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

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

April 9, 2021

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

Outline

Motivation

Basic motivation

Trivial observations

- there are already **quite a few coreference datasets** around
- but different annotation schemes applied in different coreference resources make it **virtually impossible** to run any **multilingual** experiments

Sources of inspiration

- a **snowballing** effect reached in **Universal Dependencies** that use a relatively strictly unified annotation scheme
- experience with coreference annotation in the **Prague Dependency Treebank**, in which coreference is integrated with (deep) syntax
- initial spin: discussions about **Universal Anaphora**

Convergence of coreference and dependency syntax?

Our reasons for convergence towards UD:

- not only **pragmatic**:
  - UD is a very popular brand nowadays,
  - numerous technical issues (e.g. tokenization) already somehow "standardized" in UD
- but also **theoretical**:
  - mentions often correspond to syntactically meaningful units (noun phrase, subject, ...)
  - some coreference relations manifested primarily by syntactic means (refl. and relat. 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

Variability of 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)
  - **documentation** (ideally in English)
  - **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 is readable to us** (because of very low reasons, sorry)

17 coreference datasets included in our harmonization study

free licenses

- 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

non-free licenses

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

Diversity in existing resources: mentions

One of the distinctions – representation of mention spans:

- **linear** – a sequence of tokens specified for a mention
  - the prevailing solution
  - 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)

Diversity in existing resources: mentions, cont.

	Mention repr.		Reconstructed zeros	
original corpus	linear span	syn/sem. head	null 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	×	✓	✓	✓

Table 1

Diversity in existing resources: relations

- one of important distinctions: grouping of mentions that share the same reference (identity coreference)
- two solutions:
  - **cluster-based** grouping
    - coreferential mentions marked (coindexed) by the same cluster identifier
    - slightly prevailing approach
  - **link-based** grouping
    - binary relations between corefering mentions – an anaphor pointing to its antecedent
    - 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, cont.

CorefUD dataset	Relations among mentions							
	cluster-based id.	link-based identity	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	×	✓	(✓)	(✓)	(✓)	✓	✓	×

Table 2

Our harmonizing scheme, file format, and API

Key design decisions

- maximum reuse of UD components
  - using CoNLL-U file format, so that standard UD tools can be used
  - UD-style morphological and dependency annotation added (even though only automatic in most cases, using UDPipe)
  - zeros in the UD style too (empty nodes)
- converters
  - fully automatized pipelines
  - many converters based on Python API for UD (Udapi), with newly added functionality for coreference objects

File format decisions

- really strict compliance with the CoNLL-U specification
- proved by the CoNLL-U validator
- information about mentions and coreference/bridging relations stored exclusively in the MISC column
- all information stored as `attribute=value` pairs
- all information about a mention stored on the syntactic head's line
- if multiple mentions headed in the same node, then distinguished by digits in square brackets after the attribute name
- mention identifiers thus not needed (sufficient to identify a token using UD means – sentence ID and node ordinal number)
- cluster-based representation of coreference groupings (file-wide unique identifiers of clusters)
- mention-to-cluster pointer representation of bridging relations

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)

```
# 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
```

File format example 1: multiple mentions in a node (coordination in German, nested mentions actually)

```
# text = Wenn sich Günter Grass , Christa Wolf oder Stefan Heym in politischen
        Angelegenheiten zu Wort melden ,
1       Wenn     Wenn     SCONJ   KOUS ...
2       sich     sich     PRON    PRF ...
3       Günter   Günter   PROPN   NE  ... ClusterId[1]=c77|ClusterId[2]=c83|...
                                          |MentionSpan[1]=3-10|MentionSpan[2]=3-4
4       Grass    Grass    PROPN   NE ...
5       ,        ,        PUNCT   $, ...
6       Christa  Christa  PROPN   NE ... ClusterId=c84|...|MentionSpan=6-7
7       Wolf     Wolf     PROPN   NE ...
8       oder     oder     CCONJ   KON ...
9       Stefan   Stefan   PROPN   NE ... ClusterId=c85|...|MentionSpan=9-10
10      Heym     Heym     PROPN   NE ...
11      in       in       ADP     APPR ...
12      politischen       politisch       ADJ     ADJA    ...
13      Angelegenheiten   Angelegenheit   NOUN    NN      ...
14      zu       zu       ADP     APPR    ...
15      Wort     Wort     NOUN    NN      ...
16      melden   melden   VERB    VVINF   ...
17      ,        ,        PUNCT   $,  ...
```

