Universal Anaphora

Collapsed format data tests and tricky coreference guidelines

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

Logan Peng
Georgetown University
sp1184@georgetown.edu

GEORGETOWN UNIVERSITY

Corpling@GU

CoNLL-UA (collapsed) proposal

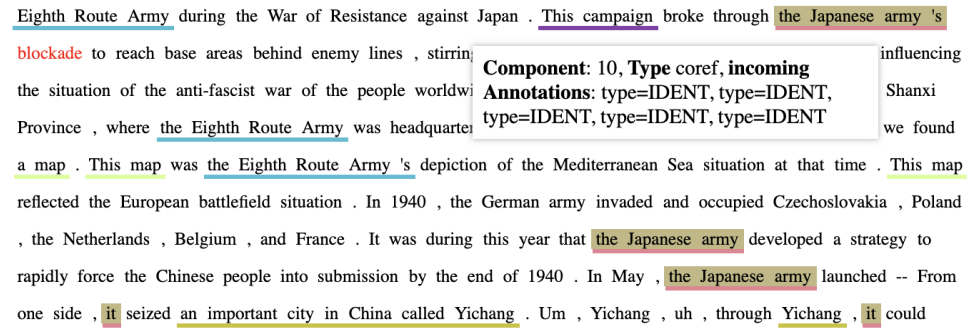
https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/UA_CONLL_U_proposal_amir.md

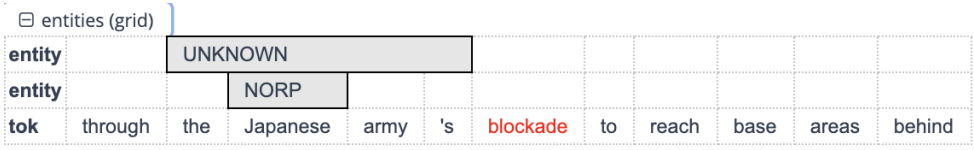
1. Uses familiar round bracket CoNLL scorer notation: e.g. Entity=(1-person)2-organization)
2. Very compact
3. Human readable
4. No co-indexing multiple annotations for markables
5. Supports spans that cross sentence boundaries
6. Supports entity linking/wikification
7. Supports metadata
8. Includes bridging/split antecedent + discontinuous markables
9. Compatible with ANNIS search

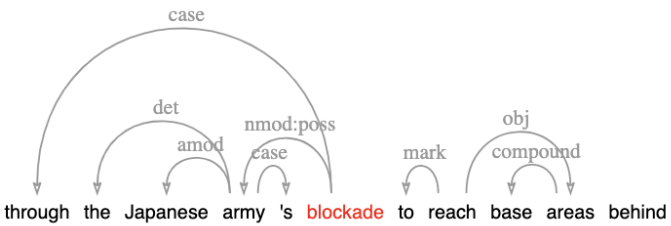
https://corpling.uis.georgetown.edu/annis/ua/
anaphora<>universal

OntoNotes - English

- Meta:
  - Followed conll 2012 train/dev/test sets and only include documents that have coref annotations
  - Since coref IDs might conflict across different sections within a document, we regard a section as a unit document
  - Tests in ON are named by *folder_doc_section (e.g. bc_cctv_0001_001)*
- Tokens & Syntax: converted ON constituent trees to UD trees using CoreNLP







