CorefUD 1.0 Coreference Meets Universal Dependencies

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

June 20-25, 2022

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Outline

Goal – Examples – Motivation

Goal

Present CorefUD 1.0, a collection of coreference datasets

- harmonized
- consistent
- multilingual

Examples of coreference

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

ANTECEDENT ANAPHOR

Other examples

(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)

Basic motivation

There are already **quite a few coreference datasets** around but annotation schemes and covered phenomena diverge broadly, even for English

- testing the methods on different languages
  - different pronoun dropping
  - definiteness of noun phrases is expressed in different ways
- attract more attention to computational modelling
- theoretical cross-lingual comparative studies

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

Background

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

Previous common formats

- CoNLL / CoNLL 2012 / SemEval 2010 (Pradhan et al., 2012, 2011, Recasens et al., 2010)
  - column-based
  - the standard for representation and evaluation of coreference
- 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
- Prague Markup Language (Pajas and Štěpánek, 2006)
- tabular format of the WebAnno tool

Variability of existing coreference data resources

Selection criteria

- We are aware of some 50 data resources in total
- Datasets are very diverse from many perspectives (domain, types of annotated relations, what is considered to be a mention, etc.)
- Clearly beyond our capacity → sampling was inescapable
- A mixture of selection criteria:
  - **data availability** (the easier access, the better)
  - **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

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)
- 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)

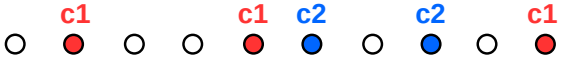
non-free licenses

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

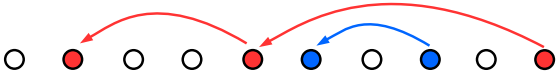
Diversity in existing resources: representation of coreference

two frequent solutions:

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



- **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)



Diversity in existing resources: relations

	Coref. grouping		Relations among mentions					
CorefUD dataset	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	×	✓	(✓)	(✓)	(✓)	✓	✓	×

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

Diversity in existing resources: mentions

	Mention representation		Reconstructed zeros	
original corpus	linear span	syn/sem. head	zero subj.	nom. ellips.
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	×	✓	✓	✓

Collection CorefUD 1.0

Publication of the resulting data

- due to individual licence limitations, only some datasets can be distributed publicly
- CorefUD 1.0 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 1.0

public edition

- Czech-PDT
- Czech-PCEDT
- English-GUM
- German-PotsdamCC
- French-Democrat
- English-ParCorFull
- German-ParCorFull
- Spanish-AnCora
- Catalan-AnCora
- Polish-PCC
- Hungarian-SzegedKoref
- Lithuanian-LCC
- Russian-RuCor

non-public add-on

- English-OntoNotes
- English-ARRAU
- Dutch-COREA
- English-PCEDT

Our file format decisions

- 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 enhanced deps, or CoNLL-U Plus)
- MISC's **attribute Entity** that identifies all mentions that begin or end at the current word
- round bracket notation (opening and ending brackets) used in this attribute
  - trivially supports nested spans and spans that cross sentence boundaries
  - discontinuous spans supported too
  - familiar to the coreference community
- **cluster-based representation** of coreference groupings
  - file-wide unique identifiers of clusters

File Format

```
# 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)
```

Example of extracted statistics: non-singleton mentions

CorefUD dataset	mentions				distribution of lengths					
	total	per 1k	length		0	1	2	3	4	5+
	count	words	max	avg.	[%]	[%]	[%]	[%]	[%]	[%]
Catalan-AnCora	62,417	128	134	4.2	10.2	34.6	19.6	7.5	4.5	23.7
Czech-PCEDT	178,475	154	79	3.4	23.0	28.5	16.1	8.3	4.1	20.0
Czech-PDT	169,644	203	99	2.9	17.2	36.4	18.7	8.5	4.1	15.1
English-GUM	22,896	170	95	2.6	0.0	54.8	20.6	8.4	3.9	12.3
English-ParCorFull	720	67	37	2.1	0.0	59.0	24.4	6.0	2.9	7.6
French-Democrat	47,172	166	71	1.7	0.0	64.2	21.7	6.4	2.5	5.3
German-ParCorFull	900	85	30	2.0	0.0	65.0	17.4	6.2	4.0	7.3
German-PotsdamCC	2,523	76	34	2.6	0.0	34.8	32.4	15.5	6.4	10.9
Hungarian-SzegedKoref	15,182	122	36	1.6	15.1	37.4	32.5	10.2	2.6	2.2
Lithuanian-LCC	4,337	117	19	1.5	0.0	69.1	16.6	11.1	1.2	2.0
Polish-PCC	82,865	154	108	2.1	0.3	68.7	14.9	5.2	2.7	8.2
Russian-RuCor	16,254	104	18	1.7	0.0	68.9	16.3	6.7	3.5	4.6
Spanish-AnCora	70,675	137	90	4.4	11.4	35.3	17.6	7.6	4.0	24.1
Dutch-COREA	8,663	62	60	2.6	0.0	42.5	33.1	8.6	4.0	11.7
English-ARRAU	31,906	139	75	2.9	0.0	45.4	26.9	10.7	4.2	12.8
English-OntoNotes	209,435	128	94	2.5	0.0	56.3	19.8	8.1	4.2	11.7
English-PCEDT	183,984	157	88	3.6	19.3	28.0	17.0	10.6	4.8	20.3

Table 1

Our solutions for...

- zeros
  - use UD mechanism for inserting empty nodes in the enhanced dependency graph to represent reconstructed zeros
- singletons
  - Both singletons and non-singletons are treated as clusters; a singleton cluster contains just a single mention
- bridging
  - in the current version, very broadly; the MISC attribute BRIDGE connects corresponding identity clusters
- split antecedents
  - The MISC attribute SplitAnte points from a cluster to two or more other clusters

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.
- YOU can start multi-lingual coreference experiments

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 1.0 data on Lindat/CLARIAH-CZ
- a link to **CRAC-2022 shared task based on the CorefUD 1.0 dataset**
- description of the file format
- a comprehensive technical report
- all our publications and presentations for CorefUD