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

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

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

- 1985-1986 - Hajičová – Panevová – Sgall, Coreference in the Grammar and in the Text
- 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

Grammatical and Textual coreference

(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

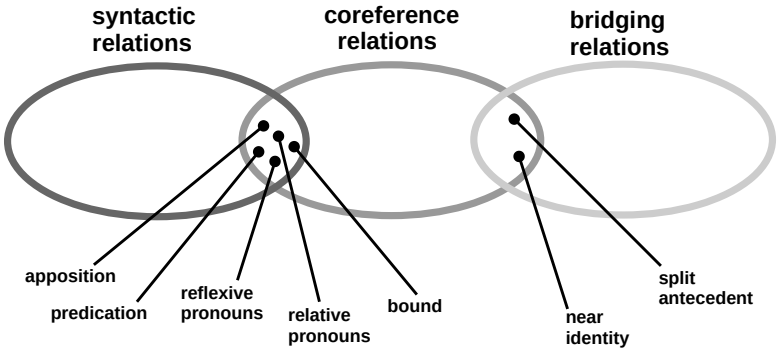


Figure 1: Types of possible relations between referring expressions, including borderline types.

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)
- 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

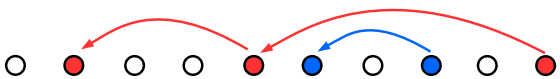
- Which domain (news, dialogues, stories...)?
- Which relations are annotated?
- What is considered to be a mention?
- Which additional linguistic information resources have (lemmatization, POS tagging, sentence segmentation, tokenization, syntactic trees, document boundaries, etc.)?

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: apposition

(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

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	×	✓	(✓)	(✓)	(✓)	✓	✓	×

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	×	✓	✓	✓

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*
Residents	[of the city]	[to themselves]	picked	**their**	way

(Novák and Nedoluzhko, 2015)

Differences in realization of coreference across languages

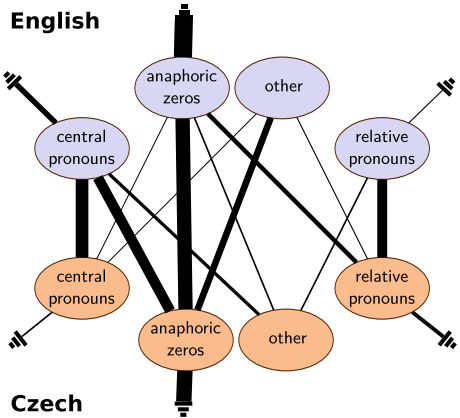


Figure 2: Correspondences between Czech and English potentially coreferential expressions

(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

```
#begin document (bc/cctv/00/cctv_0005); part 003

bc/cctv/00/cctv_0005   3   0        Yes   UH   (TOP(S(INTJ*)        -    -   -   Wang_shilin   *   (ARGM-DIS*)        *      -
bc/cctv/00/cctv_0005   3   1          ,    ,            *)        -    -   -   Wang_shilin   *           *)        *      -
bc/cctv/00/cctv_0005   3   2          I  PRP          (NP*)        -    -   -   Wang_shilin   *        (ARG0*)        *    (12)
bc/cctv/00/cctv_0005   3   3    noticed  VBD          (VP*     notice   01   1   Wang_shilin   *           (V*)        *      -
bc/cctv/00/cctv_0005   3   4       that   IN       (SBAR*        -    -   -   Wang_shilin   *        (ARG1*         *      -
bc/cctv/00/cctv_0005   3   5       many   JJ    (S(NP(NP*        -    -   -   Wang_shilin   *             *    (ARG0*      -
bc/cctv/00/cctv_0005   3   6    friends  NNS          *)        -    -   -   Wang_shilin   *             *         *      -
bc/cctv/00/cctv_0005   3   7          ,    ,            *        -    -   -   Wang_shilin   *             *         *      -
bc/cctv/00/cctv_0005   3   8     around   IN       (PP*        -    -   -   Wang_shilin   *             *         *      -
bc/cctv/00/cctv_0005   3   9         me  PRP       (NP*)))     -    -   -   Wang_shilin   *             *        *)    (12)
bc/cctv/00/cctv_0005   3  10   received  VBD          (VP*     receive   01   1   Wang_shilin   *             *      (V*)      -
bc/cctv/00/cctv_0005   3  11         it  PRP       (NP*)))))   -    -   -   Wang_shilin   *             *       *)   (ARG1*)   (119)
bc/cctv/00/cctv_0005   3  12          .    .            *))     -    -   -   Wang_shilin   *             *         *      -

bc/cctv/00/cctv_0005   3   0         It  PRP   (TOP(S(NP*)        -    -   -   Wang_shilin   *        *        *      -
bc/cctv/00/cctv_0005   3   1      seems  VBZ          (VP*     seem   01   1   Wang_shilin   *      (V*)       *      -
bc/cctv/00/cctv_0005   3   2       that   IN       (SBAR*        -    -   -   Wang_shilin   *   (ARG1*        *      -
bc/cctv/00/cctv_0005   3   3     almost   RB       (S(NP*        -    -   -   Wang_shilin   *        *   (ARG0*      -
bc/cctv/00/cctv_0005   3   4   everyone   NN            *)       -    -   -   Wang_shilin   *        *        *)     -
bc/cctv/00/cctv_0005   3   5   received  VBD          (VP*     receive   01   1   Wang_shilin   *        *      (V*)     -
bc/cctv/00/cctv_0005   3   6       this   DT       (NP*        -    -   -   Wang_shilin   *        *   (ARG1*   (119
bc/cctv/00/cctv_0005   3   7        SMS   NN        *)))))     -    -   -   Wang_shilin   *        *)        *)   119)
bc/cctv/00/cctv_0005   3   8          .    .            *))     -    -   -   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)

