Universal Anaphora: The First Three Years

Massimo Poesio,[1,2] Maciej Ogrodniczuk,[3] Vincent Ng,[4] Sameer Pradhan,[5] Juntao Yu,[1] Nafise Sadat Moosavi,[6] Silviu Paun,[7] Amir Zeldes,[8] Anna Nedoluzhko,[9] Michal Novák,[9] Martin Popel,[9] Zdeněk Žabokrtský,[9] and **Daniel Zeman[9]**

[1]Queen Mary University, UK; [2]University of Utrecht, The Netherlands;
[3]Institute of Computer Science, Polish Academy of Sciences; [4]University of Texas at Dallas;
[5]LDC, University of Pennsylvania, USA; [6]University of Sheffield, UK; [7]Amazon;
[8]Georgetown University, USA; [9]Charles University, Czechia;
m.poesio@qmul.ac.uk; maciej.ogrodniczuk@ipipan.waw.pl; vince@hlt.utdallas.edu;
pradhan@cemantix.org; j.yu@qmul.ac.uk; n.s.moosavi@sheffield.ac.uk;
spaun3691@gmail.com; Amir.Zeldes@georgetown.edu;
{nedoluzko,mnovak,popel,zabokrtsky,zeman}@ufal.mff.cuni.cz

Abstract

The aim of the Universal Anaphora initiative is to push forward the state of the art in anaphora and anaphora resolution by expanding the aspects of anaphoric interpretation which are or can be reliably annotated in anaphoric corpora, producing unified standards to annotate and encode these annotations, delivering datasets encoded according to these standards, and developing methods for evaluating models that carry out this type of interpretation. Although several papers on aspects of the initiative have appeared, no overall description of the initiative's goals, proposals and achievements has been published yet except as an online draft. This paper aims to fill this gap, as well as to discuss its progress so far.

Keywords: anaphora resolution, coreference, Universal Anaphora, CorefUD, bridging references, split-antecedent anaphora, discourse deixis, discontinuous markables, zero anaphora, dialogue

1. Introduction

In recent years, the attention of the anaphoric interpretation / coreference community in NLP has started to turn to more complex cases of anaphora, to genres, and to languages not represented in the reference OntoNotes dataset[1] (Weischedel et al., 2011; Pradhan et al., 2013). This trend is illustrated by research on anaphora whose interpretation requires some form of commonsense knowledge, tested by benchmarks for the Winograd Schema Challenge (Rahman and Ng, 2012; Liu et al., 2017; Sakaguchi et al., 2020), or on pronominal anaphors that cannot be resolved purely using gender, for which benchmarks such as GAP have been developed (Webster et al., 2018). Another fruitful line of research has been devoted to creating datasets covering genres other than news, such as conversation (Muzerelle et al., 2014; Uryupina et al., 2020; Khosla et al., 2021; Yu et al., 2022a), literature (Bamman et al., 2020) or scientific articles (Cohen et al., 2017).

Further research has been carried out on aspects of anaphoric interpretation beyond identity anaphora that are covered by datasets such as ARRAU (Poesio et al., 2018; Uryupina et al., 2020) GUM (Zeldes, 2017) and GENTLE (Aoyama et al., 2023) for English, the Prague Dependency Treebank (Nedoluzhko, 2013) for Czech, and ANCORA for Catalan and Spanish (Recasens and Martí, 2010). These include bridging reference (Clark, 1977; Hou et al., 2018; Hou, 2020; Yu and Poesio, 2020; Kobayashi and Ng, 2021), discourse deixis (Webber, 1991; Marasović et al., 2017; Kolhatkar et al., 2018) or split-antecedent anaphora (Eschenbach et al., 1989; Vala et al., 2016; Zhou and Choi, 2018; Yu et al., 2020, 2021; Paun et al., 2023).

The **Universal Anaphora** initiative, or UA,[2] was launched in 2020 to coordinate these efforts to push forward the state of the art in anaphora research. The initiative, modelled on Universal Dependencies,[3] aims to achieve this by expanding the aspects of anaphoric interpretation which are or can be reliably annotated in anaphoric corpora, producing unified standards to annotate and encode these annotations, delivering datasets encoded according to these standards, and developing methods for evaluating this type of interpretation. In parallel, the CorefUD project was also launched, with the related aim of developing standards for adding anaphoric information to corpora annotated according to Universal Dependencies (Nedoluzhko et al., 2022). These two initiatives have since collaborated closely, particularly on the UA scorer (Yu et al., 2023).[4]

Although several papers on aspects of the UA

[1] https://catalog.ldc.upenn.edu/LDC2013T19

[2] http://www.universalanaphora.org

[3] https://universaldependencies.org/

[4] We should also mention that an ISO standard for reference exists (International Organization for Stan-

initiative have appeared (Khosla et al., 2021; Yu et al., 2022a, 2023; Paun et al., 2023), no description of its goals and proposals has been published yet except the original online document from December 2020.[5] This paper aims to fill this gap, as well as discussing the progress since then, thus serving as a sort of road map to the variety of existing resources and ideas, highlighting where further progress is needed.

The structure of the paper is as follows. The objectives of Universal Anaphora are discussed in Section 2. Next, we discuss in detail the developments concerning these objectives: the proposed coverage (Section 3), the two proposals concerning markup developed since 2020 (Section 4), and the Universal Anaphora scorer (Section 7). We also briefly summarize the objectives and achievements of CorefUD (Section 6). We then discuss the activities of the initiative since 2020, including the three shared tasks organized and the repositories. Finally, we report on the key discussions on markup and scoring held as part of the initiative, and discuss open issues.

2. Objectives

The proponents of UA were very much aware that there is at still only partial agreement on the anaphoric phenomena that should be covered by such a scheme, and on the details of how they should be annotated (Zaenen, 2006; Poesio et al., 2016; Zeldes, 2022; Poesio et al., 2024).