File format example 3: bridging (part-of relation in Czech)

```
# sent_id = cmpr9410-015-p8s2
# text = Technici totiž zvládli výměnu zařízení ordinace za víkend.
1   Technici        technik NOUN ...
2   totiž   totiž   CCONJ   ...
3   zvládli zvládnout       VERB ...
4   výměnu  výměna  NOUN    ...
5   zařízení        zařízení        NOUN ...
6   ordinace        ordinace        NOUN    ...
7   za      za      ADP ...
8   víkend  víkend  NOUN...    ClusterId=c423|...|MentionSpan=7-8
9   .       .       PUNCT  ...

# sent_id = cmpr9410-015-p8s3
# text = V sobotu demontovali, v neděli ustavili zařízení nové a proškolili lékaře.
1   V       v       ADP ...
2   sobotu  sobota  NOUN    ... Bridging=c423:Part|ClusterId=c433|MentionSpan=1-2
```

Translation: *However, technicians managed the device replacement ... during the weekend. On Saturday ...*

File format example 4: zero (a pro-drop in Hungarian)

```
# sent_id = 79
# text = Ezt a lapot mára kellett behozni és rajtam kívül mindenkinél itt volt .
1       Ezt     ez      DET     ...
2       a       a       DET     ...
3       lapot   lap     NOUN    ... ClusterId=c40|MentionSpan=2-3
4       mára    mára    ADV     ...
5       kellett kell    VERB    ...
6       behozni behozik VERB    ...
7       és      és      CCONJ   ...
8       rajtam  raj     VERB    ...
9       kívül   kívül   ADP     ...
10      mindenkinél     mindenkinél     SCONJ   ...
11      itt     itt     ADV     ...
12      volt    van     AUX     ...
12.1    _       _       _       ... ClusterId=c40|EmptyType=NullSubj|MentionSpan=12.1
13      .       .       PUNCT   ...
```

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
```

Case study: comparison of two mention span notations

interval notation used in CorefUD:

- Pros
  - all information about a mention on a single line
  - discontinuities easy and intuitive to handle (e.g `MentionSpan=1-3,5`)
  - mentions identifiable without introducing their IDs
- Cons
  - layered attributes (square brackets with co-indexes) needed in the case of multiple mentions sharing the same head node
  - cumbersome in case of mention crossing sentence boundaries (but still possible)

bracketed notation in Amir Zeldes' proposal:

- Pros
  - similar to other formats using BIO encoding (CoNLL2012)
- Cons
  - mention heads not represented
  - cumbersome representation of discontinuous mentions
  - all mention attributes (incl. cluster type) need to be encoded by joining in entity IDs – difficult to extend
  - pushdown automaton needed for parsing

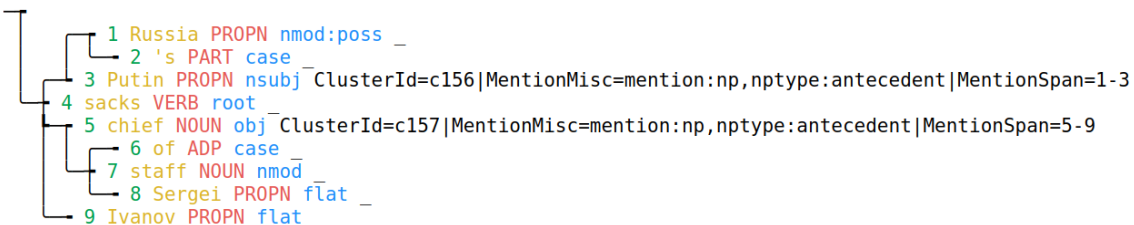
API - coreference object model added to Udapi

- 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

API - example source code

```
>>> 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
```



API - example source code