```
# 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
```

Additional annotations stored in the data

CorefUD dataset	sentence segmentation		tokenization		POS tags		lemmas		syntactic trees	
	orig.	**new**	**orig.**	**new**	**orig.**	**new**	**orig.**	**new**	**orig.**	**new**
Catalan-AnCora	✓	UD2.7	✓	kept	✓	convert	✓	convert	✓ (phr.)	convert
Czech-PCEDT	✓	kept	✓	convert	✓	convert	✓	convert	(✓) (dep.)	convert
Czech-PDT	✓	kept	✓	convert	✓	convert	✓	kept	✓ (dep.)	convert
English-GUM	✓	kept	✓	kept	✓	kept	✓	kept	✓ (dep.)	kept
English-ParCorFull	✓	kept	✓	kept	×	UDPipe	×	UDPipe	×	UDPipe
French-Democrat	(✓)	kept	(✓)	kept	(✓)	kept	(✓)	kept	(✓) (dep.)	kept
German-ParCorFull	✓	kept	✓	kept	×	UDPipe	×	UDPipe	×	UDPipe
German-PotsdamCC	✓	kept	✓	kept	×	UDPipe	×	UDPipe	×	UDPipe
Hungarian-SzegedKoref	×	rules	✓	kept	×	UDPipe	×	UDPipe	×	UDPipe
Lithuanian-LCC	×	rules	×	rules	×	UDPipe	×	UDPipe	×	UDPipe
Polish-PCC	✓	kept	✓	kept	✓	UDPipe	✓	UDPipe	×	UDPipe
Russian-RuCor	✓	kept	✓	kept	✓	UDPipe	✓	UDPipe	×	UDPipe
Spanish-AnCora	✓	UD2.7	✓	kept	✓	convert	✓	kept	✓ (phr.)	convert
Dutch-COREA	✓	kept	✓	kept	×	UDPipe	×	UDPipe	×	UDPipe
English-ARRAU	✓	kept	✓	kept	✓	UDPipe	✓	UDPipe	✓ (phr.)	UDPipe
English-OntoNotes	✓	kept	✓	kept	✓	UDPipe		UDPipe	✓ (phr.)	UDPipe
English-PCEDT	✓	kept	✓	kept	✓	convert	✓	kept	(✓) (d+p.)	convert d.

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
- 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

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

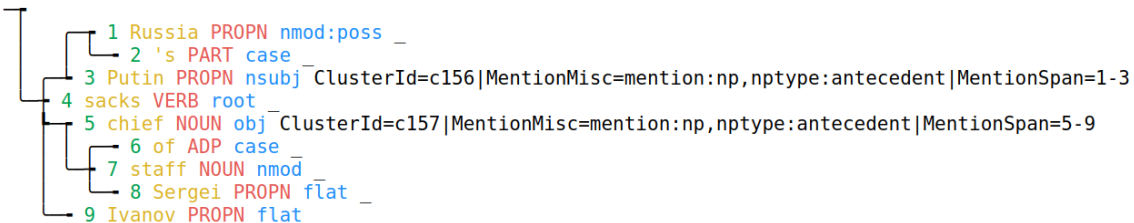
API - coreference object model added to Udapi

- 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

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
...
```