OntoNotes - English

- Included annotations:
  - Prop & Sense
  - Name (entity) annotations: entity spans and types
  - Coref annotations: spans and coref chains
- Entity types are migrated from Name annotations to Coref annotations; as well as through an IDENT coref chain
- Some spans have type=UNKNOWN

```
1   This    this    DET DT   _   2   det _   Entity=(EVENT-IDENT-5
2   campaign    campaign    NOUN    NN  _   3   nsubj       Entity=EVENT-IDENT-5)
3   broke   break   VERB    VBD _   0   root    _
Prop=break.01|Sense=break-v.16.5
4   through through ADP IN   _   9   case    _   _
5   the the DET DT   _   7   det _   Entity=(UNKNOWN-IDENT-13
6   Japanese    japanese    ADJ JJ  _   7   amod    _   Entity=(NORP-SGL-14)
7   army    army    NOUN    NN  _   9   nmod:poss    _   SpaceAfter=No
8   's  's  PART    POS _   7   case    _   Entity=UNKNOWN-IDENT-13)
9   blockade    blockade    NOUN    NN  _   3   obl
10  to  to  PART    TO  _   11  mark    _   _
11  reach   reach   VERB    VB  _   3   xcomp   _   Prop=reach.01|Sense=reach-v.1
12  base    base    NOUN    NN  _   13  compound
13  areas   area    NOUN    NNS _   11  obj _   _
14  behind  behind  ADP IN   _   16  case    _   _
15  enemy   enemy   NOUN    NN  _   16  compound    _   _
16  lines   line    NOUN    NNS _   13  nmod    _   SpaceAfter=No
17  ,   ,   PUNCT   ,   _   3   punct   _   _
```

GAP (English)

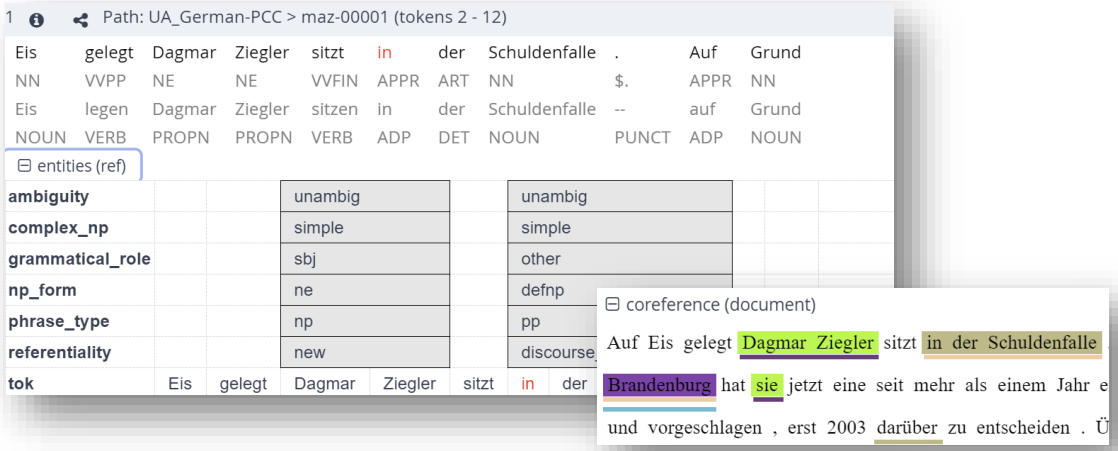
- One pronoun and two candidate NPs
- Stanza parsed UD with predicted token boundaries

```
# newdoc id = train_1738
# meta_url = http://en.wikipedia.org/wiki/Diane_Robertson
# meta_partition = train
# sent_id = train_1738-1
# text = Robertson was born in Waipukurau in 1953, the daughter of Joan Lois Coburn
and her husband Alexander Lawrence Coburn.
1	Robertson	Robertson	PROPN	NNP Number=Sing 3	nsubj:pass	_
Entity=(PERSON-SGL-0)
2	was be	AUX VBD Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	3
aux:pass	_	_
3	born	bear	VERB	VBN Tense=Past|VerbForm=Part	0	root	_	_
4	in	in	ADP IN	_	5	case	_	_
5	Waipukurau	Waipukurau	PROPN	NNP Number=Sing 3	obl _	_
6	in	in	ADP IN	_	7	case	_	_
7	1953	@card@	NUM CD	NumType=Card	3	obl _	SpaceAfter=No
8	,	,	PUNCT	,	_	10	punct	_	_
9	the the DET DT	Definite=Def|PronType=Art	10	det _	_
10	daughter	daughter	NOUN	NN	Number=Sing 3	parataxis	_	_
11	of	of	ADP IN	_	12	case	_	_
12	Joan	Joan	PROPN	NNP Number=Sing 10	nmod	_	Entity=(PERSON-IDENT-1
13	Lois	Lois	PROPN	NNP Number=Sing 12	flat	_	_
14	Coburn	Coburn	PROPN	NNP Number=Sing 12	flat	_	Entity=PERSON-IDENT-1)
15	and and CCONJ	CC	_	17	cc	_	_
16	her her PRON	PRP$	Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs
17	nmod:poss	_	Entity=(PERSON-IDENT-1)
17	husband husband NOUN	NN	Number=Sing 12	conj	_	_
18	Alexander	Alexander	PROPN	NNP Number=Sing 12	conj	_	_
19	Lawrence	Lawrence	PROPN	NNP Number=Sing 18	flat	_	_
20	Coburn	Coburn	PROPN	NNP Number=Sing 18	flat	_	SpaceAfter=No
21	.	.	PUNCT	.	_	3	punct	_	_
```

