Massimo Poesio

UNIVERSAL ANAPHORA 1.0: PROGRESS MEETING

https://universalanaphora.github.io/UniversalAnaphora/

Universal Anaphora

- There are by now a number of substantial anaphorically annotated corpora, covering several languages, according to broadly comparable guidelines, but created without real coordination
- The objective of the Universal Anaphora initiative is to attempt to replicate the success of Universal Dependencies for anaphora

Progress so far

- Website: https://universalanaphora.github.io/UniversalAnaphora/
- Mailing lists: universalanaphora@googlegroups.com, universalanaphora-markup@googlegroups.com
- Proposal regarding coverage: https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/Universal_Anaphora_1_0___Proposal_for_Discussion.pdf
  - See also the very extensive CorefUD document, https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/nedoluzhko_et_al_CorefUD20201_tr66.pdf
- Markup proposals & dataset conversions (see website)
  - CONLLU extensions: Zeldes, CorefUD
  - CONLLU Plus extension: CONLL-UA `exploded'
- CODI/CRAC Shared Task on Coreference in Dialogue https://competitions.codalab.org/competitions/30312
- UA Scorer: https://github.com/juntaoy/universal-anaphora-scorer

The plan for today

- 2:00: Intro (this talk)
- 2:20-4:00: Update on progress so far, discussion on markup proposals
- 4:20-5:40: Next steps (Ontonotes and other planned conversions, Website, ANNIS, first discussion of complex issues in markup)
- 5:40: End discussion

Universal Anaphora: proposed coverage

https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/Universal_Anaphora_1_0___Proposal_for_Discussion.pdf

Non Anaphoric Layers from CONLL-U

- Mandatory layers
  - NEWDOC, SENT_ID, ID, FORM (token)
- Optional layers:
  - NEWPAR, TEXT, TEXT_EN
  - Basic UD layers:
    LEMMA, UPOS, XPOS, FEATS (Morphosyn),
    HEAD, DEPREL

Core Anaphoric Layers

- Relations:
  - Focus on identity anaphora
  - Allow for split antecedent plurals
- Markable definition:
  - Adopt the definition of markable from MATE/GNOME/ARRAU/ONTONOTES?
  - Treat singletons as markables?
  - Allow for empty categories (e.g., verb as markable for zeros)
- Mandatory layers:
  - Identity
  - MIN
- Optional layers:
  - Sem_Type (DN/DO/expletive/ quantifier / predicate / idiom ….)
- Issues:
  - Definition of markable
  - Predication

Additional Anaphoric Layers

- Relations:
  - Associative anaphora
  - Discourse deixis
- Mandatory layers:
  - None
- Optional layers:
  - SPLIT
  - BRIDGING
  - DISCOURSE_DEIXIS (Anaphora to non-nominal antecedents)
  - DEIXIS (reference)

Non Anaphoric Layers not in CONLL-U

- Optional layers:
  - NOM_SEM (Entity Type, Genericity ….)
  - Linguistic layers relevant to anaphoric interpretation:
    CONSTITUENCY, WORDSENSE, PROPOSITION, DISCOURSE (RST, PDTB)

Issues

https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/Universal_Anaphora_1_0___Proposal_for_Discussion.pdf

Some issues

- Which layers should be mandatory and which ones shouldn't (e.g., MIN, Sem_Type)
- Notion of `Identity'
  - See Amir's presentation on Predication
- Markable definition:
  - See Prague group presentation on discontinuous markables
- Issues from multilinguality / other genres
  - E.g., coreference in dialogue
- Multiplicity of interpretations
- How to deal with differences in guidelines
  - Generics
  - Bridging
  - General idea: having a unified format would allow users to search the datasets / the scorer to work irrespective

Amir Zeldes

The Prague Group

Using UA for Anaphora Resolution: Exploded CONLL-UA and the UA Scorer

Massimo Poesio & Juntao Yu
(for the scorer also: Nafise Moosavi & Silviu Paun)

Exploded CONLL-UA

- For anaphora resolution/coreference, it is useful to have a format in which the UA layers are represented in separate columns instead of being all packed in the Misc column
  - So that, for example, a system can only use the columns relevant to Bridging/ Discourse Deixis
- This is Exploded CONLL-UA https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/UA_CONLL_U_Plus_proposal_v1.0.md

