The ARRAU 3.0 Corpus

Massimo Poesio[1], **Maris Camilleri**[1], **Paloma Carretero Garcia**[1],
Juntao Yu[1], **Mark-Christoph Müller**[2]
[1]Queen Mary Univ., UK [2]Leibniz-Institut für Deutsche Sprache
{m.poesio,m.camilleri,p.carreterogarcia,juntao.yu}@qmul.ac.uk
mark-christoph.mueller@ids-mannheim.de

Abstract

ARRAU is an anaphorically annotated corpus designed to cover a variety of aspects of anaphoric reference in a variety of genres, including both written text and spoken language. The objective of this annotation project is to push forward the state of the art in anaphoric annotation, by overcoming the limitations of current annotation practice and the scope of current models of anaphoric interpretation, which in turn may reveal other issues. The resulting corpus is still therefore very much a work in progress almost twenty years after the project started. In this paper, we discuss the issues identified with the coding scheme used for the previous release, ARRAU 2, and through the use of this corpus for three shared tasks; the proposed solutions to these issues; and the resulting corpus, ARRAU 3.

1 Introduction

Although the scope and ambition of anaphoric annotation projects has enormously increased in the last twenty years (Poesio, 2004; Hinrichs et al., 2004; Pradhan et al., 2007, 2012; Poesio and Artstein, 2008; Uryupina et al., 2020; Recasens and Martí, 2010; Rahman and Ng, 2012; Nedoluzhko, 2013; Muzerelle et al., 2014; Cohen et al., 2017; Zeldes, 2017; Webster et al., 2018; Bamman et al., 2020; Sakaguchi et al., 2020; Khosla et al., 2021; Yu et al., 2022a; Nedoluzhko et al., 2022) a number of open questions about anaphoric annotation remain, and many if not most of the existing corpora have limitations either in size or coverage.

The ARRAU annotation (Poesio and Artstein, 2008; Uryupina et al., 2020; Poesio et al., 2018) is a long-term project to expand the range of anaphoric annotation by creating an anaphorically annotated corpus covering a wide variety of aspects of anaphoric reference (Poesio, 2016). The annotation project started in 2004 as the result of a series of studies of the reliability of 'difficult' aspects of anaphoric annotation (Poesio, 2004; Poesio and Artstein, 2005b,a; Artstein and Poesio, 2006, 2008) and the first release was primarily focused on anaphoric reference in dialogue (Poesio and Artstein, 2008). The scope of the annotation then broadened both in terms of linguistic aspects that were annotated and in terms of genres, resulting in a second release in 2013 (Uryupina et al., 2020). This second release was then used as the core dataset for the 2018 CRAC Shared Task (Poesio et al., 2018), the first shared task for anaphora resolution covering also identification of non-referring expressions, bridging reference and discourse deixis; and as additional material for the 2021 and 2022 CODI-CRAC shared tasks on anaphora resolution in dialogues (Khosla et al., 2021; Yu et al., 2022a). These shared tasks highlighted the need to revise the annotation guidelines for a range of phenomena including discourse deixis and genericity and reference in dialogues. They also revealed a number of issues with tokenization and markup. We therefore started an extensive reannotation and cleaning up, resulting in a third, substantially revised release of the corpus.

In this paper we discuss the issues identified with the previous annotation, the revised annotation scheme and guidelines, the cleaning up procedure, and the new corpus resulting from this effort.

2 Anaphoric Annotation

We review in this Section the aspects of anaphoric interpretation captured in the ARRAU annotation.

Identity Anaphora Most modern anaphoric annotation projects cover identity anaphora as in (1).

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

However, many other types of identity anaphora exist, as well as other types of anaphoric relations, discussed below.

Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), pages 127–138

Split-antecedent anaphora In most corpora, plural reference is only marked when the antecedent is mentioned by a single noun phrase. But in **split-antecedent anaphors** (Eschenbach et al., 1989; Kamp and Reyle, 1993) such as (2), plural pronoun *they* refers to a set composed of two entities introduced by separate noun phrases.

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

Such references are not annotated in many corpora, or *They* is treated as a bridging reference.

The semantic function of noun phrases The nominal expressions in (1) are examples of **referring** noun phrases, which either introduce new entities in a discourse (first mention of Mary and the new dress) or link to previously introduced entities (pronouns *it* and *her*). But NPs can serve different functions. **Quantificational** NPs such as *No one* in *No one would put the blame on him/herself* (Partee, 1972) do not refer to an individual or set of individuals, but can still participate in anaphoric relations even though anaphoric reference to quantifiers has distinctive properties (Partee, 1972) and is subject to semantic constraints (Karttunen, 1976). **Predicative** noun phrases express properties of objects: for instance, in sentence (3), the NP *a busy place* does not introduce a new discourse entity or refer back to an existing discourse entity, but expresses a property. Finally, in languages like English, forms like *it* and *there* can also be used to express semantically vacuous **expletives** as well as pronouns, like the *it* in *It is four o'clock*. Distinguishing referring from non-referring nominals is a part of the task of interpreting anaphoric expressions which cannot be evaluated in corpora where non-referring expressions are not annotated.