As a result, more modest initial objectives were set. First of all, to **catalogue** the aspects of anaphoric reference annotated in the great number of existing projects worldwide. Second, to come up with an agreed-upon **markup scheme** that could be used to encode such anaphoric information; this would in turn enable the creation of a collection of corpora all encoded using the same scheme. Third, and crucially, to develop a **scorer** extending the reference CoNLL scorer (Pradhan et al., 2014) and able to evaluate the interpretation not just of identity anaphora, but also of the other aspects of anaphoric interpretation included in the coverage, such as the identification of split-antecedent plurals, non-referring expressions, bridging reference, and discourse deixis. The hope was that, as with Universal Dependencies, these initial developments would prove useful starting points for further discussion on the annotation schemes as well. These objectives have largely been achieved; we discuss them in turn in the following Sections.

dardization, 2019). The standard does not catalogue anaphoric reference phenomena in great detail, focusing instead primarily on the specification of a TEI-compatible XML markup format, but this latter is based on the same conceptual framework for (anaphoric) reference assumed here.

[5] https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/documents/Universal_Anaphora_1_0___Proposal_for_Discussion.pdf

3. Aspects of anaphoric reference covered by Universal Anaphora

The specification of the aspects of anaphoric reference to be covered by Universal Anaphora in the original proposal was designed (i) to cover all aspects of anaphoric information currently annotated in existing projects (see (Poesio et al., 2016) for a review and (Nedoluzhko et al., 2021) for a detailed discussion of which aspects are covered by which corpora) (ii) identifying some of these aspects as required, but (iii) without requiring all projects to annotate all of this information, and (iv) leaving room for future extensions (e.g., to cover ellipsis).

3.1. The definition of markable

The first aspect to consider is what is to be counted as 'anaphoric expression' or **markable**[6]. In Universal Anaphora 1.0, it is assumed that markables are defined on syntactic grounds but no further restrictions are specified, because differences remain, as illustrated by the following examples.

Reference and generic terms In corpora such as OntoNotes, a hybrid approach to reference with generic terms is adopted. Generic reference using pronouns and nominals to generic antecedents, as in (1), is annotated; but generic reference via bare plurals, as in (2), is not.

(1) [Meetings]$_i$ are most productive when [they]$_i$ are held in the morning. [Those meetings]$_i$, however, generally have the worst attendance.

(2) Allergan Inc. said it received approval to sell the PhacoFlex intraocular lens, the first foldable silicone lens available for [cataract surgery]$_i$. The lens' foldability enables it to be inserted in smaller incisions than are now possible for [cataract surgery]$_i$.

In ARRAU and in GUM, on the other end, all of these types of coreference are uniformly annotated. Universal Anaphora 1.0 does not legislate on whether / which generic references should be annotated, but it allows for an attribute encoding this information such as ARRAU's `GENERIC` attribute on the core layer of the Universal Anaphora scheme, `Identity` (see Section 4).

[6] The term 'mention' is also used in this sense, e.g. in the CorefUD initiative.

Reference and prenominal modification Another difference between the definition of markables in existing corpora is whether identity between references to kinds expressed as bare nominals in premodifier position, as in (3), is annotated. Such premodifier nominals are considered as markables in both ARRAU and GUM/GENTLE, whereas in ONTONOTES only proper nouns are treated as markables (e.g. ***Hong Kong*** *government* can contain a reference to ***Hong Kong***).

(3) Even the volatility created by [stock]$_i$ index arbitrage and other computer-driven trading strategies isn't entirely bad, in Mr. Connolly's view.For the long-term investor who picks [stocks]$_i$ carefully, …

Again, UA 1.0 does not legislate on this issue, but prenominal markables are currently allowed (and included in the datasets in the current Universal Anaphora collection, see Section 5).

Discontinuous markables Agreement is still lacking on how to annotate many aspects of anaphoric reference in dialogue. One such aspect that was extensively discussed during the development of the Universal Anaphora format and in connection with the shared tasks, ultimately leading to revisions to the Universal Anaphora scorer, is the fact that in dialogue referential expressions often cannot be clearly associated with contiguous syntactic constituents. Such discontinuous markables are exemplified, for instance, by (4) (from the `trains` subset of the ARRAU corpus (Poesio et al., 2024)), where the referring expression *a tanker ... of orange juice* is started by M in u1, interrupted by S's acknowledgment in u2, and completed in u3.

(4)
u1	M	[a tanker
u2	S	yeah
u3	M	of orange juice]

Universal Anaphora 1.0 is agnostic as to whether discontinuous markables should be allowed in a corpus or not, but the markup should be designed to support the representation in Universal Anaphora format of corpora that do contain them, and the scorer should be able to evaluate systems performing anaphora resolution in such datasets, as discussed later in the paper.

Zero anaphors such as ∅ in (5) are one of the most common forms of anaphoric reference in languages which allow unrealized arguments such as Arabic, Chinese, Czech, Italian, or Japanese.

(5) [IT] [Giovanni]$_i$ è in ritardo, così [∅]$_i$ mi ha chiesto se posso incontrarlo al cinema.
[EN] [John]$_i$ is late so [he]$_i$ asked me if I can meet him at the movies. ((Poesio et al., 2016), ex. 9, p. 29)

Zero anaphora is annotated in Arabic and Chinese ONTONOTES, as well as the ANCORA corpus for Catalan and Spanish (Recasens and Martí, 2010), the LIVEMEMORIES corpus for Italian (Rodriguez et al., 2010), the NAIST corpus for Japanese (Iida et al., 2017), the Prague Dependency Treebank (Hajič et al., 2020) and Czech-English Dependency Treebank (Nedoluzhko et al., 2016), among others, but using different markup methods, some of which assume the existence of other layers of annotation, such as dependency structure.