Exploded CONLL-UA

- An extension of UD's 'CONLL-U Plus' https://universaldependencies.org/ext-format.html
- Five anaphora/reference columns:
  - IDENTITY
  - SPLIT
  - BRIDGING
  - DISCOURSE_DEIXIS
  - DEIXIS
- Plus a NOM_SEM column
  - ENTITY_TYPE, GENERICITY

Identity

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC IDENTITY SPLIT BRIDGING DISCOURSE_DEIXIS REFERENCE NOM_SEM
# newdoc id = GUM_voyage_tulsa
# sent_id = GUM_voyage_tulsa-1
# text = Tulsa
1       Tulsa   _       _       _       _       _       _       _       _       (EntityID=1|MarkableID=markable_1|Min=1)

# sent_id = GUM_voyage_tulsa-2
# text = Tulsa is in the Green Country region of Oklahoma
2       Tulsa   _       _       _       _       _       _       _       _       (EntityID=1|MarkableID=markable_2|Min=2)
3       is      _       _       _       _       _       _       _       _       _
4       in      _       _       _       _       _       _       _       _       _
5       the     _       _       _       _       _       _       _       _       (EntityID=2|MarkableID=markable_3|Min=7
6       Green   _       _       _       _       _       _       _       _       _       _
7       Country _       _       _       _       _       _       _       _       _
8       region  _       _       _       _       _       _       _       _       _
9       of      _       _       _       _       _       _       _       _       _
10      Oklahoma        _       _       _       _       _       _       _       (EntityID=3|MarkableID=markable_4|Min=10))
11      .       _       _       _       _       _       _       _       _

# sent_id = GUM_voyage_tulsa-3
# text = It is also called "T-town"
12      It      _       _       _       _       _       _       _       _       (EntityID=1|MarkableID=markable_5|Min=12)
13      is      _       _       _       _       _       _       _       _       _
14      also    _       _       _       _       _       _       _       _       _
15      called  _       _       _       _       _       _       _       _       _
16      “       _       _       _       _       _       _       _       _       (EntityID=4-Pseudo|MarkableID=markable_6|Min=17
17      T-town  _       _       _       _       _       _       _       _
18      ”       _       _       _       _       _       _       _       _       )
```

Example from GUM, Courtesy Amir Z

Identity+SemType (ARRAU)

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC IDENTITY SPLIT BRIDGING DISCOURSE_DEIXIS REFERENCE NOM_SEM
# newdoc id = GUM_voyage_tulsa
# sent_id = GUM_voyage_tulsa-1
# text = Tulsa
1       Tulsa   _       _       _       _       _       _       _       _       (EntityID=1|MarkableID=markable_1|Min=1|SemType=dn)

# sent_id = GUM_voyage_tulsa-2
# text = Tulsa is in the Green Country region of Oklahoma
2       Tulsa   _       _       _       _       _       _       _       _       (EntityID=1|MarkableID=markable_2|Min=2|SemType=do)
3       is      _       _       _       _       _       _       _       _       _
4       in      _       _       _       _       _       _       _       _       _
5       the     _       _       _       _       _       _       _       _       (EntityID=2|MarkableID=markable_3|Min=7|SemType=dn
6       Green   _       _       _       _       _       _       _       _       _       _
7       Country _       _       _       _       _       _       _       _       _
8       region  _       _       _       _       _       _       _       _       _
9       of      _       _       _       _       _       _       _       _
_
10      Oklahoma        _       _       _       _       _       _       _       (EntityID=3|MarkableID=markable_4|Min=10|SemType=dn))
11      .       _       _       _       _       _       _       _       _

# sent_id = GUM_voyage_tulsa-3
# text = It is also called "T-town"
12      It      _       _       _       _       _       _       _       _       (EntityID=1|MarkableID=markable_5|Min=12|SemType=do)
13      is      _       _       _       _       _       _       _       _       _
14      also    _       _       _       _       _       _       _       _       _
15      called  _       _       _       _       _       _       _       _       _
16      “       _       _       _       _       _       _       _       _       (EntityID=4-
Pseudo|MarkableID=markable_6|Min=17|SemType=predicate
17      T-town  _       _       _       _       _       _       _       _
18      ”       _       _       _       _       _       _       _       _       )
```