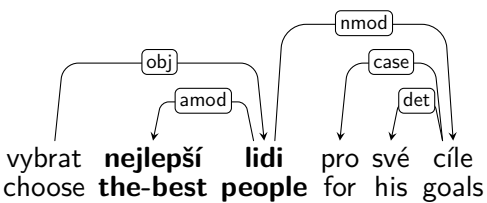
Case study 1: discontinuous mentions

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



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
  - 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

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

Statistics on discontinuous/non-treelet mentions

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

English-ARRAU

- ∼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

German-PotsdamCC

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

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

'...that parents can choose from **daycare centers that offer different educational concepts**.'

- shared modifier in a coordination

(18) *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

(19) *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

(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**

Head UPOS distribution [%]

CorefUD dataset	NOUN	PRON	PROPN	DET	ADJ	VERB	ADV	NUM	other
Catalan-AnCora	51.1	14.7	24.9	2.5	0.5	1.4	0.0	4.9	0.0
Czech-PCEDT	43.3	27.5	7.0	13.4	1.1	2.9	1.3	0.7	2.9
Czech-PDT	47.5	20.0	11.7	9.5	6.0	2.1	1.7	0.9	0.6
English-GUM	53.9	21.8	17.0	0.0	0.8	1.7	0.3	4.0	0.5
English-ParCorFull	24.1	46.1	24.2	0.7	0.3	2.3	0.7	0.8	0.8
French-Democrat	52.9	27.6	8.2	7.2	0.4	1.7	0.8	0.3	0.8
German-ParCorFull	27.5	47.0	18.8	1.3	0.3	2.6	1.3	0.2	0.9
German-PotsdamCC	66.7	15.7	10.1	0.6	1.4	0.5	3.3	0.0	1.7
Hungarian-SzegedKoref	50.6	13.4	6.2	1.7	2.1	3.6	6.9	0.2	15.4
Lithuanian-LCC	42.5	13.0	22.9	4.9	0.3	2.7	1.1	0.8	12.0
Polish-PCC	60.4	8.1	9.2	1.9	3.7	11.9	0.9	0.8	3.2
Russian-RuCor	39.2	26.4	23.4	8.2	0.9	0.7	0.2	0.5	0.4
Spanish-AnCora	51.4	15.7	22.3	3.5	0.9	2.1	0.0	4.0	0.0
Dutch-COREA	63.1	11.6	11.4	1.4	2.7	5.0	1.6	1.2	1.9
English-ARRAU	55.8	10.7	18.6	0.7	2.7	3.8	0.7	3.5	3.5
English-OntoNotes	27.6	41.6	24.9	0.6	0.7	2.5	0.3	1.0	0.9
English-PCEDT	31.4	30.7	22.7	9.4	0.6	2.3	0.6	1.2	1.1

Case study 2: inducing linear mentions from trees

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

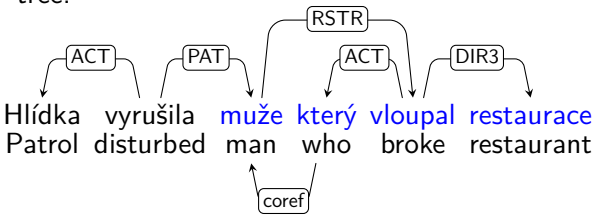
- Assumption: **Mention** corresponds to **complete** tectogrammatical **subtree** of a node
  - This does not necessarily hold in the corresponding UD tree!

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?