Zero anaphora is included among the aspects of anaphoric reference that should be covered by Universal Anaphora 1.0, but without requiring corpora to annotate it. Also, there are currently no official specifications of how it should be encoded in the markup, but the UA scorer only supports one format, as discussed in Section 7.

3.2. Referring vs. non-referring markables

Another difference between the definition of markable in different corpora is whether only anaphoric expressions are marked, or all nominal expressions. In ONTONOTES, for instance, references to entities only mentioned once (**singletons**) are not annotated, and neither are **expletives**, the subclass of **non-referring expressions** consisting of semantically vacuous noun phrases, such as *it* in (6). On the other end, a second sub-class of non-referring expressions, **predicate nominals** such as *a busy place* in (6), were annotated. Singletons, expletives and predicate nominals have all been annotated in more recent corpora.

(6) [It] seems to be [a busy place]

The Universal Anaphora 1.0 specification recommended to mark all nominal phrases, and optionally to include in the `Identity` layer an attribute (`SemType`) specifying whether a nominal phrase is referring, predicative, or an expletive. Such attribute is used by the UA scorer to evaluate a system's ability to recognize non-referring nominals.

3.3. Anaphoric relations

Identity Anaphora Most modern anaphoric annotation projects cover basic identity anaphora as in (7). UA 1.0 requires all cases of basic identity anaphora to be marked in the Identity layer.

(7) [Mary]$_i$ bought [a new dress]$_j$ but [it]$_j$ didn't fit [her]$_i$.

However, many other types of identity anaphora exist and are annotated in other corpora. **Split-antecedent anaphors** (Eschenbach et al., 1989; Kamp and Reyle, 1993) are cases of plural identity anaphora as in (8), where plural anaphor *they*

refers to a set of two or more entities introduced by separate noun phrases. Such references are annotated in, e.g., ARRAU (Uryupina et al., 2020), GUM (Zeldes, 2017), GENTLE (Aoyama et al., 2023) and *Phrase Detectives* (Poesio et al., 2019).

(8) [John]$_1$ met [Mary]$_2$. [He]$_1$ greeted [her]$_2$. [They]$_{1,2}$ went to the movies.

Split-antecedent plural reference was not evaluated by the Reference Coreference Scorer (Pradhan et al., 2014). UA 1.0 does not require for such cases to be annotated, but the markup allows them to be encoded, and the scorer can evaluate their interpretation, as discussed below.

Discourse deixis In ONTONOTES, **event anaphora**, a subtype of **discourse deixis** (Webber, 1991; Kolhatkar et al., 2018) is marked, as exemplified by *that* in (9), which refers to the event of a white rabbit with pink ears running past Alice; but abstract anaphors such as *this*, which refers to the fact that the Rabbit was able to talk, are not. A more extensive annotation of event anaphora is found in corpora such as the multi-sentence AMR corpus (O'Gorman et al., 2018) and more complex discourse deictic references are marked in, e.g., ANCORA and ARRAU.

(9) ... when suddenly a White Rabbit with pink eyes ran close by her. There was nothing so VERY remarkable in [that]; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at [this], but at the time it all seemed quite natural); ....

Universal Anaphora 1.0 does not require discourse deixis to be annotated, but it specifies that discourse deixis should be annotated in a separate layer, but following the same markup format as for other types of identity anaphora (see Section 4) so that the scorer can evaluate discourse deixis using the same metrics as for other types of identity anaphora, as discussed in Section 7.

Non-identity anaphora Possibly the most studied case of non-identity anaphora is **bridging reference** or **associative anaphora** (Clark, 1977; Hawkins, 1978; Prince, 1981) as in (10), where bridging reference / associative anaphora *the roof* refers to an object which is related to / associated with, but not identical to, the *hall*.

In UA, marking bridging references is not mandatory, but the markup format allows for such types of anaphoric reference to be encoded in a separate layer, and the scorer can evaluate such types of anaphoric reference. The layer for non-identity anaphora is also used to encode ***other* anaphora** such as *the other* in (11).

(10) There was not a moment to be lost: away went Alice like the wind, and was just in time to hear it say, as it turned a corner, 'Oh my ears and whiskers, how late it's getting!' She was close behind it when she turned the corner, but the Rabbit was no longer to be seen: she found herself in [a long, low hall, which was lit up by a row of lamps hanging from [the roof]].

(11) There were doors all round the hall, but they were all locked; and when Alice had been all the way down [one side] and up [the other], trying every door, she walked sadly down the middle, wondering how she was ever to get out again.

Identity of sense, as in *one*-anaphora, is exemplified by *John bought a red shirt, and Bill [a blue one]* (Poesio, 2016). In ARRAU and in GUM, *one*-anaphora is marked as a type of non-identity anaphora. This is the approach followed in Universal Anaphora 1.0 as well.

4. Markup

The markup format proposed in UA, called CoNLL-UA,[7] is based on the CoNLL-U-Plus tabular format proposed in Universal Dependencies for corpora containing additional linguistic annotations.[8]

The key modification is the introduction of new layers devoted to the representation of anaphoric information. The format specifies the following layers in addition to those defined in UD:

- `Identity` (required), specifying the–possibly discontinuous–markables (noun phrases, nominal modifiers, zeros, etc.), and the entity a markable refers to in the case of a referring markable (as in the CoNLL coreference scheme). In addition to coreference information, this layer may contain additional optional attributes specifying whether the markable is referring or not (attribute `SemType`), and what its head is (attribute `Min`). This layer is also used for split antecedents, to indicate the set they belong to.

- `Bridging` (optional), specifying the anchor, its most recent mention, and, optionally, the associative relation.