```
>>> 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
...
```

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 (UFAL-internal) add-on
    - 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
- German-PotsdamCC
- Hungarian-SzegedKoref
- Lithuanian-LCC
- Polish-PCC
- Russian-RuCor
- Spanish-AnCora

Non-public add-on:

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

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

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
- we can extend the harmonization further
  - by harmonizing annotation of more phenomena (such as mention type)
  - by adding more datasets for more languages
- we will be happy to share the know-how with others and to deliver the data for a coreference-related shared task
- convergence with UA would be great!

Acknowledgements

We would like to thank all our colleagues from different 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 Maciej Ogrodniczuk, Massimo Poesio, Sameer Pradhan, Veronika Vincze, Amir Zeldes, Svetlana Toldova, Olga Uryupina, Carole Tiberius, Iris Hendrickx, Bob Boelhouwer and others.

Thank you

If interested in CorefUD, please have a look at https://ufal.mff.cuni.cz/corefud

- 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

Discontinuities in CorefUD 0.1

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 continuous subgraph of the dependency tree.
  - 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

Preliminary statistics on discontinuities

CorefUD dataset	discontinuous mentions [%]
Czech-PCEDT	4.1
Czech-PDT	3.1
English-ParCorFull	0.7
German-ParCorFull	0.3
German-PotsdamCC	6.3
Hungarian-SzegedKoref	0.4
Polish-PCC	1.0
Russian-RuCor	0.5
Dutch-COREA	0.3
English-ARRAU	1.2
English-PCEDT	2.8

English-ARRAU

- $\sim$100%[1] shared modifier in a coordination
  (1) information about **stock purchases** and sales **by corporate insiders**.
  (2) **U.S.** analysts and **money managers**

[1] all the following proportions are estimated on <30 randomly selected examples for each language

German-PotsdamCC

- >60% punctuation not included in a span (already in the source annotation)
- verb or separable prefix in a gap

  (3) *... dass Eltern* ***unter Kindertagesstätten*** *wählen können ,* ***die***
  ... that parents from daycare-centers choose can , that
  unterschiedliche pädagogische Konzepte bieten *.*
  different educational concepts offer .
  '... that parents can choose from **daycare centers that offer different educational concepts**.'

- shared modifier in a coordination

  (4) *der Kampf* ***gegen*** *den Top-Terroristen und* ***seine Helfer***
  the fight against the top-terrorist and his helpers
  'the fight **against** the top terrorist and **his helpers**'

Polish-PCC

- ∼50% shared modifier in a coordination

  (5) *ostoję* ***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

  (6) *... 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

  (7) ***dar*** *to* ***trudny*** ***niekiedy*** ***do*** ***przyjęcia***
  gift it difficult sometimes to accepting
  '**a gift sometimes difficult to accept**'

Czech-PDT

- shared modifier in coordination

  (8) *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

  (9) *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

  (10) *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'

Extra slide: statistics on discontinuities and head POS

CorefUD dataset	mention type [%]			distribution of head UPOS [%]								
	w/empty	w/gap	non-tree	NOUN	PRON	PROPN	DET	ADJ	VERB	ADV	NUM	other
Catalan-AnCora	7.1	0.0	0.0	51.1	14.7	24.9	2.5	0.5	1.4	0.0	4.9	0.0
Czech-PCEDT	30.9	4.1	9.7	43.3	27.5	7.0	13.4	1.1	2.9	1.3	0.7	2.9
Czech-PDT	19.6	3.1	2.8	47.5	20.0	11.7	9.5	6.0	2.1	1.7	0.9	0.6
English-GUM	0.0	0.0	1.5	53.9	21.8	17.0	0.0	0.8	1.7	0.3	4.0	0.5
English-ParCorFull	0.0	0.7	2.6	24.1	46.1	24.2	0.7	0.3	2.3	0.7	0.8	0.8
French-Democrat	0.0	0.0	2.0	52.9	27.6	8.2	7.2	0.4	1.7	0.8	0.3	0.8
German-ParCorFull	0.0	0.3	1.9	27.5	47.0	18.8	1.3	0.3	2.6	1.3	0.2	0.9
German-PotsdamCC	0.0	6.3	5.4	66.7	15.7	10.1	0.6	1.4	0.5	3.3	0.0	1.7
Hungarian-SzegedKoref	15.2	0.4	3.3	50.6	13.4	6.2	1.7	2.1	3.6	6.9	0.2	15.4
Lithuanian-LCC	0.0	0.0	4.7	42.5	13.0	22.9	4.9	0.3	2.7	1.1	0.8	12.0
Polish-PCC	0.5	1.0	13.5	60.4	8.1	9.2	1.9	3.7	11.9	0.9	0.8	3.2
Russian-RuCor	0.0	0.5	4.5	39.2	26.4	23.4	8.2	0.9	0.7	0.2	0.5	0.4
Spanish-AnCora	8.5	0.0	0.0	51.4	15.7	22.3	3.5	0.9	2.1	0.0	4.0	0.0
Dutch-COREA	0.0	0.3	5.9	63.1	11.6	11.4	1.4	2.7	5.0	1.6	1.2	1.9
English-ARRAU	0.0	1.2	13.1	55.8	10.7	18.6	0.7	2.7	3.8	0.7	3.5	3.5
English-OntoNotes	0.0	0.0	6.0	27.6	41.6	24.9	0.6	0.7	2.5	0.3	1.0	0.9
English-PCEDT	29.3	2.8	2.9	31.4	30.7	22.7	9.4	0.6	2.3	0.6	1.2	1.1