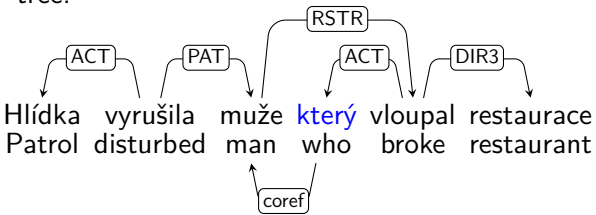
Function Words Are Not Nodes in T-trees

- T-tree:



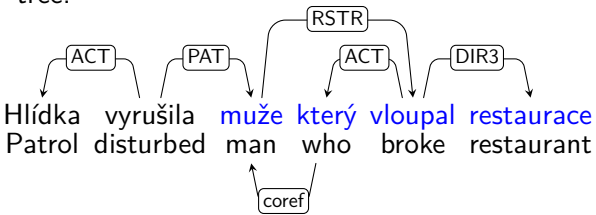
Function Words Are Not Nodes in T-trees

- T-tree:

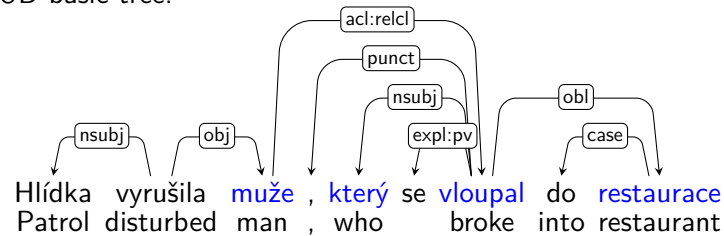


Function Words Are Not Nodes in T-trees

- T-tree:

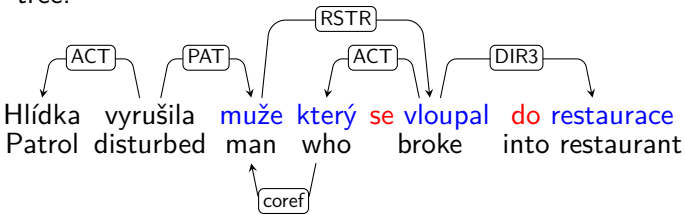


- UD basic tree:

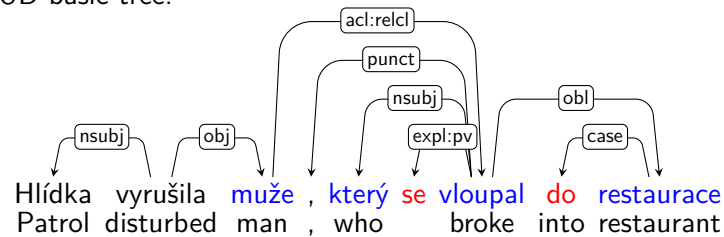


Function Words Are Not Nodes in T-trees

- T-tree:

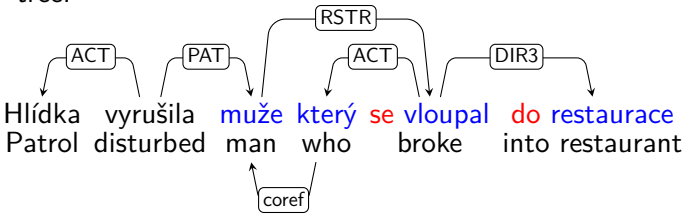


- UD basic tree:

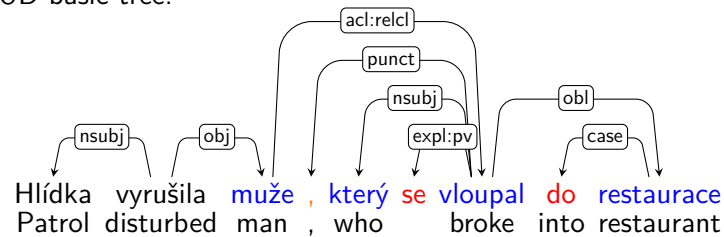


Function Words Are Not Nodes in T-trees

- T-tree:

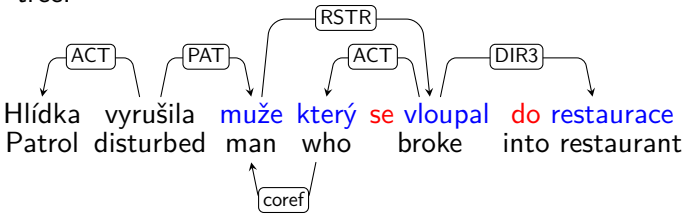


- UD basic tree:

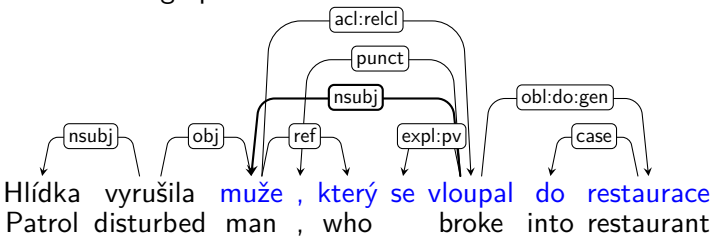


Enhanced Graph Is Not a Tree

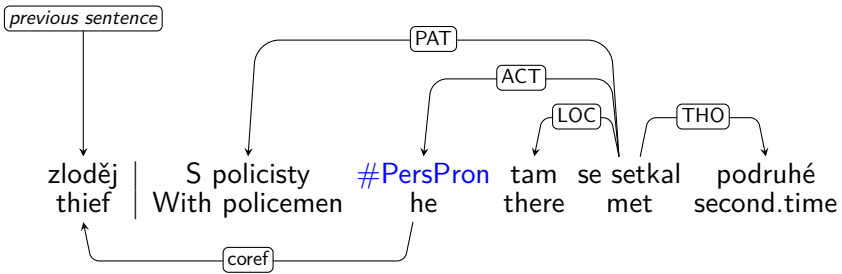
- T-tree:

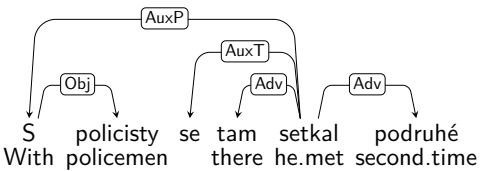


- UD enhanced graph:

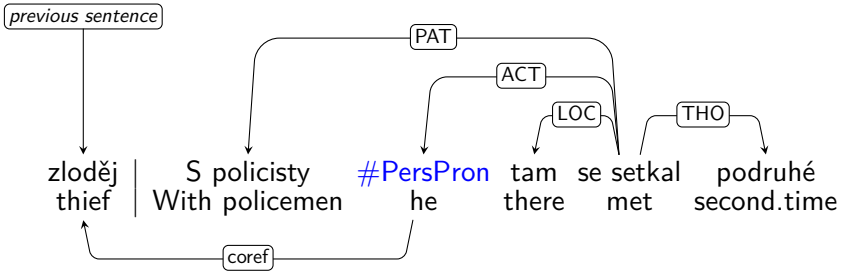


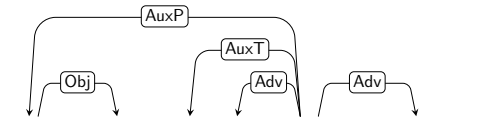
Empty Nodes

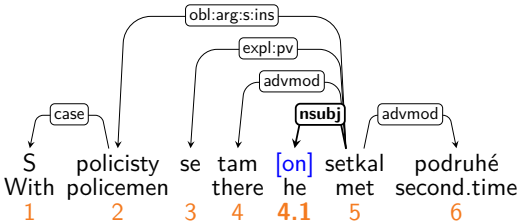

previous sentence
PAT
ACT
LOC
THO
zloděj
thief
S policisty
With policemen
#PersPron
he
tam
there
se
setkal
met
podruhé
second.time
coref


AuxP
AuxT
Obj
Adv
Adv
S
With
policisty
policemen
se
tam
there
setkal
he.met
podruhé
second.time

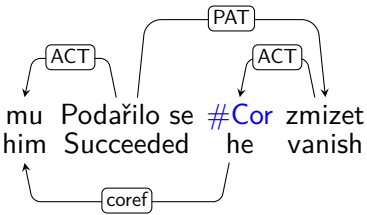
Empty Nodes

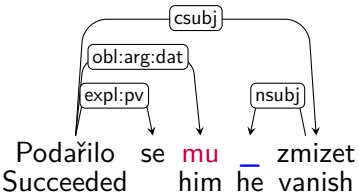

previous sentence
PAT
ACT
LOC
THO
zloděj | S policisty #PersPron tam se setkal podruhé
thief | With policemen he there met second.time
coref


AuxP
AuxT
Obj
Adv
Adv
S policisty se tam setkal podruhé
With policemen there he.met second.time