(3) [This] seems to be [a busy place]

Discourse deixis The term 'anaphoric reference' covers a wide variety of phenomena, not all of which are annotated in all corpora. **Event anaphora** is the type of anaphoric reference exemplified by *that* in (4), which does not refer to an entity introduced by a nominal, but to the event of a white rabbit with pink ears running past Alice.

(4) ... 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); ....

Event anaphora is a subtype of the more complex phenomenon of **discourse deixis** (Webber, 1991; Kolhatkar et al., 2018) which also includes references like *this* in (4), which refers to the fact that the Rabbit was able to talk. Not many corpora attempt to cover the entire range of discourse deixis.

Bridging references and other non-identity anaphora Possibly the most studied type of non-identity anaphora is **bridging reference** or **associative anaphora** (Clark, 1977; Hawkins, 1978; Prince, 1981) as in (5), where bridging reference *the roof* refers to an object which is related to / associated with, but not identical to, the *hall*.

(5) 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]].

Other types of non-identity anaphora also exist, besides bridging references. Examples include *other* anaphora like *the other* in (6), as well as **identity of sense** anaphora such as *a blue one* in (7) (Poesio, 2016).

(6) John gave one book to Mary, and [the other] to Bill.

(7) John bought a red ball, and Mary [a blue one].

The interplay between anaphora and other semantic properties of nominals Often, whether two mentions corefer depends on how they get semantically interpreted in other respects. In (8), for instance, whether the mention *bananas* in 40.2 is interpreted as coreferring with mention *bananas* in 37.8 depends on whether these bare plurals are taken to be references to the generic kind **bananas** (Carlson and Pelletier, 1995). If those mentions are interpreted as non-generic, they would not corefer. Some anaphoric corpora therefore include an annotation of noun phrases' genericity (Uryupina et al., 2020; Nedoluzhko, 2013).

(8)

37.1	M:	all right
37.2	:	and then at the same time ...
37.5	:	E2 was zipping over to Bath to pick up a boxcar
37.6	:	heading down to Avon
37.7	:	to
37.8	:	collect [bananas]$_i$
37.9	:	and then shipping [em]$_i$ back to Corning
37.10	:	shortest route ...
38.1	S:	okay so
38.2	:	E2
38.3	:	goes to Corning
38.4	:	then
38.5	:	on to Bath
38.6	:	and gets a boxcar
39.1	M:	m hm
40.1	S:	then on to Avon
40.2	:	load [bananas]$_?$

Anaphoric reference in dialogue Anaphora resolution in dialogue requires systems to handle grammatically incorrect language suffering from disfluencies and mentions jointly created across utterances (Poesio and Rieser, 2010) or whose function is to establish common ground rather than refer (Clark and Brennan, 1990; Heeman and Hirst, 1995). Dialogue contains more deictic reference, vaguer anaphoric and discourse deictic reference, or speaker grounding of pronouns. These complexities are normally absent from news or Wikipedia articles, which form the bulk of current datasets for coreference resolution (Poesio et al., 2016). There has been some research on coreference in dialogue in English (Byron, 2002; Eckert and Strube, 2001; Müller, 2008), but very limited in scope (primarily pronominal interpretation), due to the lack of suitable corpora, although the situation is better for other languages (Muzerelle et al., 2014; Grobol, 2020).

3 ARRAU 1 and 2

3.1 Genres

The ARRAU corpus[1] (Poesio and Artstein, 2008; Uryupina et al., 2020) was designed to cover a variety of genres. Initially, the corpus was meant to focus on anaphoric reference in dialogue and spoken language (Poesio and Artstein, 2008). Its TRAINS sub-corpus includes all the task-oriented dialogues in the TRAINS-93 corpus[2] (Heeman and Allen, 1995) already used in Byron's work on pronominal reference in dialogue (Byron and Allen, 1998; Byron, 2002) as well as the pilot dialogues in the so-called TRAINS-91 corpus. The PEAR sub-corpus consists of the complete collection of spoken narratives in the Pear Stories that provided some of the early evidence on salience and anaphoric reference (Chafe, 1980).[3] Subsequently, the corpus was extended to cover a substantial amount of written text, including news text in a sub-corpus called RST, consisting of the entire subset of the Penn Treebank (Marcus et al., 1993) that was annotated in the RST treebank (Carlson et al., 2003).[4] The GNOME sub-corpus covers documents from the medical and art history genres covered by the GNOME corpus (Poesio, 2004).