[7] https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/documents/UA_CONLL_U_Plus_proposal_v1.0.md

[8] https://universaldependencies.org/ext-format.html

- `Discourse_Deixis` (optional), whose markables specify the non-nominal antecedents of discourse deixis, represented exactly as in the `Identity` layer. This makes it possible to adopt for discourse deixis the same metrics used for identity anaphora.

- `Nom_Sem` (optional), for information about nominal semantics not already included in the CoNLL-U layers - ontological category, genericity, etc.

The CoNLL-UA format was designed to provide a way to specify anaphoric information independent from other layers, but compatible with the UD format. However, at present the UD validation software does not allow the UD-released datasets to use the CoNLL-U-Plus format. Thus, UA collaborated with CorefUD to design a more 'compact' format that could be used to pack the anaphoric information representable in CoNLL-UA in the '`MISC`' column of the CoNLL-U format, and is fully compatible with the Universal Dependencies. The two formats are mutually interchangable, and the UA scorer can read either format. The CorefUD format is discussed in Section 6.

5. The Universal Anaphora Datasets

5.1. Existing Datasets

A number of existing datasets have been converted to CoNLL-UA or the equivalent, more compact CorefUD format. The copyright-free subcorpora of ARRAU and the *Phrase Detectives* corpora are available from the Universal Anaphora GitHub. 17 datasets for 12 languages are available from the CorefUD repository (see Section 6).

5.2. The CODI-CRAC 2022 Corpus

In addition, several new datasets are available in CoNLL-UA format from the Universal Anaphora repository. Of these, the most widely used is the CODI-CRAC 2022 corpus, created for the CODI-CRAC Shared Task on anaphora resolution in dialogue (see Section 8.1).

The corpus created for CODI-CRAC 2021 and 2022 consists of conversations from well-known conversational datasets: the AMI corpus (Carletta, 2006), the LIGHT corpus (Urbanek et al., 2019), the PERSUASION corpus (Wang et al., 2019) and SWITCHBOARD (Godfrey et al., 1992). For each of these datasets, documents for about 15K tokens were annotated in 2021 for development according to (an extended version of) the ARRAU annotation scheme, and about the same number of tokens were annotated for testing. An additional 15K of data were annotated in 2022 to create new test sets for CODI-CRAC 2022, and the 2021 development sets became training data.

The annotation effort involved in the creation of these datasets led to the rethinking of several aspects of the ARRAU annotation scheme and, more in general, of the handling of anaphora in dialogue within Universal Anaphora. Aspects of particular focus were the treatment of first and second person pronouns, and more in general of deictic reference; and the treatment of referring expressions involved in grounding (see (Poesio et al., 2024) for some details). Also, the abundance of discontinuous markables in such corpora led to extending the original UA scorer to handle such markables.

Some basic statistics about the CODI-CRAC dataset are provided in Table 1. For each dataset, the Table reports number of documents, size in tokens, number of markables, and how many of these are Discourse Old (Identity Coreference) anaphors (DO), bridging references, and discourse deixis. With a total of 214,625 tokens and 60,993 markables, the CODI-CRAC dataset is to our knowledge the largest dataset annotated for anaphoric interpretation in dialogue in English. It is also one of the largest datasets annotated for bridging references.

The AMI, LIGHT and PERSUASION subsets are freely available from the Shared Task Codalab site and from the Universal Anaphora Github. SWITCHBOARD is distributed by LDC, like the copyrighted subsets of ARRAU.[9]

6. The CorefUD collection

The CorefUD initiative (Nedoluzhko et al., 2022) was launched in parallel with UA to build a collection of corpora annotated with coreferential and other anaphoric relations using a harmonized schema and format. Its current version CorefUD 1.1 (Novák et al., 2023) consists of 17 datasets for 12 languages in its publicly available edition, plus 4 more datasets with non-public licences. See Table 2 for the data sizes.

As its name suggests, CorefUD is inspired by the Universal Dependencies (UD) project. Similarly to UD, the aim is to continuously extend the collection with new datasets and languages, which can be directly utilized for training and testing automatic resolution systems. While the main focus is on harmonizing identity coreference, driven primarily by the shared task co-organized by the CorefUD authors (Section 8.2), the collection contains also other anaphoric relations and phenomena related to anaphora.

[9] ARRAU is also freely available to any group that purchased the Penn Treebank and TRAINS-93 corpora from LDC.

		documents	tokens	entities	markables	discourse old	predications	expletives	bridging	discourse deixis
LIGHT	train	20	11495	1804	3907	2132	143	74	381	72
	dev	21	11824	1790	3941	2181	147	62	424	84
	test	38	22017	3596	7330	3770	234	156	812	128
AMI	train	7	33741	4396	8918	4579	327	243	853	230
	dev	3	18260	2552	4870	2350	144	143	638	118
	test	3	16562	2004	3990	2007	151	95	432	118
PERSUASION	train	21	9185	1513	2743	1242	121	68	248	95
	dev	27	12198	1996	3697	1715	142	105	316	133
	test	33	14719	2142	4233	2111	134	81	304	105
SWITCHBOARD	train	11	14992	2362	4024	1679	139	138	589	128
	dev	22	35027	5438	9392	3991	323	378	1165	265
	test	12	14605	2296	3888	1606	143	172	464	107
Total		218	214625	31889	60933	29363	2148	1715	6626	1583

Table 1: Statistics about the CODI-CRAC 2022 corpus

Another relation to the UD project is COREFUD's strict compatibility with the CONLL-U format. It implies that the COREFUD collection also includes UD-like morphosyntactic annotation—either manual, if available in the original sources, or generated using the UDPipe parser (Straka, 2018). With regard to coreference and anaphora, COREFUD 1.0 can encode[10] essentially the same information as CONLL-UA, but this information is packed in the `MISC` column, which makes it possible to pass level 2 of the official UD validation.[11] One remaining difference is that COREFUD, being from its very beginning designed to represent existing data including dependency syntax, can capture zero expressions by stipulating 'empty tokens' and referencing them using enhanced dependency graphs. In contrast, CONLL-UA does not require dependency layers and binds empty tokens to the surface tokens by their relative position.

