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,3 Martin Popel,9 Zdeněk Žabokrtský,9 and Daniel Zeman9

1Queen Mary University, UK; 2University of Utrecht, The Netherlands; 3Institute of Computer Science, Polish Academy of Sciences; 4University of Texas at Dallas; 5LDC, University of Pennsylvania, USA; 6University of Sheffield, UK; 7Amazon; 8Georgetown University, USA; 9Charles University, Czechia; m.poesio@qmul.ac.uk; maciej.ogrodniczuk@ipipan.waw.pl; vince@hit.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 dataset1 (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 ANCORAl-2 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

---

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

2http://www.universalanaphora.org

3https://universaldependencies.org/

4We should also mention that an ISO standard for reference exists (International Organization for Stan-
The proponents of UA were very much aware of its goals and proposals has been published yet except the original online document from December 2020. 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 CoNerUD (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.

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. 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], are most productive when [they], are held in the morning. [Those meetings], 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]. The lens’ foldability enables it to be inserted in smaller incisions than are now possible for [cataract surgery].

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).

5 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], 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], 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]
(5) u2 S yeah
(6) 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 void 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, è in ritardo, così [it], mi ha chiesto se posso incontrarlo al cinema. [EN] John, is late so [he], 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 (Rodríguez 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 (Semantic Type) specifying whether a nominal phrase is referring, predicative, or an expletive. Such an 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], bought [a new dress], but [it], didn’t fit [her].

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.

---

⁸https://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-U 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, ud collaborated with COREFUD to design a more ‘compact’ format that could be used to pack the anaphoric information representable in CoNLL-U in the ‘MISC’ column of the CoNLL-U format, and is fully compatible with the Universal Dependencies. The two formats are mutually interchangeable, and the ud 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-U 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-U 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 ud 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**.

6. The COREFUD collection

The COREFUD initiative (Nedoluzhko et al., 2022) was launched in parallel with ud 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.

---

ARRAU is also freely available to any group that purchased the Penn Treebank and TRAINS-93 corpora from LDC.
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 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. 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 (Popel et al., 2017) that facilitates manipulation with the data in COREFUD format as well as its visualization.

Table 1: Statistics about the CODI-CRAC 2022 corpus

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

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 (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, 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

---

Footnotes:

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
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 score-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, accommo-
dated 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 coref-
rence, antecedents are introduced using nominal phrases, whereas in discourse deixis they are in-
roduced 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 met-
rics, split antecedent discourse deixis can be handled as well. This is exactly how the UA scorer eval-
uates discourse deixis: it computes the same MUC, B3, 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 re-
ports 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. In-
stead, a system’s interpretation is deemed correct as long as any mention of the correct anchor
(EntityAnchor) is found, as done e.g., in Poe-
sio 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 dis-
continuous markable is labeled by its ordinal number and the total number of parts in square brack-
et 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 marked alignment strategies during the evaluation: ‘strict’ and ‘par-
tial’. 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 speci-
fied 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 (Zabokrtský 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 version 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 (Zabokrtský 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 24-1132S (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


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.


Michal Novák, Martin Popel, Zdeněk Žabokrtský, Daniel Zeman, Anna Nedoluzhko, Kutay Acar, Peter Bourgonje, Silvie Cinková, Gülşen Cebiroğlu Eryiğit, Jan Hajíc, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Käsn, Pauline Kriélke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, Petter Mæhlum, M. Antonia Martí, Marie Mikulová, Anders Nøklestad,
Massimo Poesio, Lilja Øvrelid, Tuğba Pa-
may Arslan, Marta Recasens, Per Erik Sol-
tal library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University.

Tim O’Gorman, Michael Regan, Kira Griffitt, Ulf Hermjacob, Kevin Knight, and Martha Palmer. 2018. Amr beyond the sentence: the multi-

Silviu Paun, Juntao Yu, Nafise Moosavi, and Mas-
simo Poesio. 2023. Scoring coreference chains with split-antecedent anaphors and other enti-
ties constructed from a discourse model. Dia-
logue and Discourse, 14(2).

Massimo Poesio. 2016. Linguistic and cogni-

Massimo Poesio and Ron Artstein. 2005. The reli-

Massimo Poesio, Maris Camilleri, Paloma Car-
retero Garcia, Juntao Yu, and Mark-Christoph Müller. 2024. The ARRAU 3.0 corpus. In Pro-
cedings of the 5th Workshop on Computational Approaches to Discourse (CODI ), pages 127–

Massimo Poesio, Jon Chamberlain, Udo Kr-
uschwitz, Silviu Paun, Alexandra Uma, and Jun-
tao Yu. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric inter-
pretation. In Proc. of NAACL, page 1778–1789, Minneapolis. Association for Computational Lin-
guistics (ACL).

Massimo Poesio, Jon Chamberlain, Udo Kr-
uschwitz, Livio Robaldo, and Luca D Buccesi. 2013. Phrase detectives: Utilizing collective in-
telligence for internet-scale language resource creation. ACM Transactions on Intelligent Inter-
active Systems, 3(1).

Massimo Poesio, Yulia Grishina, Varada Kol-
hatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olgta Uryupina, Juntao Yu, and Helke Zimsmeis-
tics.

Massimo Poesio, Sameer Pradhan, Marta Rec-
sources and Applications, chapter 4. Springer.

Martin Popel, Zdeněk Žabokrtský, and Martin Vo-
ida 2017 Workshop on Universal Dependencies (UDW 2017), pages 96–101, Gothenburg, Swe-
eden. Association for Computational Linguistics.

Sameer Pradhan, Xiaoliang Luo, Marta Rec-
casens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementa-
tion. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 30–35, Balti-
more, Maryland. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nian-
enteenth Conference on Computational Natural Language Learning, pages 143–152, Sofia, Bul-
garia. 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 Pro-
cedings of the Sixteenth Conference on Com-
putational Natural Language Learning (CoNLL 
2012), Jeju, Korea.

Ellen F. Prince. 1981. Toward a taxonomy of given-
new information. In P. Cole, editor, Radical Prag-

tational Natural Language Learning, pages 777–
789, Jeju Island, Korea. Association for Compu-
tational Linguistics.