Split Antecedents

```
# newdoc id = Artificial_example_2
# sent_id = Artificial_example-1
                Identity                                                    Split

1     John      (EntityID=1|MarkableID=markable_1|Min=1|SemType=dn)     EntityID=1|ElementOf=3
2     met
3     Mary      (EntityID=2|MarkableID=markable_2|Min=2|SemType=dn)     EntityID=2|ElementOf=3
4     .
5     They      (EntityID=3|MarkableID=markable_3|Min=2|SemType=do)
6     went
7    to
```

Bridging

```
# newdoc id = Artificial_example_3
# sent_id = Artificial_example-1
Form        Identity                                Bridging

the         (EntityID=1|MarkableID=markable_1
house       )
.
.
the         (EntityID=2|MarkableID=markable_2    MarkableID=markable_2|Rel=poss|MentionAnchor=markable_1|EntityAnchor=1
door         )
```

Discourse Deixis

```
# newdoc id = Artificial_example_4
# sent_id = Artificial_example-1
# text = John met Mary.
1       John    _       _       _       _       _       _       _       _       (EntityID=1|MarkableID=markable_1|Min=1|SemType=dn)
_       _       (EntityID=1-DD|MarkableID=dd_markable_1|Min=2|SemType=dn
2       met     _       _       _       _       _       _       _       _       _
_       _       _
3       Mary    _       _       _       _       _       _       _       _       (EntityID=2|MarkableID=markable_2|Min=3|SemType=dn)
_       _       _
4       .       _       _       _       _       _       _       _       _                       _               _       _       )


# sent_id = Artificial_example-2
# text = That is not true .
5       That    _       _       _       _       _       _       _       _
(EntityID=3|MarkableID=markable_3|Min=5|SemType=dn)             _       _       (EntityID=1-DD|MarkableID=dd_markable_2|Min=5|SemType=do)
6       is      _       _       _       _       _       _       _       _                       _               _       _       _       _
7       not     _       _       _       _       _       _       _       _                       _               _       _       _       _
8       true    _       _       _       _       _       _       _       _                       _               _       _       _       _
9       .       _       _       _       _       _       _       _       _                       _               _       _       _       _
```

Deixis (reference)

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC IDENTITY SPLIT BRIDGING DISCOURSE_DEIXIS REFERENCE NOM_SEM
# newdoc id = GUM_voyage_tulsa
# sent_id = GUM_voyage_tulsa-1
# text = Tulsa
1       Tulsa   _       _       _       _       _       _       _       _       (EntityID=1|MarkableID=markable_1|Min=1|SemType=dn)
_       _       _       Tulsa,_Oklahoma