3.2 Annotation scheme

The same coding scheme was used for all subcorpora, but separate guidelines were written for the spoken dialogue and written language subcorpora. The original annotation scheme used for Release 1 (Poesio and Artstein, 2008), focused on dialogue, is distributed with the dataset and is also available from the ARRAU corpus page. For the second release (Uryupina et al., 2020), the guidelines for bridging were extended and genericity was also annotated using the GNOME guidelines, but a complete new manual was not produced. However, a fairly extensive description can be found in Uryupina et al. (2020).

Markable definition Many older anaphorically annotated corpora impose syntactic, semantic or discourse-based restrictions on markables. For instance, in ONTONOTES neither expletives nor singletons are annotated (Poesio et al., 2016). By contrast, in ARRAU *all* NPs are considered as markables, including non-referring expressions (e.g., expletives such as *it* or predicative NPs such as *a busy place*) in (3), and expressions do not corefer with any other markable ('singletons'). Moreover, in ARRAU non-referring markables are manually subclassified into expletives, predicative, and quantifiers. In addition, all generic references are marked, including premodifiers when the entity referred to is mentioned again, e.g., in the case of the proper name *US* in (9), and premodifiers that refer to a kind, like *exchange-rate* in (10).

(9) ...The Treasury Department said that the [US]$_1$ trade deficit may worsen next year

[1] http://www.arrauproject.org/corpus

[2] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC95S25

[3] https://www.linguistics.ucsb.edu/research/pear-film

[4] https://catalog.ldc.upenn.edu/LDC2002T07

after two years of significant improvement... The statement was the [US]$_1$'s government first acknowledgment ...

(10) The Treasury report, which is required annually by a provision of the 1988 trade act, again took South Korea to task for its [exchange-rate]$_1$ policies. "We believe there have continued to be indications of [exchange-rate]$_1$ manipulation ...

A distinctive feature of ARRAU's definition of markables is that, due to its initial focus on dialogue, it also allows **discontinuous** markables such as the collaborative constructed *three ... loaded boxcars* in (11), building on (Müller, 2008) and leveraging MMAX2's support for such markables.

(11)
S: okay um if you can only pull three loaded boxcars
U: [three]^1
S: yeah [loaded boxcars]^1

Referential status A markable can be marked as semantically non-referring (an expletive, a predicate, a quantifier, a coordination, an idiom, or incomplete) or referring (either `discourse new` or `discourse old`). Discourse new mentions introduce new entities and thus are not marked as being coreferent with an entity already introduced (**antecedent**). For discourse-old markables, the annotation of different types of anaphoric relations is supported. The antecedent of discourse-old mentions can be either of type `phrase` (if the antecedent was introduced using a nominal markable) or `segment` (not introduced by a nominal markable, for **discourse deixis**).[5] In addition, referring NPs can be marked as **related** to a previously mentioned discourse entity to identify them as examples of associative (**bridging**) anaphora.

Bridging references Annotating — indeed, even identifying — bridging references in a reliable way is difficult, which is one of the reasons why so few large-scale corpora for anaphora include this type of annotation (Poesio et al., 2016; Kobayashi and Ng, 2020). The ARRAU guidelines for bridging anaphora are based on experiments that ran from (Poesio and Vieira, 1998) to (Poesio, 2004). The ARRAU Release 1 and 2 guidelines followed the GNOME guidelines, but with an extension and a simplification. Annotators were asked to mark a markable as `related` to a particular antecedent if it stood to that antecedent in one of the GNOME relations or in the two additional relations

- `other`, for *other* NPs, broadly following the guidelines in Modjeska (2003);
- an `undersp-rel` relation for 'obvious cases of bridging that didn't fit any other category'.

However, the actual relations were not marked in ARRAU 1. Relation annotation started with ARRAU 2, but only for the `RST` portion. One of the objectives for ARRAU 3 was to annotate the relations underlying bridging reference for all sub-corpora.

Discourse deixis Discourse deixis in its full form is a very complex form of reference, both to annotate and to resolve (Kolhatkar et al., 2018) . Very few anaphoric annotation projects have attempted to annotate discourse deixis in its entirety (Kolhatkar et al., 2018). More typical is a partial annotation, as in (Byron and Allen, 1998; Navaretta, 2000), who annotated pronominal reference to abstract objects; in ONTONOTES, where event anaphora was marked (Pradhan et al., 2007); and in (Kolhatkar and Hirst, 2014), which focused on so-called shell nouns. In ARRAU, a coder specifying that a referring expression is discourse-old is asked whether its antecedent was introduced using a `phrase` (markable) or a `segment` (discourse segment). Coders who choose `segment` have to mark a sequence of *predefined* clauses as antecedent.

