td>15.7</td>
<td>22.3</td>
<td>3.5</td>
<td>0.9</td>
<td>2.1</td>
<td>0.0</td>
<td>4.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Dutch-COREA</td>
<td>63.1</td>
<td>11.6</td>
<td>11.4</td>
<td>1.4</td>
<td>2.7</td>
<td>5.0</td>
<td>1.6</td>
<td>1.2</td>
<td>1.9</td>
</tr>
<tr>
<td>English-ARRAU</td>
<td>55.8</td>
<td>10.7</td>
<td>18.6</td>
<td>0.7</td>
<td>2.7</td>
<td>3.8</td>
<td>0.7</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>English-OntoNotes</td>
<td>27.6</td>
<td>41.6</td>
<td>24.9</td>
<td>0.6</td>
<td>0.7</td>
<td>2.5</td>
<td>0.3</td>
<td>1.0</td>
<td>0.9</td>
</tr>
<tr>
<td>English-PCEDT</td>
<td>31.4</td>
<td>30.7</td>
<td>22.7</td>
<td>9.4</td>
<td>0.6</td>
<td>2.3</td>
<td>0.6</td>
<td>1.2</td>
<td>1.1</td>
</tr>
</tbody>
</table>
Case study 2: inducing linear mentions from trees
- Prague Dependency Treebank
  - Tectogrammatical layer (t-trees): coreference annotated here
  - Analytical layer (a-trees): so far the only source for Czech Universal Dependencies

- Assumption: **Mention** corresponds to **complete** tectogrammatical **subtree** of a node
  - This does not necessarily hold in the corresponding UD tree!
Conversion of Czech-PDT

- Prague Dependency Treebank
  - Tectogrammatical layer (t-trees): coreference annotated here
  - Analytical layer (a-trees): so far the only source for Czech Universal Dependencies

- Assumption: **Mention** corresponds to **complete** tectogrammatical **subtree** of a node
  - This does not necessarily hold in the corresponding UD tree!

- Universal Dependencies
  - Basic tree
  - Enhanced graph
    - Empty nodes
    - Reentrancies
    - Even cycles!
    - What would “subtree of a node” mean?
Function Words Are Not Nodes in T-trees

- T-tree:

```
ACT          PAT          RSTR
Hlída vyrušila muže který vloupal restaurace
Patrol disturbed man who broke restaurant
```

Coref in nutshell

Resource variability Our scheme CorefUD 0.1 API Case study 1 Case study 2 Conclusions
Function Words Are Not Nodes in T-trees

- T-tree:

```
ACT      PAT      RSTR
Hlídka vyrušila muže který vloupal restaurace
Patrol disturbed man who broke restaurant
```

Coref in nutshell

Resource variability

Our scheme

CorefUD 0.1

API

Case study 1

Case study 2

Conclusions
Function Words Are Not Nodes in T-trees

- T-tree:

```
Hlídka vyrušila muže, který vloupal restaurace
Patrol disturbed man, who broke restaurant
```

- UD basic tree:

```
Hlídka vyrušila muže, který se vloupal do restaurace
Patrol disturbed man, who broke into restaurant
```
Function Words Are Not Nodes in T-trees

• T-tree:

```
ACT   PAT
Hlída vyrůšila muže, který se vloupal do restaurace
Patrol disturbed man, who broke into restaurant
```

• UD basic tree:

```
nsubj  obj
Hlída vyrůšila muže, který se vloupal do restaurace
Patrol disturbed man, who broke into restaurant
```
Function Words Are Not Nodes in T-trees

- T-tree:

```
ACT  PAT  ACT  DIR3
Hlídka vyrušila muže který se vloupal do restaurace
Patrol disturbed man who broke into restaurant
```

- UD basic tree:

```
Hlídka vyrušila muže , který se vloupal do restaurace
Patrol disturbed man , who broke into restaurant
```

Coref in nutshell  Resource variability  Our scheme  CorefUD 0.1  API  Case study 1  Case study 2  Conclusions
Enhanced Graph Is Not a Tree

- T-tree:

  - Hlída vyrušila muže, který se vloupal do restaurace
  - Patrol disturbed man, who broke into restaurant

- UD enhanced graph:

  - Hlída vyrušila muže, který se vloupal do restaurace
  - Patrol disturbed man, who broke into restaurant
With policemen he there met second time.

Case study 1
Case study 2
Conclusions
zloděj S policisty tam se setkal podruhé

thief With policemen he there met second.time
Control Verb Constructions

mu Podařilo se #Cor zmizet
him Succeeded he vanish

Podařilo se mu zmizet
Succeeded him vanish

Coref in nutshell
Resource variability Our scheme CorefUD 0.1 API Case study 1 Case study 2 Conclusions
Coordination

#Gen při dělení kvót na republiku Českou a na Slovensko by nemělo dojít k potížím by

CASE study 1

53/ 57
Coordination

#Gen při dělení kvót na republiku Českou a na Slovensko by nemělo dojít k potížím by .someone during division of quotas between republic Czech and between Slovakia should not come to troubles
Coordination

by someone during division of quotas between republic Czech and between Slovakia should not come to troubles
Coordination

#Gen při dělení kvót na republiku Českou a na Slovensko by nemělo dojít k potížím by someone during division of quotas between republic Czech and between Slovakia should not come to troubles
Spurious Discontinuity

Podniky slovenské \#Oblfm vyvážely totéž co \#Oblfm vyvážely podniky české companies Slovak somewhere export same.thing what somewhere export companies Czech

Abý by slovenské podniky vyvážely totéž, co české companies export same.thing, what Czech
Conclusions
Our contributions

We have

• analyzed variability of coreference annotations in wide range of resources,
• designed a common scheme, built on top of the UD standards,
• converted the 17 resources into this scheme,
• released a subset of the collection publicly.
Future plans

• we can eventually start multi-lingual coreference experiments
• YOU can eventually start multi-lingual coreference experiments
• we can fix some imperfections in the harmonization
• we can extend the harmonization further
  • by harmonizing annotation of more phenomena (such as mention type)
  • by adding more datasets for more languages
• we hope for future convergence with the Universal Anaphora effort
Thank you

If interested in CorefUD, have a look at

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

where you will find

• a link to the CorefUD 0.1 data on Lindat/CLARIAH-CZ
• a short description of the file format (5 pages)
• a comprehensive technical report (some 60 pages)
• this presentation

We would like to thank all our colleagues from various annotation projects who were so kind to give us access to their datasets, comments and advise on the data and annotation structure. We especially thank Ekaterina Lapshinova-Koltunski, Maciej Ogrodniczuk, Massimo Poesio, Sameer Pradhan, Veronika Vincze, Amir Zeldes, Svetlana Toldova, Olga Uryupina, Carole Tiberius, Iris Hendrickx, and Bob Boelhouwer.