# sent_id = GUM_voyage_tulsa-2
# text = Tulsa is in the Green Country region of Oklahoma
2       Tulsa   _       _       _       _       _       _       _       _       (EntityID=1|MarkableID=markable_2|Min=2|SemType=do)
3       is      _       _       _       _       _       _       _       _       _
4       in      _       _       _       _       _       _       _       _       _
5       the     _       _       _       _       _       _       _       _       (EntityID=2|MarkableID=markable_3|Min=7|SemType=dn
_       _       _       Green_Country
6       Green   _       _       _       _       _       _       _       _       _       _
7       Country _       _       _       _       _       _       _       _       _
8       region  _       _       _       _       _       _       _       _       _
9       of      _       _       _       _       _       _       _       _
_
10      Oklahoma        _       _       _       _       _       _       _       (EntityID=3|MarkableID=markable_4|Min=10|SemType=dn))
_       _       _       Oklahoma
11      .       _       _       _       _       _       _       _       _
```

Corpora converted so far

- The CONLL-UA `Exploded' format is the format used in the CODI/CRAC 2021 Shared Task (see later talk by Sopan Khosla)
- Training / dev: ARRAU (in particular TRAINS but also Pear Stories)
- Testing: new data
  - Switchboard-3
  - Light
  - AMI
  - Persuasion

Scoring anaphora resolution systems

- In order for these datasets to be used for anaphora resolution, we need to be able to score systems carrying out this type of interpretation
- For Identity, we have a standard - the Reference Scorer from Pradhan et al (2014) https://github.com/conll/reference-coreference-scorers
- Ported to Python, added LEA (Moosavi & Strube, 2016) https://github.com/ns-moosavi/LEA-coreference-scorer
- For CRAC 2018,
  - the Python version of the Reference Scorer was extended by Moosavi to also score Non-Referring Markables identification
  - A separate script implementing the Bridging scoring from Hou et al was developed
- No scorers for
  - Split Antecedents
  - Discourse Deixis

The new Universal Anaphora scorer

- https://github.com/juntaoy/universal-anaphora-scorer
- Python-based
- An extension of the Python version of the Reference Scorer
- Incorporates the generalization of coreference scoring to split-antecedents by Paun, Yu, Moosavi and Poesio (in preparation)
- Incorporates the CRAC 2018 Bridging scorer
- Uses the Identity scorer for discourse deixis
- Will be tested in the CODI/CRAC Shared Task

Generalizing coreference metrics to unmentioned entities

- A key property of discourse models is that they allow references to an EXTENDED UNIVERSE of entities besides those explicitly mentioned:
  - Entities in the `Implicit Focus' (AKA associative reference – bridging - of various kinds)
  - Entities created as the result of actions
    - Mix the flour with some water, then let THE DOUGH rest for 15'
  - Abstract entities that can be `extracted' from what has been said, aka discourse deixis
  - Entities created by various operations on the existing entities, as in split antecedent anaphora
    - John met Mary. THEY had not seen each other in years.

Coreference chains with split antecedents

Single antecedent entity representation (coreference chain)

$K = \{m1, m5, \ldots.\}$

Single+Split antecedent entity representation

$K = K^S + K^M = \{m1, m5, \ldots.\}$

where

$K^S = \{K1, K16, \ldots.\}$

Example

John met Mary.
THEY had not seen each other in years.

K_1 = {John} K_2= {Mary}

$K_3 = K^S (= \{K_1, K_2\}) + K^M (= \{They\})$

Generalizing MUC Recall

Alignment: tau(K^S) = R^S

$$\delta_i = 1 - \mathbf{MUC}_{\mathbf{Recall}}\left(K_i^s, R_j^s\right)$$

$$\mathbf{Recall} = \frac{\sum_i |K_i| - |\mathcal{P}(K_i^m; R^m)| - \delta_i}{\sum_i |K_i| - 1}$$

Scoring discourse deixis

Approach: follow practice from event coreference resolution of using same metrics as entity coreference (Lu & Ng, 2018)

```
# newdoc id = Artificial_example_4
# sent_id = Artificial_example-1
# text = John met Mary.
1       John    _       _       _       _       _       _       _       (EntityID=1|MarkableID=markable_1|Min=1|SemType=dn)
_       _       (EntityID=1-DD|MarkableID=dd_markable_1|Min=2|SemType=dn
2       met     _       _       _       _       _       _       _       _
_       _       _
3       Mary    _       _       _       _       _       _       _       (EntityID=2|MarkableID=markable_2|Min=3|SemType=dn)
_       _       _
4       .       _       _       _       _       _       _       _                       _               _       _       )