Genericity ARRAU is not a multi-layer corpus like ANCORA, GUM, ONTONOTES or the Prague Dependency Treebank, meaning that other linguistic information relevant for the study of anaphora (morphosyntax, dependency structure, semantics) also has to be annotated within the anaphoric layer. We only discuss in this paper genericity, as it's the one among these attributes for which the guidelines changed in ARRAU 3.

The ARRAU scheme and guidelines for genericity build on the studies of genericity reliability carried out as part of the GNOME annotation (Poesio et al., 2004). This scheme is based on a generalised notion of scopal dependence for nominals covering both genericity and scopal dependence on a range of operators including conditionals, quantifiers, and temporal adverbials. More specifically, according to the guidelines used for ARRAU 1 and 2, the annotation of the `generic` attribute is carried out following a decision tree going from the

[5] Identity anaphora also includes split antecedent plural anaphoric reference.

easiest cases to the more complex ones. Coders are first asked to check whether the nominal is in the syntactic scope of an *explicit* operator such as a conditional like *if* (as in (12)) or an individual quantifier such as *every* or *most* (`iquant`) In these cases, the nominal is *not* marked as generic, but as being in the scope of the appropriate operator. If no such explicit quantifier/operator is present, coders are asked to check whether the nominal refers to semantic objects whose genericity is left underspecified, such as substances (e.g., *gold*), as in (13) Finally, the annotator is asked whether the sentence in which the markable occurs is generic, and in this case, to mark the nominal as `generic-yes` if it refers generically, as in (14), or `generic-no` otherwise. With these instructions, reasonable intercoder agreement was achieved ($\kappa = .82$) (Poesio, 2004).

(12) New York State Comptroller Edward Regan predicts a $ 1.3 billion budget gap for the city ‘s next fiscal year, a gap that could grow if there is [a recession]$^{operator-conditional}$.“

(13) Not that [oil]$^{undersp-substance}$ suddenly is a sure thing again .

(14) In its report to Congress on [international economic policies]$^{generic-yes}$, the Treasury said that any improvement in the broadest measures of trade, known as the current account.

3.3 Annotation procedure

ARRAU 1 and 2 were annotated using MMAX2 (Müller and Strube, 2006). All annotation was carried out by trained (computational) linguists. ARRAU 1 was primarily annotated at the University of Essex between 2004 and 2007 under the direction of Ron Artstein, who also designed the MMAX2 style, and in collaboration with Mark-Christoph Müller. The initial annotation was then extended and checked as part of the Johns Hopkins 2007 Workshop on Entity Disambiguation (ELERFED).

ARRAU 2 was annotated at the University of Trento between 2008 and 2016 under the coordination of Kepa Rodriguez, Francesca Delogu, Federica Cavicchio, and Olga Uryupina. Most of the annotation was carried out by Antonella Bristot.

3.4 Use in shared tasks

In recent years, ARRAU was used for three shared tasks: the CRAC 2018 shared task on anaphora resolution with the ARRAU corpus (Poesio et al., 2018), and the 2021 and 2022 CODI-CRAC shared tasks on anaphora resolution in dialogue (Khosla et al., 2021; Yu et al., 2022a).

The use of the corpus for such tasks was enabled by two improvements brought about by the Universal Anaphora initiative.[6] The first of these was the development of a tabular markup format extending the CONLL-U tabular format used for the CONLL 2011 and 2012 shared tasks on coreference (Pradhan et al., 2012) with ways to represent the additional types of anaphoric information encoded in ARRAU, but consistent with it so that modellers would understand it better. And second, the development of scorers extending the Coreference Reference scorer (Pradhan et al., 2014) with ways of scoring the interpretation of these additional phenomena (Poesio et al., 2018; Yu et al., 2022b).

4 ARRAU 3: Summary of the Revisions

The CRAC 2018 shared task revealed a number of issues with the ARRAU 2 annotation - first of all with the annotation of bridging references and discourse deixis- that prompted a first round of revisions to the annotation scheme and the annotation guidelines. More issues about the annotation of anaphoric reference in dialogue were revealed when the data were used for the CODI-CRAC 2021 shared task, resulting in a second round of revisions. During the CODI-CRAC shared task we also discovered issues with tokenization and with the way the RST portion had been converted. As a result, we started revising the corpus by: (i) revising annotation schemee and guidelines (ii) fixing the issues with tokenization and with conversion. In the following two sections, we discuss each of these revisions in detail.