Combining morphosyntactic and anaphoric annotation is motivated not only pragmatically (popularity of UD and standards for numerous technical issues), but it is also grounded theoretically. For instance, entity mentions often correspond to syntactically relevant notions (e.g. noun phrase, subject), some coreference relations are manifested mainly by syntactic means (e.g. reflexive and relative constructions), and zero expressions (e.g. pro-drops) are vital for coreference in many languages.

The COREFUD collection is accompanied with API implemented within the Udapi framework[12] (Popel et al., 2017) that facilitates manipulation with the data in COREFUD format as well as its visualization.

7. The Universal Anaphora Scorer

The Universal Anaphora (UA) scorer is a Python scorer for the varieties of anaphoric reference in the scope of the Universal Anaphora catalogue. The scorer builds on the original Reference Coreference scorer [13] (Pradhan et al., 2014) developed for use in the CONLL 2011 and 2012 shared tasks on the ONTONOTES corpus (Pradhan et al., 2012) and its reimplementation in Python by Moosavi,[14] developed for the CRAC 2018 shared task (Poesio et al., 2018). The first version of the scorer (Yu et al., 2022b), used in the CODI-CRAC shared tasks (Khosla et al., 2021; Yu et al., 2022a), covered identity reference, split antecedent plurals, identification of non-referring expressions, bridging reference, and discourse deixis. This version

[10] The file format used since COREFUD 1.0 (Nedoluzhko et al., 2022) is described at https://ufal.mff.cuni.cz/~popel/corefud-1.0/corefud-1.0-format.pdf. Previous versions of COREFUD used a different format.

[11] https://universaldependencies.org/validation-rules.html#levels-of-validity. Passing the higher levels is not possible with automatically predicted POS tags and dependency relations.

[12] https://github.com/udapi/udapi-python

[13] https://github.com/conll/reference-coreference-scorers

[14] https://github.com/ns-moosavi/coval

CorefUD dataset	documents	tokens (k)	entities (k)	markables (k)	singletons (k)	appositions	predications	split antecedents	discourse deixis	bridging
Catalan-AnCora	1,298	429	18	62	0	✓	✓	✓	✓	✗
Czech-PCEDT	2,312	1,156	49	168	3	(✓)	(✓)	✓	✓	✓
Czech-PDT	3,165	835	47	155	32	(✓)	(✓)	✓	✓	✓
English-GUM	195	187	7	32	20	✓	✓	✓	✓	✓
English-ParCorFull	19	11	0	1	0	✓	(✓)	✓	✓	✗
French-Democrat	126	285	7	46	32	✗	✗	✗	✗	✗
German-ParCorFull	19	11	0	1	0	✓	(✓)	✓	✓	✗
German-PotsdamCC	176	33	1	3	3	✓	✓	✗	✓	✗
Hungarian-KorKor	94	25	1	4	0	?	?	✗	?	✗
Hungarian-SzegedKoref	400	124	5	15	0	✓	?	✗	✓	✓
Lithuanian-LCC	100	37	1	4	0	✗	✗	✓	✗	✗
Norwegian-BokmaalNARC	346	246	6	27	48	?	?	✓	?	✓
Norwegian-NynorskNARC	394	207	5	22	40	?	?	✓	?	✓
Polish-PCC	1,828	539	22	83	106	✓	✓	✗	✓	✓
Russian-RuCor	181	157	4	16	0	✓	✓	✗	✗	✗
Spanish-AnCora	1,356	458	19	71	1	✓	✓	✓	✓	✗
Turkish-ITCC	24	55	1	4	0	?	?	✗	?	✗
Dutch-COREA	844	140	3	9	25	✓	✓	✗	✓	✓
English-ARRAU (ARRAU 1)	413	229	8	32	40	✓	✓	✓	✓	✓
English-OntoNotes	3,493	1,632	51	209	0	✓	✗	✗	✓	✗
English-PCEDT	2,312	1,174	39	139	15	(✓)	(✓)	✓	✓	✗

Table 2: CorefUD 1.1 statistics. The left part shows the number of documents, words, entities excluding singletons, mentions (markables) excluding singletons, and singletons. All the numbers except for documents are reported in thousands. The right part shows which types of relations among mentions are present in the data (in addition to identity). Brackets around the check sign mean that this kind of information has not been completed manually within the annotation of coreference-related phenomena, but it can be obtained from other annotation layers (mostly, from the syntactic annotation). The 4 datasets in the bottom part are not released publicly because their licences do not allow redistribution.

was extended to include handling of discontinuous markables in the CorefUD 1.0 scorer, developed for the CRAC 2022 shared task. The Universal Anaphora 2.0 scorer (Yu et al., 2023) merges the two versions, and adds a scoring mechanism for zero anaphors encoded using special symbols in the markup layer, as done, e.g., in OntoNotes.

7.1. Identity Reference

The scorer computes all major metrics for identity reference including MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF (Luo, 2005), CoNLL (the unweighted average of MUC, B^3, and CEAF) (Pradhan et al., 2014), BLANC (Luo et al., 2014; Recasens and Hovy, 2011), and LEA (Moosavi and Strube, 2016) scores. The scorer preserves the settings used in the Reference Coreference scorer, and its scores are consistent with those of that scorer. Three scores reporting options are available: The first option mirrors the evaluation used in the CoNLL shared tasks (Pradhan et al., 2012) which excludes singletons and split-antecedents from evaluation. The second option was used in the identity anaphora sub-task of the CRAC shared task (Poesio et al., 2018). This evaluation includes singletons, but not split-antecedents. Finally, the scorer can include both singletons and split-antecedent anaphors, as done in CODI-CRAC (Khosla et al., 2021; Yu et al., 2022a).