# sent_id = Artificial_example-2
# text = That is not true .
5       That    _       _       _       _       _       _       _
(EntityID=3|MarkableID=markable_3|Min=5|SemType=dn)             _       _       (EntityID=1-DD|MarkableID=dd_markable_2|Min=5|SemType=do)
6       is      _       _       _       _       _       _       _                       _               _       _       _       _
7       not     _       _       _       _       _       _       _                       _               _       _       _       _
8       true    _       _       _       _       _       _       _                       _               _       _       _       _
9       .       _       _       _       _       _       _       _                       _               _       _       _       _
```

Markup discussion

Markup discussion

- The word from Jan Haijc, Joakim Nivre and Marie-Catherine de Marneffe:
  "we tried to fix the format early on, evolving and refining it as the need arose, rather than explore multiple formats in parallel, and I am personally convinced that this was a key to success"
- Can we choose one single format?

An emerging consensus?

- Two mutually convertible formats
  - One compact format compatible with CONLLU for storage in UD repositories / visualization through ANNIS
  - One exploded format compatible with CONLLU+ for anaphora resolution / datasets without other linguistic layers

Issues:

- CONLLU vs. CONLLU+
  - AKA, storing in UD repositories vs. facilitating anaphora resolution work
- CONLLU-compatible variants
- The issue of ambiguity complicates matters

The CODI/CRAC Shared Task

Break

Sameer Pradhan

Practical matters

Practical matters

- Github
  - Reorganization (docs/datasets/software/events)
  - Store all the (open access) data there?
  - Coordinating multiple repositories
- Papers
  - The UA format (CODI/CRAC? LRE?)
  - The UA scorer
  - The UA data repository (LREC? LRE?)
  - CODI/CRAC Shared Task?

Amir Zeldes - ANNIS

Towards 2.0: Guidelines

Towards UA 2.0

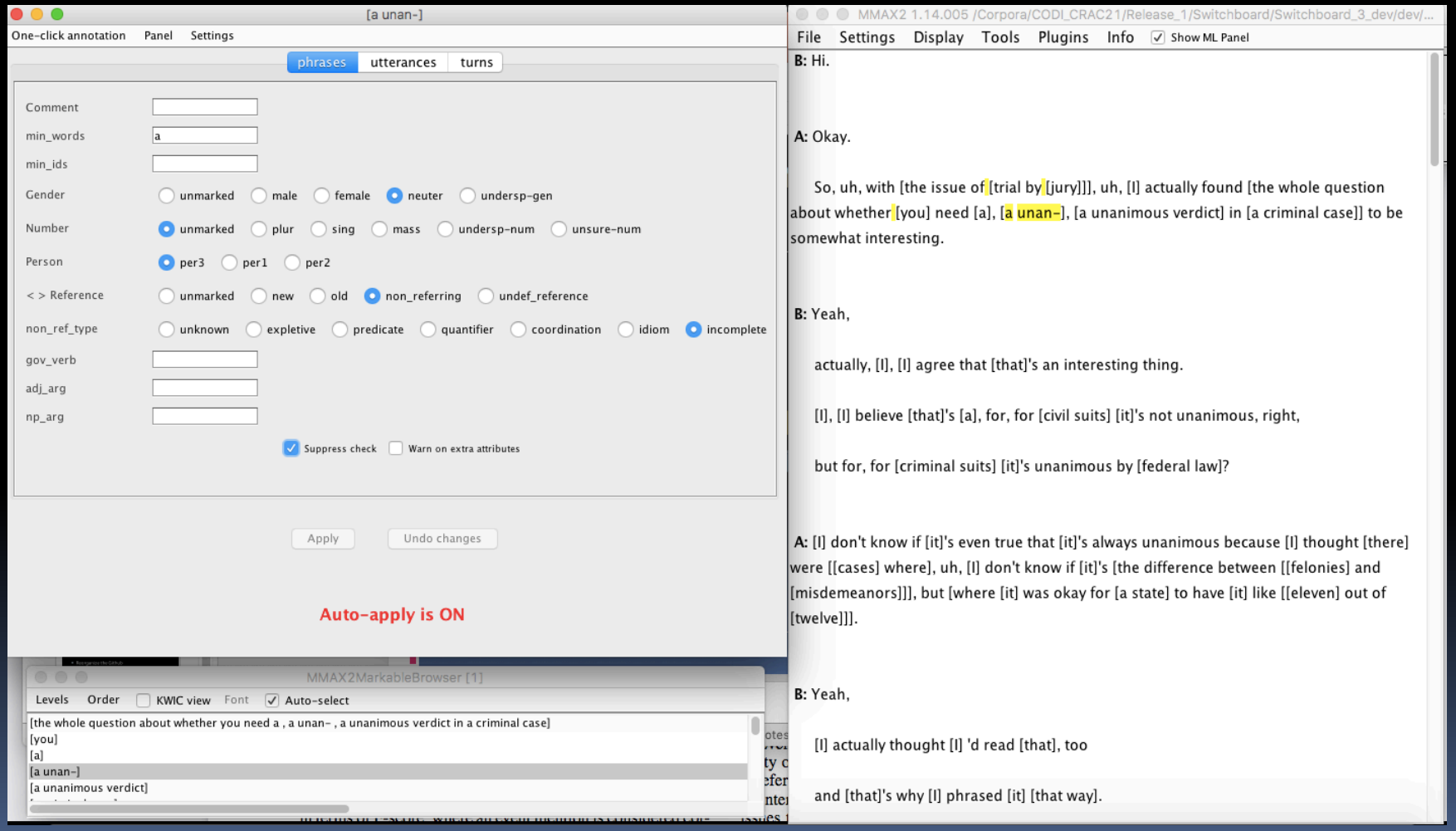
- Is it possible to reach an agreement on some of the aspects of annotation on which there are still differences between existing datasets, such as:
  - Markable definition (allow for discontinuous markables (issue primarily for the scorer)? How to mark zeros (UD approach)?)
  - Non-referring expressions and predication
  - Guidelines for bridging ('relational' vs `cohesive' approaches)
  - Guidelines for discourse deixis
  - Dialogue phenomena
  - Ambiguity