5 Revised Guidelines and Re-annotation

5.1 Revised annotation scheme and guidelines

The changes to the annotation scheme and guidelines between ARRAU 2 and ARRAU 3 can be summarized as follows: (i) alternative schemes especially for the more complex aspects of the annotation (e.g., bridging reference, genericity, discourse deixis) were carefully analyzed and the annotation scheme and guidelines for these aspects were (partially) revised at the light of the solutions proposed in this work; (ii) a more *semantic* approach was adopted for the annotation of certain aspects that

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

had been previously annotated following purely syntactic guidelines (e.g., predication, genericity); (iii) for the dialogue sub-corpora, more attention was paid to aspects of reference in dialogue that previously had not been sufficiently considered (e.g., deictic first and second person pronouns, or the use of referring expressions for grounding purposes).

Predicative NPs The ARRAU 2 guidelines for predicative NPs were not very explicit and essentially relied on syntactic information, marking as predicates object NPs in copular clauses (*Antonio Conte was [an Italian prime minister]*) and clauses with verbs such as *become* (*Antonio Conte became [the Italian prime minister]* as well as appositions (*Antonio Conte, [the Italian prime minister], arrived in London for talks today*).

However, the decision whether an NP is predicative cannot always be made on syntactic grounds alone (Zeldes, 2022). For instance, in *[The Italian prime minister, [Antonio Conte]], arrived in London for meetings today*, it is the NP in appositive position (*Antonio Conte*) that acts as term-denoting, whereas the outside NP has a predicative function. In so-called **specificational** copular clauses, it is the subject that is predicative, whereas the object is generally taken to be referential:

(15) [The director of Anatomy of a Murder] is Otto Preminger

Whereas in so-called **identificational** copular clauses, both the subject and object are generally taken to be referring:

(16) [That woman] is [Sylvia]

Some of these cases were covered in the previous guidelines, but not systematically. The annotation guidelines were therefore thoroughly revised, to make the decision about whether a clause is predicative depend more on semantic criteria.

Non-identity anaphora The first objective of the revision of the bridging reference annotation for ARRAU 3 was to add information about the semantic relation for all subcorpora.

Equally importantly, however, we intended to produce much more explicit guidance. One issue was highlighted by the CRAC 2018 shared task (Poesio et al., 2018). Following her participation to the shared task, in which she found that the approach proposed by Hou et al (Hou et al., 2014, 2018) for the ISNOTES corpus (Markert et al., 2012) achieved very poor results on ARRAU (Roesiger, 2018), Ina Rösiger et al carried out a detailed analysis of the difference between the annotation of bridging references in the two corpora (Roesiger et al., 2018), concluding that very different notions of 'bridging' were used. In ISNOTES, only what they called **referential** bridging references were annotated, such as *the door* in (17)–cases where the anaphoric expression contains an implicit anaphoric argument (*the door [of the house]*). (We think the term 'referential' is misleading, so we will call these bridging references **implicitly anaphoric**, or IA.) In ARRAU, in addition to implicitly anaphoric bridging references, a second category of referring expressions was also annotated as bridging references, that Rösiger et al called **lexical** bridging references. One example is *Dubrovnik* in (18): the NP is not implicitly anaphoric, but it establishes entity coherence with its anchor *Croatia* through shared knowledge. (We will call this category of bridging references **coherence-establishing**, or CE.) Rösiger et al disagreed with this broader definition of bridging reference, but also pointed out that several examples of both IA and CE bridging references were not actually annotated in ARRAU 2.

(17) John walked towards the house. [The door] was open.

(18) Croatis's tourism industry has been booming. The number of yearly visitors to [Dubrovnik] grew to over 2 million by 2019.

Following that discussion, the annotation guidelines for bridging were expanded to provide more explicit information about these types of bridging references. Explicit instructions were also added to mark split-antecedent plurals not as bridging references, but using the separate multiple antecedent mechanism offered by MMAX2. Furthermore, explicit instructions about identity of sense anaphora weree added. Further instructions were also added requiring attributes to be marked as bridging (e.g., *income* in *Kellogg reported its financial results for the year yesterday. [Income] grew to ....*).

Genericity Another issue observed while running the shared tasks was that the guidelines for genericity followed in ARRAU 1 and 2 has resulted in an excessively syntactic interpretation of scope in general and genericity in particular. Consider for instance the contrast between (19) and (20), from the TRAINS corpus. We consider instructions as

introducing an implicit modal operator, and our guidelines therefore required to annotate NPs in such utterances as `operator-instruction`. This is appropriate for both *a boxcar from Elmira* and *oranges* in (19). However, not all such NPs are in fact in the scope of the implicit modal operator–for instance, *the boxcar from Elmira* refers deictically to an entity in the visual scence (the TRAINS world map). As a result, we changed the guidelines to only annotate NPs in utterances containing implicit or explicit operators when they were actually in the *semantic* scope of the operator.