7.2. Split Antecedent Anaphora

The evaluation metrics for split antecedent anaphora proposed in previous work (e.g. Vala et al. (2016); Zhou and Choi (2018)) are not entirely satisfactory. The UA scorer implements a new method for scoring split-antecedent anaphora proposed by Paun et al. (2023), based on the idea

of treating the antecedents of split-antecedent anaphors as a new type of mention, **accommodated sets**–set denoting entities which have the split antecedents as elements. So for instance, in example (8), split-antecedent anaphor $[They]_{1,2}$ is encoded as belonging to a coreference chain whose first element is the accommodated set {1,2} with the coreference chains for *John* and *Mary* as elements. This extension to include accommodated entities could be potentially used to handle other types of anaphoric reference using accommodation (Beaver and Zeevat, 2007), such as context change accommodation (Webber and Baldwin, 1992; Fang et al., 2021, 2022). See Paun et al. (2023) for details.

7.3. Non-referring expressions

Non-referring expressions are not treated as singletons. Instead, non-referring expressions are separated from identity references when inputted to the scorer, using the `SemType` attribute. The scorer can then compute an F1 score for non-referring expressions only. The F1 score for non-referring expression is reported separately from the F1 scores for identity reference.

7.4. Discourse Deixis

The UA scorer supports the extension to discourse deixis proposed in version 1.0 of the Universal Anaphora specification by implementing an entirely new approach to evaluation of discourse deixis supporting the evaluation.

The implementation is based on the observation that discourse deixis is similar to coreference, in that both form clusters by linking the anaphors to their antecedents. Another important similarity is that in both cases we can have split-antecedent anaphors that refer to multiple antecedents–in fact, split antecedent reference is the norm for discourse deixis. The main difference is that, in coreference, antecedents are introduced using nominal phrases, whereas in discourse deixis they are introduced using non-nominal phrases (segments).

In CoNLL-UA, discourse deixis is specified in the separate `Discourse_Deixis` layer, but using the exact same attributes as the `Discourse_deixis` column of the 'exploded' format, and the same attributes are used as for the `Identity` column.

This representation enables the application of coreference metrics to evaluate discourse deixis–and given that our new scorer provides a way to incorporate split-antecedents into the standard metrics, split antecedent discourse deixis can be handled as well. This is exactly how the UA scorer evaluates discourse deixis: it computes the same MUC, B^3, CEAF, CoNLL, BLANC and LEA metrics as for identity anaphora.

7.5. Bridging References

For bridging references, the scorer follows the approach introduced by (Hou et al., 2018). It reports three scores: the two metrics computed by the scorer used for CRAC 2018 shared task–mention-based F1 and entity-based F1–and, in addition, anaphora recognition F1. Mention-based F1 for bridging evaluates a system's ability to predict the correct anaphora and the mention of the anchor specified in the annotation. Entity-based F1 is more relaxed than mention-based F1, and does not require the system to predict exactly the same mention as the gold annotation. Instead, a system's interpretation is deemed correct as long as any mention of the correct anchor (`EntityAnchor`) is found, as done e.g., in Poesio et al. (2018). Finally, anaphora recognition F1 is used to assess the system's ability to identify bridging anaphors.

7.6. Discontinuous Markables

In CoNLL-UA, discontinuous markables can be used in both the `Identity` and `Discourse_Deixis` columns by sharing the `MarkableID` between the different sub-spans of a discontinuous markable. The scorer can then recognise the discontinuous markables from the text. For example, if a discontinuous markable consists of two continuous spans, the two spans will have the same `Identity` column, e.g. same `EntityID`, `MarkableID`, `Min` and `SemType`.

The CorefUD format does not assign IDs to markables. Instead, each continuous part of a discontinuous markable is labeled by its ordinal number and the total number of parts in square brackets just after the cluster ID: `Entity=(10[1/2]` ... `Entity=10[1/2])` ... `Entity=(10[2/2]` ... `Entity=10[2/2])`.

7.7. Strict and partial matching

The scorer provides two markable alignment strategies during the evaluation: 'strict' and 'partial'. In a 'strict' setting markables are aligned only if all parts of the discontinuous markables are recognised correctly by the system. In the 'partial' setting, markables can be aligned using a specified fuzzy matching algorithm. To use the 'partial' matching, the `Min/head` span for each markable needs to be specified in the key files.

7.8. Zero Anaphora

In both CoNLL-UA and CorefUD format, zeros are represented using the UD standard of empty nodes, in which the first column (`ID`, word index) is indicated using the decimal numbers. For instance, if we have a zero anaphora right after a token whose `ID` is 5, we index the zero with 5.1 instead of 6

used for a normal token. The scorer identifies the zeros by the decimal indexing and has the option to include zeros in the evaluation.

When zeros are included in the evaluation, again we need to align them between the key and response. Currently, alignment is based on the position of the zeros– i.e. zeros are aligned if they are located in the same position in the sentences.

7.9. Formats

The scorer supports three formats: CoNLL 2012, CoNLL-UA and CorefUD.

8. Shared Tasks

In the years since the launch of the initiative, two shared tasks using data annotated according to CoNLL-UA have been run as a collaboration between the CODI and CRAC series of workshops in 2021 and 2022, and an additional one using data from the CorefUD repository was run in collaboration with CODI 2023. All these shared tasks used versions of the Universal Anaphora scorer. In this Section we briefly discuss these shared tasks and the datasets employed in, or produced for, them.