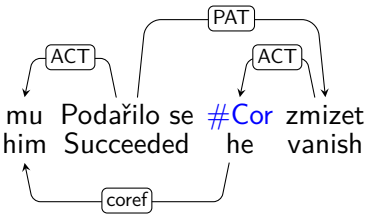

obl:arg:s:ins
expl:pv
advmod
case
nsubj
advmod
S policisty se tam [on] setkal podruhé
With policemen there he met second.time
1 2 3 4 4.1 5 6

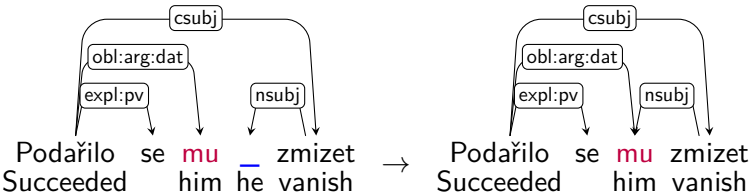
Control Verb Constructions



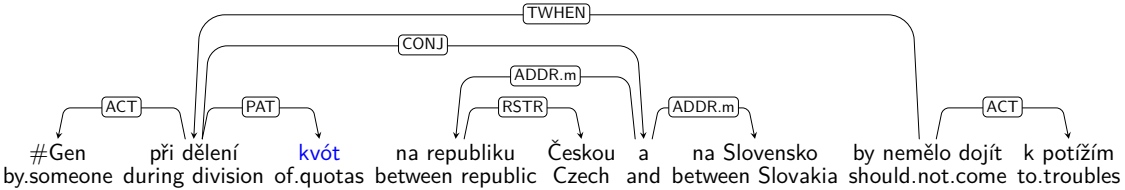


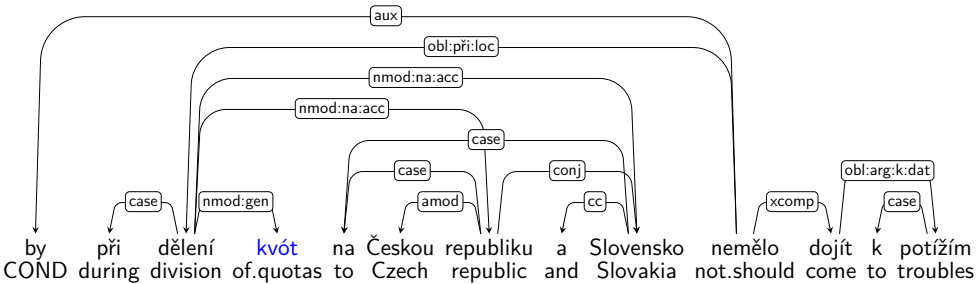
Control Verb Constructions



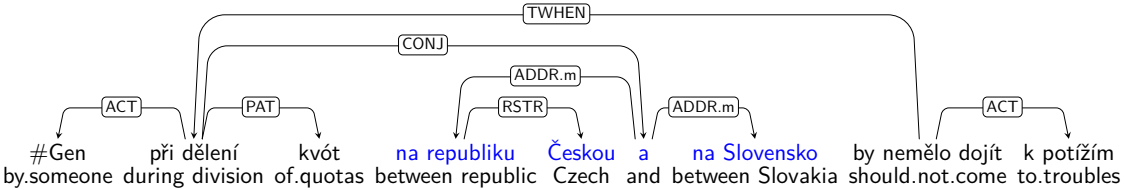