(19) take [a boxcar from Elmira]$_i$ and load [it]$_i$ with [oranges]

(20) take [the boxcar from Elmira]$_i$ and load [it]$_i$ with [oranges]

Reference in dialogue One issue with the previous guidelines that emerged in particular from the annotation for the CODI-CRAC dataset was that many aspects of reference in dialogue were not covered, or covered only in part.

The first such issue was the annotation of **first and second person pronouns**. Such pronouns were not annotated in the TRAINS sub-corpora in ARRAU 1 and 2, based on the belief that they were all deictic and referring to one or the other speaker, such as the instance of *you* in (21).

(21) S: hello how can I help [you]

However, this belief proved incorrect; first and second person pronouns are used in a number of other ways. E.g., in (22) the two instances of *you* in the first utterance are most likely interpreted *generically*–U is asking about what is possible in the task. We revised the guidelines providing directions for distinguishing between the uses.

(22) U: an [you] do can [you] do things simultaneously here or do they have to be done like can I have the same time having it the engine

Another issue that had not been sufficiently considered in previous releases was the relation between a wh-NP like *how long* in (23) and the answer to the question, *eight hours*. Clearly, this is not a case of coreference. However, even though wh-NP are annotated as quantifiers in ARRAU, it's not a case of bound anaphora either (as in *[No student]$_i$ forgot [their]$_i$ passport*). In the end, we decided to mark such cases as cases of associative references of type `element`, given that it may be argued that the wh-NP denotes a set (the set of possible answers) of which the answer is an element; but this decision may be reconsidered in the future.

(23)
U : so [how long] will it take if I take the two boxcars
. . .
S [eight hours]
U eight hours

A new manual A revised version of the annotation guidelines was produced.[7] These new guidelines were also used for the annotation of the documents included in the CODI-CRAC dataset used for the 2021 and 2022 shared tasks.

5.2 Re-Annotation

The revision proceeded in two passes. In the first pass we checked the more settled aspects of the annotation: the attributes encoding morphosyntactic information, referentiality (non referring / referring), identity anaphora, and bridging references (including e.g., checking split antecedent anaphora). The second pass was devoted to the more complex forms of annotation, including in particular genericity, ambiguity, and discourse deixis. In this second pass, we also reconsidered the annotation of the dialogue corpora at the light of the experience with the CODI-CRAC annotation. In both passes all documents were checked and possibly corrected; and each document was completely checked by each annotator.

6 Correcting tokenization and conversion errors

ARRAU 3 fixes a couple of errors and inconsistencies in the markup in previous versions. If the corrections resulted in modifications to the underlying text (the *basedata* in MMAX2 parlance), existing annotations were adapted such that they were still valid. Depending on the complexity of the corrections, they were performed in a fully or semi-automatic manner (based on scripts using pyMMAX2 (Müller, 2020)), with manual checks afterwards.

6.1 Tokenization

Tokenization, i.e. splitting of text into basedata elements, was improved for all sub-corpora by using a more fine-grained splitting scheme than the previous one, which was only sensitive to white space

[7] https://github.com/arrauproject/data/blob/main/ARRAU_3_Annotation_Manual_1.0.pdf

and punctuation. Most notably, basedata is split at word-internal non-word characters, including hyphens. As a result, hyphenated words (e.g. noun compounds and other hyphenated multi-word expressions) will be separated into several contiguous basedata elements, allowing for more fine-grained annotation. At the same time, tokenization keeps track of the original input string composition, including white space, and stores, for every basedata element, the number of leading white space characters. This way, the original text appearance can be reproduced in the the annotation tool MMAX2, allowing for a better-to-read, more natural and less distracting rendering of the display.

6.2 PRD Conversion Errors

Some errors were found in the RST portion of the dataset. The RST portion was originally converted from the Penn Treebank PRD format. During the first round of checks, we discovered that this conversion had introduced a couple of errors. NPs for numbers which contained commas as separators (Example (24), from WSJ_0012) were incorrectly truncated, resulting in only the first number (*4* in the example) to be imported into the ARRAU data.

(24)
```
...
(NP (NP average circulation)
    (PP of
    (NP (NP 4,393,237))))
...
```

Sentence annotations, which are instrumental for structuring the annotation tool display by adding sentence-final line breaks, are derived from the PRDs top-level S-bracketings. In previous versions of ARRAU, sentence annotations frequently left out trailing punctuations, causing both sentence-final markables to be incomplete, and the display to be incorrect. Yet another class of errors in previous versions of ARRAU were caused by imperfect creation of the PRD files from the original raw files, in cases where the original text contained slashes. Example (25), from WSJ_0207, shows the rendering in the PRD file (which is also used in ARRAU. In the original file, however, which is also distributed with ARRAU, the actual text reads "11 1/2 minutes".