8.1. The Shared Tasks on Anaphora Resolution in Dialogue

CODI-CRAC 2021 and 2022 consisted of three tasks covering identity anaphora, bridging anaphora, and discourse deixis. The Universal Anaphora scorer was used for all the tasks. In 2021, a total of 55 individual participants registered for the CODI-CRAC shared task on CodaLab. Five teams submitted results for Task 1, three submitted results for Task 2, and two submitted results for Task 3. In 2022, five teams submitted a total of 36 runs to the official scoreboard.

8.2. The Shared Tasks on Multilingual Coreference Resolution

The first edition of The Shared Task on Multilingual Coreference Resolution was organized in association with the CRAC workshop in 2022. Shared task participants were supposed to both (a) identify mentions in texts and (b) predict which mentions belong to the same coreference cluster (i.e., refer to the same entity or event). The public version of CorefUD version 1.0 (Nedoluzhko et al., 2022), i.e. 13 datasets for 10 languages (as described in Section 6), was used as the source of the training, development, and evaluation data; evaluation data was published only without gold coreference annotations. Five systems competed in the shared task in 2022. The winner system (Straka and Straková, 2022) outperformed the baseline by 12 percentage points, in terms of the primary metrics averaged across all datasets. More information about the participating systems and their results can be found in (Žabokrtský et al., 2022).

The second edition of the shared was organized in 2023 and held with the CRAC workshop again, and had a very similar scheme. The most important differences were as follows: (1) this shared task edition made use of CorefUD v. 1.1(Novák et al., 2023) with 17 datasets for 12 languages, and (2) the head-matching score was used instead of partial matching. Seven systems competed in the shared task in 2023, with the same team delivering the winner system. See (Žabokrtský et al., 2022) for a summary of the findings of the second edition.

9. Conclusion and Future Work

Phase 1 of the Universal Anaphora initiative has achieved most of its initial goals, including the development of markup formats suitable to encode the anaphoric phenomena in its coverage in multiple languages, and of a scorer that can be used to evaluate models carrying out more complex forms of anaphoric interpretation.

UA 2.0 has now begun, with no less ambitious objectives. A first objective is to further develop the markup in order to cover more aspects of anaphoric interpretation, such as ambiguity and quasi-coreference (Poesio and Artstein, 2005; Recasens et al., 2011; Poesio et al., 2013); further specify the methods for marking in deictic reference in visual contexts, building e.g. on the proposals in (Loáiciga et al., 2022); and several as yet poorly understood aspects of anaphoric reference in dialogue. A second, and much harder, objective is to start attempting developing common guidelines, as done in UD, ideally in collaboration with the linguistic community. Finally, and as importantly, we hope to expand our community to include more researchers, from the computational and the linguistic fields.

Acknowledgements

The work of Juntao Yu and Massimo Poesio was funded by the ARCIDUCA project, UK EPSRC grant EP/W001632/1. The work of Michal Novák, Martin Popel and Daniel Zeman was funded by the Grant 20-16819X (LUSyD) of the Czech Science Foundation (GAČR); LM2023062 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic. The work of Anna Nedoluzhko was funded by the Grant 24-11132S (HVar) of the Czech Science Foundation (GAČR).

10. Limitations

The nature of this initiative is inherently incremental, along several dimensions. A number of aspects of anaphoric reference are still not captured by the proposals made so far. Also, there still are serious discrepancies among the annotation guidelines used in the different corpora, that we hope to address in the next phase of the initiative.

11. Bibliographical References

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. GENTLE: A genre-diverse multilayer challenge set for English NLP and linguistic evaluation. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 166–178, Toronto, Canada. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation (LREC) - Workshop on linguistics coreference*, volume 1, pages 563–566. ACL.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proc. of LREC*. European Language Resources Association (ELRA), Association for Computational Linguistics (ACL).

David Beaver and Henk Zeevat. 2007. Accommodation. In G. Ramchand and C. Reiss, editors, *The Handbook of Linguistic Interfaces*, pages 503–536. Oxford.

Jean Carletta. 2006. Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5.

Herbert H. Clark. 1977. Bridging. In P. N. Johnson-Laird and P.C. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press, London and New York.

Kevin Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A. Baumgartner Jr., Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E. Hunter. 2017. Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. *BMC Bioinformatics*, 18(372).

Carola Eschenbach, Christopher Habel, Michael Herweg, and Klaus Rehkämper. 1989. Remarks on plural anaphora. In *Proceedings of the fourth conference on European chapter of the Association for Computational Linguistics*, pages 161–167. Association for Computational Linguistics.

Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. 2022. What does it take to bake a cake? the RecipeRef corpus and anaphora resolution in procedural text. In *Findings of the Association for Computational Linguistics: ACL 2022*, page 3481–3495. Association for Computational Linguistics.

Biaoyan Fang, Christian Druckenbrodt, Saber A Akhondi, Jiayuan He, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU-Ref: A corpus for modeling anaphora resolution in the chemical domain. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, page 1362–1375. Association for Computational Linguistics.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: telephone speech corpus for research and development . acoustics,. In *IEEE International Conference on Speech, and Signal Processing, ICASSP-92*, volume 1, pages 517–520.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague dependency treebank - consolidated 1.0. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 5208–5218. European Language Resources Association.

John A. Hawkins. 1978. *Definiteness and Indefiniteness*. Croom Helm, London.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

Ryu Iida, Mamoru Komachi, Naoya Inoue, Kentaro Inui, and Yuiji Matsumoto. 2017. NAIST text corpus: Annotating predicate-argument and coreference relations in Japanese. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, volume 2, pages 1177–1196. Springer.