PCC (German)

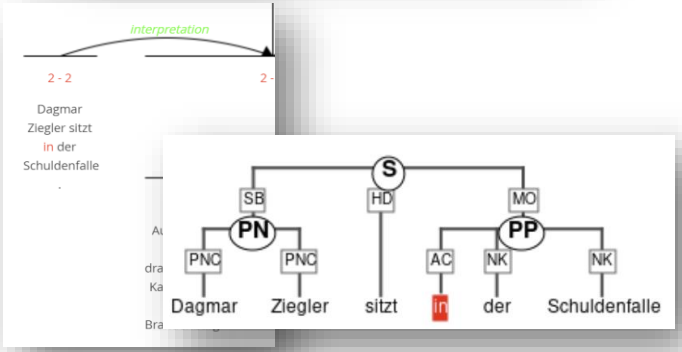
Represented in collapsed UA format

- Coreference (mmax) with markable annotations:
  - phrase_type
  - complex_np
  - np_form
  - referentiality
  - grammatical
  - role-ambiguity



Merged into ANNIS corpus

- Syntax (TigerXML constituent trees)
- RST discourse parses



PCC (German)

```
# newdoc id = maz-00001
# global.Entity = GRP-phrase_type-complex_np-np_form-referentiality-grammatical_role-ambiguity
# meta::id = 00001
# meta::date = 30.10.2002
# meta::section = UNKNOWN
# meta::author = STEPHAN BREIDING
# meta::header = Auf Eis gelegt
# sent_id = maz-00001-1
1	Auf	auf	ADP	APPR	_	2	case	_	_
2	Eis	Eis	NOUN	NN	_	3	obl	_	_
3	gelegt	legen	VERB	VVPP	_	0	root	_	_

# sent_id = maz-00001-2
1	Dagmar	Dagmar	PROPN	NE	_	3	nsubj	_	Entity=(1-np-simple-ne-new-sbj-unambig
2	Ziegler	Ziegler	PROPN	NE	_	1	flat	_	Entity=1-np-simple-ne-new-sbj-unambig)
3	sitzt	sitzen	VERB	VVFIN	_	0	root	_	_
4	in	in	ADP	APPR	_	6	case	_	Entity=(2-pp-simple-defnp-discourse_cataphor-other-unambig
5	der	der	DET	ART	_	6	det	_	_
6	Schuldenfalle	Schuldenfalle	NOUN	NN	_	3	obl	_	Entity=2-pp-simple-defnp-discourse_cataphor
7	.	--	PUNCT	$.	_	3	punct	_	_
```

Markable annotation names

- UD parse by Stanza from gold POS tagged/tokenized input
- Metadata merged from original corpus
- Coref and markable annos from mmax (incl. singletons)