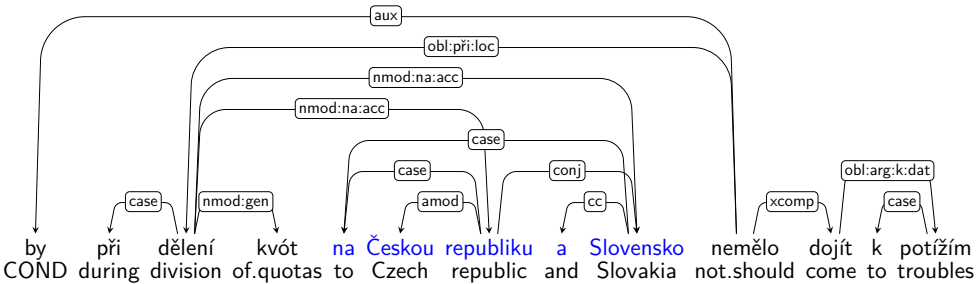
Coordination



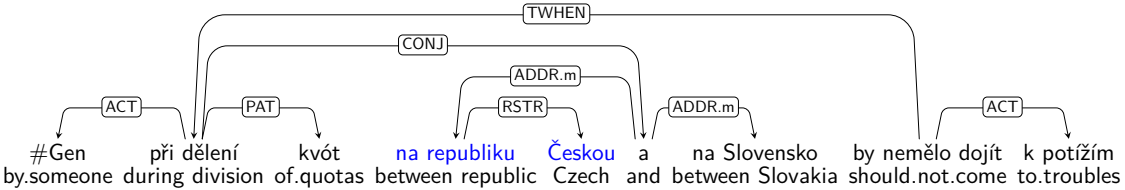


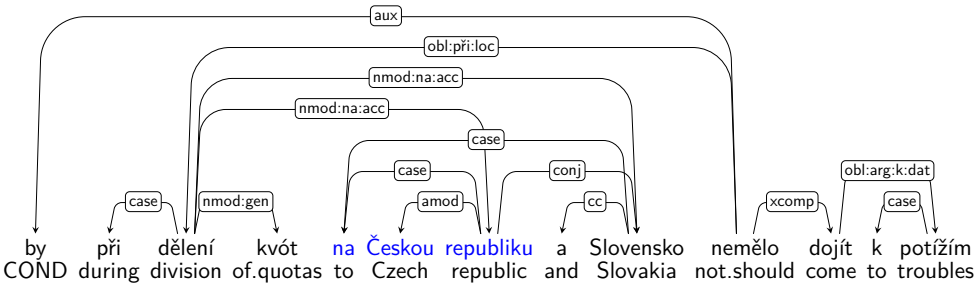
Coordination



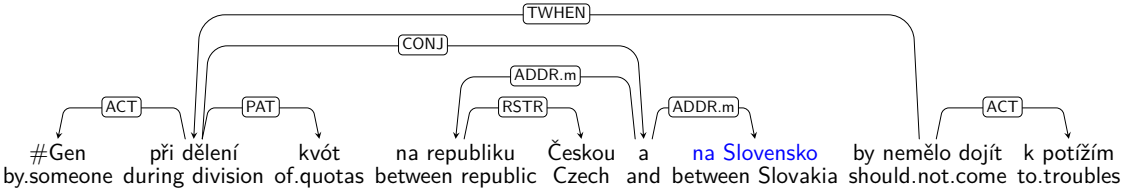


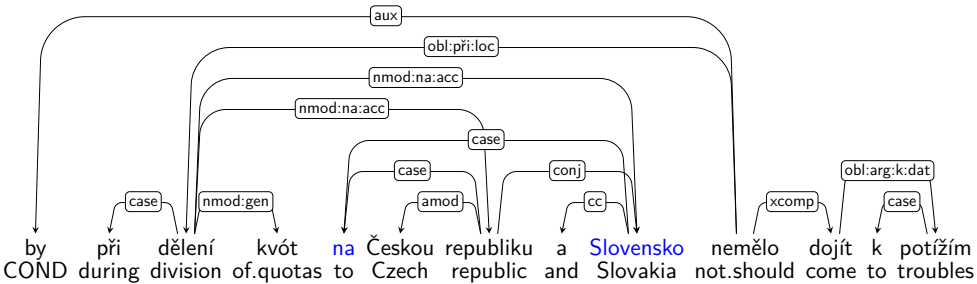
Coordination



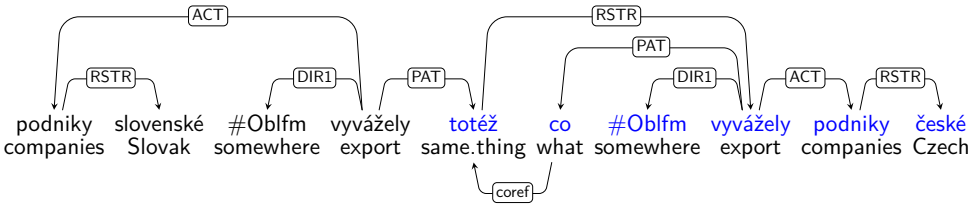


Coordination





Spurious Discontinuity



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.