International Organization for Standardization. 2019. *Language Resource Management - Semantic Annotation Framework - Part 9, Reference Annotation Framework (RAF)*, ISO 24617-9:2019 edition. International Organization for Standardization, Vernier, Geneva, Switzerland.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. D. Reidel, Dordrecht.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The CODI-CRAC 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proc. of the CODI/CRAC Shared Task Workshop*.

Hideo Kobayashi and Vincent Ng. 2021. Bridging resolution: Making sense of the state of the art. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1652–1659, Online. Association for Computational Linguistics.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Anaphora with non-nominal antecedents in computational linguistics: a Survey. *Computational Linguistics*, 44(3):547–612.

Quan Liu, Hui Jiang, Andrew Evdokimov, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2017. Cause-effect knowledge acquisition and neural association model for solving a set of winograd schema problems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2344–2350.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2022. Anaphoric phenomena in situated dialog: A first round of annotations. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 31–37.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. 2014. An extension of BLANC to system mentions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29, Baltimore, Maryland. Association for Computational Linguistics.

Ana Marasović, Leo Born, Juri Opitz, and Anette Frank. 2017. A mention-ranking model for abstract anaphora resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 221–232, Copenhagen, Denmark. Association for Computational Linguistics.

Nafise S. Moosavi and Michael Strube. 2016. A proposal for a link-based entity aware metric. In *Proc. of ACL*, pages 632–642, Berlin.

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. 2014. ANCOR_Centre, a large free spoken French coreference corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

Anna Nedoluzhko. 2013. Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 103–111.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC*, Portoroz. ELRA.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. Corefud 1.0: Coreference meets universal dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, page 4859–4872. European Language Resources Association.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2021. Coreference meets universal dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages. ÚFAL Technical Report TR-2021-66, Charles University, Prague.

Michal Novák, Martin Popel, Zdeněk Žabokrtský, Daniel Zeman, Anna Nedoluzhko, Kutay Acar, Peter Bourgonje, Silvie Cinková, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, Petter Mæhlum, M. Antònia Martí, Marie Mikulová, Anders Nøklestad,

Maciej Ogrodniczuk, Lilja Øvrelid, Tuğba Pamay Arslan, Marta Recasens, Per Erik Solberg, Manfred Stede, Milan Straka, Svetlana Toldova, Noémi Vadász, Erik Velldal, Veronika Vincze, Amir Zeldes, and Voldemaras Žitkus. 2023. Coreference in universal dependencies 1.1 (CorefUD 1.1). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Tim O'Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 2018. Amr beyond the sentence: the multi-sentence amr corpus. In *Proc. of COLING*, pages 3693–3702, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Silviu Paun, Juntao Yu, Nafise Moosavi, and Massimo Poesio. 2023. Scoring coreference chains with split-antecedent anaphors and other entities constructed from a discourse model. *Dialogue and Discourse*, 14(2).

Massimo Poesio. 2016. Linguistic and cognitive evidence about anaphora. In M. Poesio, R. Stuckardt, and Y. Versley, editors, *Anaphora Resolution: Algorithms, Resources and Applications*, chapter 2. Springer.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

Massimo Poesio, Maris Camilleri, Paloma Carretero Garcia, Juntao Yu, and Mark-Christoph Müller. 2024. The ARRAU 3.0 corpus. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI)*, pages 127–138. Association for Computational Linguistics.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Silviu Paun, Alexandra Uma, and Juntao Yu. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proc. of NAACL*, page 1778–1789, Minneapolis. Association for Computational Linguistics (ACL).

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Intelligent Interactive Systems*, 3(1).

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana. Association for Computational Linguistics.

Massimo Poesio, Sameer Pradhan, Marta Recasens, Kepa Rodriguez, and Yannick Versley. 2016. Annotated corpora and annotation tools. In M. Poesio, R. Stuckardt, and Y. Versley, editors, *Anaphora Resolution: Algorithms, Resources and Applications*, chapter 4. Springer.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press, New York.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Marta Recasens and Ed Hovy. 2011. BLANC: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Marta Recasens, Ed Hovy, and M. Antonia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Marta Recasens and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.

Kepa-Joseba Rodriguez, Francesca Delogu, Yannick Versley, Egon Stemle, and Massimo Poesio. 2010. Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC*. European Language Resources Association (ELRA).

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8732–8740.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2022. ÚFAL CorPipe at CRAC 2022: Effectivity of Multilingual Models for Coreference Resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*. Association for Computational Linguistics.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, Samuel Humeau, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. ArXiv preprint arXiv:1903.03094.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Journal of Natural Language Engineering*, 26(1).

Hardik Vala, Andrew Piper, and Derek Ruths. 2016. The more antecedents, the merrier: Resolving multi-antecedent anaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2287–2296, Berlin, Germany. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Bonnie L. Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Bonnie Lynn Webber and Breck Baldwin. 1992. Accommodating context change. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 96–103, Newark, Delaware, USA. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. *OntoNotes: A Large Training Corpus for Enhanced Processing*.

Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, and Massimo Poesio. 2022a. The CODI/CRAC 2022 shared task on anaphora resolution, bridging and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*.

Juntao Yu, Sopan Khosla, Nafise Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. 2022b. The Universal Anaphora scorer 1.0. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2020. Free the plural: Unrestricted split-antecedent anaphora resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages

6113–6125, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2021. Stay together: A system for single and split-antecedent anaphora resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Juntao Yu, Michal Novák, Abdulrahman Aloraini, Nafise Sadat Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. 2023. The Universal Anaphora scorer 2.0. In *Proc. of the International Workshop on Computational Semantics (IWCS)*.

Juntao Yu and Massimo Poesio. 2020. Multitask learning based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3534–3546, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Annie Zaenen. 2006. Mark-up barking up the wrong tree. *Computational Linguistics*, 32(4):577–580.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes. 2022. Can we fix the scope for coreference? *Dialogue and Discourse*, 13(1):41–62.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.