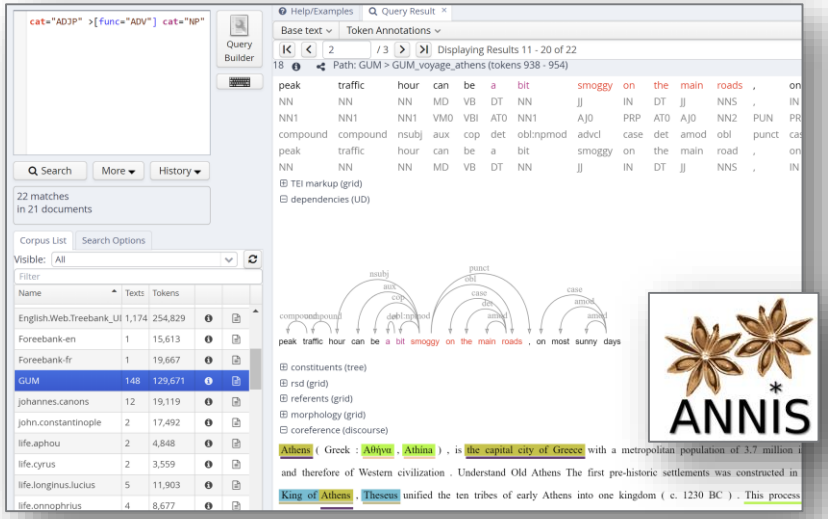
Tests so far

- UA_English-GAP (gender balanced personal pronouns)
- UA_English-GUM (broad multilayer coverage, 12 genres)
- UA_English-OntoNotes (huge multilayer benchmark data)
- UA_French-Democrat1921 (coref long wikis and articles)
- UA_German-PCC (multilayer editorials corpus)

Now looking at ARRAU for English, more to come!

Search in ANNIS (details later today)

- We use ANNIS (Krause & Zeldes 2016) to concurrently search and visualize UA and other annotations
- Support for importing compact CoNLL-UA format
- Can merge annotations in other supported formats if tokenization matches (RST, constituents and more)
  - Supported formats: https://corpus-tools.org/pepper/knownModules.html
  - Goal: same as UD search on http://match.grew.fr/)



https://corpling.uis.georgetown.edu/annis/ua/
anaphora<>universal

Coref, genericity and predication

Principles I would like to see UA commit to:

1. a focus on semantics, not morphosyntax
2. cross-linguistic generalizability
3. a separation of **identity** and **scope** (later)

Indefinites != generics != non-referring

In ON, indefinites are seen as generic and excluded as anaphors. But:

1. Indefinite mentions are often neither generic nor underspecified or abstract ("[participants] comprised [15 women and 10 men]")
2. It is not immediately obvious that we should not want coreference information even for mentions which are generic, abstract, etc.
3. In context it's hard to know whether a pronoun is generic, and even if it is, generic pronouns can still form multiple distinct clusters
4. Many languages **do not have widespread articles** to identify 'generic' mentions, even if we agreed that all indefinites should be considered generic

Marbles is the first social media star to have [a wax figure]$_i$ displayed in Madame Tussauds ... In 2015 , Marbles unveiled [a wax figure of herself]$_i$ at Madame Tussauds
(GUM_bio_marbles)

Indefinites != generics != non-referring

- Not annotated in ON:
  - [Program trading] is “a racket,” ... [program trading] creates deviant swings
  - [you] couldn’t start unless [you] knew that the replacement heart would make it to the operating room
- Do we want these?
  - I have [a mini-MMPI] ... I have [a chart that I’ll go through]
  - [You]$_1$ feel like [you]$_1$’re prepared , [you]$_1$’re in a , [you]$_2$ know , in a relationship ... [you]$_2$ know
    (1=someone who became pregnant; 2=any listener)

What about other languages?

- “死刑(capital punishment), 世界(world – as distinct from “the world” meaning the planet Earth), 社会(society) are considered generic nouns” (BBN Technologies, 2008, 8)

- 我们能不能发展的快一些、好一些，实现经济快速发展和[社会] 全面进步，并且保 持[社会] 稳定，十分重要(ON, cnr_0016)
- Whether our development can progress faster and better to make it possible for the economy to grow quickly and for [society] to make progress across all metrics as well as to maintain the stability of [society] is very crucial

Compound modifiers

Thought to be “anaphoric islands”

*[Animal]i hunters tend to like [them]i. (Postal, 1969, 230)

- Only included in English(!) ON if they are proper nouns:
  - The [Hong Kong] government’s jurisdiction is the [Hong Kong] Special Administrative Region (included in ON)
  - small investors seem to be adapting to greater [stock market] volatility . . . Glenn Britta . . . is “factoring” [the market’s] volatility “into investment decisions.” (**NOT** annotated in ON)
- Annotated in GUM based solely on semantics:
  - [Cinnamon] basil really does smell like [the sweet spice] (GUM_whow_basil)

What about other languages?