(25)
```
...
(VP lasts
 (NP-TMP (QP 11 1) minutes))
...
```

While none of these issues are critical, correcting them may also help a future integration in the corpus of other types of annotation available for the RST subset, in particular discourse structure but also for instance PropBank information.

7 ARRAU3: Statistics and Availability

Basic Statistics Table 1 compares the three releases of ARRAU in terms of total number of documents, tokens, and markables. ARRAU3 is only slightly larger than ARRAU 2 in terms of documents (**DC**) (558 vs 552) tokens (**TK**) (359,500 vs 348,072) and markables (**MK**) (106,700 vs 99,582). The number of non-referring expressions and discontinuous markables in ARRAU 3 is also similar to that in ARRAU 2, suggesting that this aspect of the annotation is by now fairly stable.

Complex forms of anaphoric reference Table 2 shows that the difference between ARRAU 3 and ARRAU 2 is much more substantial when considering more complex cases of anaphoric reference. The figures for discourse deixis (**DD**) and split-antecedent plurals (**SP**) didn't change much - suggesting again that these annotations are fairly stable. However, the number of generic markables (**GE**), bridging references (**BG**) and markables identified as ambiguous (**AMB**) are much higher.

Formats The corpus is available in the native MMAX XML format as well as in the Universal Anaphora format.

Availability Like the previous version, all of ARRAU 3 will be available through LDC, whereas the copyright-free subcorpora (GNOME, PEAR, and TRAINS-91) will also be available through the Universal Anaphora repository.

8 Conclusion and Future Work

ARRAU is a long-term project to push forward the state of the art in anaphoric annotation. During each phase of the annotation we discovered new issues that were then corrected in the subsequent version. So while we think the newest release is much improved over ARRAU 2, a number of issues were identified in the last round of annotation, that we hope to correct in future releases. They include in particular several issues related to reference in dialogue (e.g., how to annotate repairs) as well as more complex forms of discourse deixis.

		ARRAU1			ARRAU2			ARRAU3		
		DC	**TK**	**MK**	**DC**	**TK**	**MK**	**DC**	**TK**	**MK**
RST	train				335	182031	57686	333	182424	57489
	dev				18	12845	3986	18	12845	3962
	test				60	33225	10341	60	33225	10319
	overall	204	146512	45990	413	228901	72013	411	228494	71770
TRAINS	91				16	14496	2884	16	14496	3706
	93				98	69158	14115	98	69158	17262
	overall	35	25783	5198	114	83654	16999	114	83654	20968
PEAR		20	14059	3881	20	14059	4008	20	14059	4023
GNOME	2	5	21599	6215	5	21458	6562	5	21458	6571
	2001							8	11835	3368
	overall	5	21599	6215	5	21458	6562	13	33293	9939
Total		264	184,748	60884	552	348,072	99582	558	359,500	106,700

Table 1: Size comparison between ARRAU 3 and previous releases in terms of documents (DC), tokens (TK), and markables (MK)

		ARRAU2					ARRAU3				
		GE	**BG**	**DD**	**SP**	**AMB**	**GE**	**BG**	**DD**	**SP**	**AMB**
RST	train	753	2797	496	346	68	5149	5398	578	353	430
	dev	198	277	36	27	0	814	431	49	26	38
	test	487	703	99	63	14	907	966	98	69	110
	overall	1438	3777	631	436	82	6870	6795	725	448	578
TRAINS	91	98	74	154	48	22	107	176	163	59	168
	93	635	636	708	182	99	651	1007	725	257	245
	overall	733	710	862	230	121	758	1183	888	316	413
PEAR		74	333	67	30	31	175	346	71	32	63
GNOME	2	12	692	73	43	16	814	737	74	53	78
	2001						800	396	9	0	11
	overall	12	692	73	43	16	1614	1133	83	53	89
Total		2257	5512	1633	739	250	9417	9457	1767	849	1143

Table 2: Complex types of anaphora in ARRAU 3 and the previous release ARRAU 2. GE=generic, BG=bridging, DD=discourse deixis, SP=split-antecedent plurals, AMB=ambiguous.

Bibliographical References

References

Ron Artstein and Massimo Poesio. 2006. Identifying reference to abstract objects in dialogue. In *Proc. of BRANDIAL*, Potsdam.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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).

Donna Byron. 2002. Resolving pronominal references to abstract entities. In *Proc. of the ACL*, pages 80–87.