Amir Zeldes - Predication

The Prague Group – Discontinuous markables

Anaphora in dialogue & ambiguity

Anaphoric reference in dialogue



Anaphoric reference in conversation

- Some issues:
  - Markables:
    - Fragments
    - Discontinuous markables
  - References to the visual situation
  - More & looser discourse deixis
  - More ambiguity

Ambiguity

1.1	M	:	all right system
1.2		:	we've got a more complicated problem
1.4		:	first thing _I'd_ like you to do
1.5		:	is send engine E2 off with a boxcar to Corning to pick up oranges
1.6		:	uh as soon as possible
2.1	S	:	okay
3.1	M	:	and while it's there it should pick up the tanker
4.1	S	:	okay
4.2		:	and that can get
4.3		:	we can get that done by three
5.1	M	:	good
5.3		:	can we please send engine E1 over to Dansville to pick up a boxcar
5.4		:	and then send it right back to Avon
6.1	S	:	okay
6.2		:	it'll get back to Avon at 6

Disagreements in the PD corpus

RB ne75965

The day came that had been fixed for the marriage. The bridegroom arrived and also a large company of guests, for the miller had taken care to invite all his friends and relations. As [they] sat at the feast, each guest in turn was asked to tell a tale; the bride sat still and did not say a word.

DO ne75948 {for the miller had taken care to invite [all his friends and relations]} (11,3,1,13),
DO ne75945 {a large company of [guests]} (2,2,2,2),
DN (10,3,1,12),
DO ne75936 {the girl}, ne75942 {[the bridegroom]}, ne75945, ne75948 (1,1,3,-1),
DO ne75942, ne75946 {[the miller]}, ne75948 (2,2,2,2),
DO ne759370001 {[her]}, ne75942 {[the bridegroom]}, ne75945 {[the large company of guests]},
ne759490001 {his (the miller)} (1,0,4,-3,e2,e18),
DO ne75942 ne75948 ne759370001 ne75946 (1,3,1,3),
DO ne75942 ne75948 ne759370001 (1,2,2,1),
DO ne75942 ne759370001 ne75945 (1,0,4,-3),
DO ne75948 ne75946 (2,1,3,0),
DO ne75942 ne75948 ne759370001 ne75945 ne75946 (2,1,3,0),
DO ne75936 ne75937 {her father aka the miller} ne75942 ne75948 (1,0,4,-3)

+ 2 not_selectable, 3 skips

81 A+V, 5 comments skips, Total: 86 judgments

Next steps

Next steps

- More recruitment
- Reorganize the Github
- Agree on a markup format
- Convert datasets & start populating the UA repository
- Explore using ANNIS as a way of visualizing the UA data
- Paper on the UA scorer
- Set up sub-groups for discussion of guidelines
- Organize UA workshop at LREC 2022? Or stay under the umbrella of CODI/CRAC?

Final discussion

Discussion

- Do all three formats encode the same information?
- CorefUD vs UA?