- Construct state compound modifiers regularly included in Arabic:

محاكمة ضباط روس بتهمة الاهمال ... محاكمة ... لثلاثة ضباط روس (ON, ann_0006)

muḥākamatu ḍubbātin rūsi bituhmati l-ihmāli... muḥākamatun... li-ṯalāṯatin ḍubbātin rūsi

[Russian Officer]$_i$ Trial on Charges of Negligence... a trial... for [three Russian officers]$_i$

- What would happen if we develop multilingual applications with coreference resolution? Notice guidelines conflict here **within OntoNotes**
- Can we even identify compound modifiers unambiguously across languages?

Predication

- Kicked out of coref after ACE (van Deemter & Kibble 2006, Zaenen 2006):
  - [Henry Higgins, who was formerly [sales director of Sudsy Soaps]i]i, became [president of Dreamy Detergents]i
  - If [Beyoncé]i were [the Queen of England]i, [she]i would....
- Handled heterogeneously in ON guideline (BBN Technologies, 2007, 27):
  - [Elizabeth II]i is the Queen of England. [She]i ...
  - The Queen of England is [Elizabeth II]i. [She]i ...
  - She was crowned [Elizabeth II]i in 1953. [She]i ...

Is the problem actually predication?

- The scope problems don’t really come from copula predication as a syntactic construction
- We can create semantic clashes using plain definite NPs without a copula:
  - [A fresh major in the Swedish army,] in 1812 [Gordon] went to war . . . In 1875 [the now general in the Russian army] was ready to pursue [his] ultimate achievement. . . [Gordon] is buried in. . .
- But, we all understand that this is about **the same person**

Isn’t syntax enough?

No. For example:

- Modals:
  - a. [He] **would be** a Libertarian today (no coref)
  - b. [The principles governing an F-E translation] **would then be**: [reproduction of grammatical units; consistency in word usage; and meanings in terms of the source] (coref)
- Subtance predication:
  - a. [This coffee table] is glass
  - b. [This ice here] of course is [water] (coref, part of a chemistry demonstration in which the speaker literally identifies an ice cube as being the same water in a solidified state)
- Complex negation:
  - [He] was not the leaf-collecting doctor, but [an altogether strange man, with silver eyebrows in his smooth face and long fine-knuckled hands]
- Spatio-temporal:
  - This town is 35 minutes from the harbor
  - But Christmas is still the whole winter to wait

Unexiling predication

- The problems with predication are not reason enough to throw out **Identity Predication** (*I am Amir*)
- The vast majority of non-identity predication are also not problematic (*Kim is a teacher who lives on 22 Main Street with 2 cats*)
- Many entities change over the course of a text and we still allow their mentions to corefer

- Kicking these out is throwing out the baby with the bath water!

What about scope?

- Some corpora have attempted to address scope (RED, O'Gorman et al. 2016, other scoping initiatives see Hendrickx et al. 2012, Nissim et al. 2013, Rubinstein et al. 2013 for modality; temporal scope: Pustejovsky and Stubbs 2011, Styler IV et al. 2014, Pustejovsky et al. 2019)
- Basically:
  - If <coref id="1">Beyoncé</coref> were <coref id="1" scope="s1">the Queen of England </coref>, <coref id="1" scope="s1">she</coref> would conduct the annual swan upping.

- I think this is the right direction, but we can annotate predication for now without doing this
- Leave scope as advanced research topic, like bridging etc.

How big are these problems?

expression type	instances per 1K tokens	% of total
pron. anaphora	59.84	44.98%
cataphora	1.39	1.05%
nominal predicate	4.94	3.71%
compound mod.	5.79	4.36%
split antecedent	1.22	0.92%
apposition	3.81	2.86%
coref. name	19.84	14.92%
other indef. NP	14.53	10.92%
other coref	21.67	16.29%
total	133.02	100%

Table 1: Coreference link type distribution in GUM