Donna Byron and James Allen. 1998. Resolving demonstrative anaphora in the TRAINS-93 corpus. In *Proc. of the Second Colloquium on Discourse, Anaphora and Reference Resolution*. University of Lancaster.

Greg N. Carlson and Francis J. Pelletier, editors. 1995. *The Generic Book*. University of Chicago Press.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. Kuppevelt and R. Smith, editors, *Current Directions in Discourse and Dialogue*, pages 85–112. Kluwer.

Wallace L. Chafe. 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Ablex, Norwood, NJ.

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.

Herbert H. Clark and Susan E. Brennan. 1990. Grounding in communication. In L. B. Resnick, J. Levine, and S. D. Behrend, editors, *Perspectives on Socially Shared Cognition*. APA.

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).

Miriam Eckert and Michael Strube. 2001. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*.

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.

Loïc Grobol. 2020. *Coreference resolution for spoken French*. Ph.D. thesis, Université Sorbonne Nouvelle.

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

Peter A. Heeman and James F. Allen. 1995. The TRAINS-93 dialogues. TRAINS Technical Note TN 94-2, University of Rochester, Dept. of Computer Science, Rochester, NY.

Peter A. Heeman and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational Linguistics*, 21(3):351–382.

Erhard W. Hinrichs, Sandra Kübler, Karin Naumann, Heike Telljohann, and Julia Trushkin. 2004. Recent developments in linguistic annotations of the TüBa-D/Z treebank. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, Tübingen, Germany.

Yufang Hou, Katja Markert, and Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2082–2093, Doha, Qatar. Association for Computational Linguistics.

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

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

Lauri Karttunen. 1976. Discourse referents. In J. McCawley, editor, *Syntax and Semantics 7 - Notes from the Linguistic Underground*, pages 363–385. Academic Press, New York.

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. 2020. Bridging resolution: A survey of the state of the art. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3708–3721, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Varada Kolhatkar and Graeme Hirst. 2014. Resolving shell nouns. In *Proc. of EMNLP*, pages 499–510, Doha, Qatar.

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.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proc. of the ACL*, Juju island, Korea.

Natalia N. Modjeska. 2003. *Resolving other anaphors*. Ph.D. thesis, University of Edinburgh.

Mark-Christoph Müller. 2008. *Fully Automatic Resolution of It, This And That in Unrestricted Multy-Party Dialog*. Ph.D. thesis, Universität Tübingen.

Mark-Christoph Müller. 2020. pyMMAX2: Deep access to MMAX2 projects from python. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 167–173, Barcelona, Spain. Association for Computational Linguistics.

Mark-Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In S. Braun, K. Kohn, and J. Mukherjee, editors, *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, volume 3 of *English Corpus Linguistics*, pages 197–214. Peter Lang.

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 *Proc. of LREC*.

Costanza Navarretta. 2000. Abstract anaphora resolution in Danish. In *Proc. of the 1st SIGdial Workshop on Discourse and Dialogue*, pages 56–65. ACL.

Anna Nedoluzhko. 2013. Generic noun phrases and annotation of coreference and bridging relations in the prague dependency treebank. In *Proc. of LAW*, pages 103–111.

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.

Barbara Hall Partee. 1972. Opacity, coreference, and pronouns. In D. Davidson and G. Harman, editors, *Semantics for Natural Language*, pages 415–441. D. Reidel, Dordrecht, Holland.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proc. of the ACL Workshop on Discourse Annotation*, pages 72–79, Barcelona.

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. 2005a. Annotating (anaphoric) ambiguity. In *Proc. of the Corpus Linguistics Conference*, Birmingham.

Massimo Poesio and Ron Artstein. 2005b. 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 and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proc. of LREC*, Marrakesh.

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, Rahul Mehta, Alex Maroudas, and Janet Hitzeman. 2004. Learning to solve bridging references. In *Proc. of ACL*, pages 143–150, Barcelona.

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.

Massimo Poesio and Hannes Rieser. 2010. Completions, coordination, and alignment in dialogue. *Dialogue and Discourse*, 1(1):1–89.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

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, 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.

Sameer S. Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Indentifying entities and events in ontonotes. In *Proc. IEEE International Conference on Semantic Computing (ICSC)*, Irvine, CA.

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

Ina Roesiger. 2018. Rule- and learning-based methods for bridging resolution in the ARRAU corpus. In *Proc. of CRAC*.

Ina Roesiger, Arndt Riester, and Jonas Kuhn. 2018. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3516–3528, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

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*.

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

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.

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 *Proc. of CODI/CRAC Shared Task*.

Juntao Yu, Sopan Khosla, Nafise Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. 2022b. The universal anaphora scorer 1.0. In *Proc. of LREC*.

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.