Anaphora Resolution with the ARRAU Corpus

Massimo Poesio,[1] Yulia Grishina,[2] Varada Kolhatkar,[3] Nafise Sadat Moosavi,[4] Ina Roesiger,[5] Adam Roussel,[6] Fabian Simonjetz,[6] Alexandra Uma[1], Olga Uryupina[7], Juntao Yu,[1] Heike Zinsmeister[8]

[1]Queen Mary University of London, [2]University of Potsdam, [3]Simon Fraser University,
[4]HITS Heidelberg, [5]University of Stuttgart, [6]Ruhr University Bochum,
[7]University of Trento, [8]University of Hamburg

Abstract

The ARRAU corpus is an anaphorically annotated corpus of English providing rich linguistic information about anaphora resolution. The most distinctive feature of the corpus is the annotation of a wide range of anaphoric relations, including bridging references and discourse deixis in addition to identity (coreference). Other distinctive features include treating all NPs as markables, including non-referring NPs; and the annotation of a variety of morphosyntactic and semantic mention and entity attributes, including the genericity status of the entities referred to by markables. The corpus however has not been extensively used for anaphora resolution research so far. In this paper, we discuss three datasets extracted from the ARRAU corpus to support the three subtasks of the CRAC 2018 Shared Task–identity anaphora resolution over ARRAU-style markables, bridging references resolution, and discourse deixis; the evaluation scripts assessing system performance on those datasets; and preliminary results on these three tasks that may serve as baseline for subsequent research in these phenomena.

1 Introduction

The release of the ONTONOTES coreference corpus (Pradhan et al., 2007a) and the organization of two CONLL shared tasks based on the dataset (Pradhan et al., 2012) have resulted in a substantial increase in coreference research, both in terms of quantity and in terms of quality. We expect ONTONOTES to remain a key resource for the field for many years.

However, ONTONOTES also has a number of frequently mentioned limitations, including:

- Not all NPs of relevance to anaphora resolution are treated as markables. For instance, expletives are not annotated.
- And even among referring markables, singletons are not annotated, nor are references to abstract objects or many types of generic objects (Pradhan et al., 2012).

Furthermore, anaphora resolution involves a number of phenomena besides 'coreference', such as bridging reference (Clark, 1975) and discourse deixis (Webber, 1991). Only a simple form of discourse deixis, event anaphora, is annotated in ONTONOTES; bridging reference was not annotated, although a subset of the corpus has been annotated with this information by Markert et al. (2012).

A number of these limitations are overcome in the ARRAU corpus (Uryupina et al., In press). In ARRAU, all NPs are considered markables, including expletives and singletons. Both discourse deixis and bridging reference have been annotated.

The corpus however, hasn't been widely used for anaphora resolution research yet, with a few exceptions (Rodriguez, 2010; Uryupina and Poesio, 2012; Marasović et al., 2017). There are a number of reasons for this, ranging from the fact that research in both bridging reference and discourse deixis is still limited, to the unusual markup format. The objective of this paper is to introduce the community to the three datasets extracted from the ARRAU corpus to support this year's CRAC18 Shared task, the first evaluation campaign based on ARRAU. Our hope is that making such datasets available may, on the one hand, facilitate the use of ARRAU; on the other, increase the community of researchers working on these aspects of anaphora resolution.

2 The ARRAU Corpus

2.1 Genres

The ARRAU corpus includes a substantial amount of news text in the sub-corpus called RST, con-

Proceedings of the Workshop on Computational Models of Reference, Anaphora and Coreference, pages 11–22
New Orleans, Louisiana, June 6, 2018.

sisting of the entire subset of the Penn Treebank (Marcus et al., 1993) that was annotated in the RST treebank (Carlson et al., 2003). News data were annotated so that researchers could compare results on ARRAU with results on other news datasets; and these documents were chosen because they had already been annotated in a number of ways—not only syntactically (e.g., through the Penn Treebank (Marcus et al., 1993)) and for their argument structure (e.g., through Propbank (Palmer et al., 2005)) but also for rhetorical structure (Carlson et al., 2003). But one of the objectives of the ARRAU annotation was to cover genres other than news, so, in addition to RST, ARRAU includes three more sub-corpora. The TRAINS sub-corpus includes all the task-oriented dialogues in the TRAINS-93 corpus;[1] 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); and the GNOME sub-corpus covers documents from the medical and art history genres covered by the GNOME corpus (Poesio, 2000a, 2004b) used to study both local and global salience (Poesio et al., 2004, 2006). The same coding scheme was used for all sub-corpora, but separate guidelines were written for the textual and the spoken dialogue sub-corpora. Table 1 provides basic statistics about the four ARRAU sub-corpora. Note in particular the large number of non-referring markables. RST, TRAINS and PEAR were used for the CRAC 2018 shared task.

2.2 Markables

Markable definition Many, especially among the older, anaphorically annotated corpora impose syntactic, semantic or discourse-based restrictions on markables. For instance, in ONTONOTES neither expletives nor singletons are annotated (for a discussion of the state of the art in anaphoric annotation, see (Poesio et al., 2016)). By contrast, in ARRAU *all* NPs are considered as markables, also when they are non-referring because either expletives such as *it* or predicative NPs such as *a busy place* in (1), or when they do not corefer with any other markable and thus form a singleton coreference chain. Moreover, non-referring markables are manually sub-classified into expletives, predicative, and quantifiers. In addition, possessive pronouns are marked as well, and all premodifiers are marked when the entity referred to is mentioned again, e.g., in the case of the proper name *US* in (2), and when the premodifier refers to a kind, like *exchange-rate* in (3).

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

(2) . . . The Treasury Department said that the $[\text{US}]_1$ trade deficit may worsen next year after two years of significant improvement. . . The statement was the $[\text{US}]_1$'s government first acknowledgment of what other groups, such as the International Monetary Fund, have been predicting for months.

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

In ARRAU, the full NP is marked with all its modifiers; in addition, a MIN attribute is marked, as in the MUC corpora. For nominal markables, MIN is the head noun, whereas for (modified or not) named entities MIN is the entire proper name.

(4) $[^{min}[\text{Alan Spoon}]^{min}$, recently named Newsweek president] , said
Newsweek's ad rates would increase 5% in January.

Markable properties All markables are manually annotated for a variety of properties according to the GNOME guidelines (Poesio, 2000b): these include morphosyntactic agreement (gender, number and person), grammatical function, and the semantic type of the entity. The guidelines and reliability studies leading to this scheme are discussed in (Poesio, 2000a, 2004a; Uryupina et al., In press). We will only mention one attribute here, the reference attribute, that specifies a combination of information about the logical form status of the NP (referring, expletive, quantificational, or predicative), and can be used to distinguish between referring and non-referring markables.

2.3 Types of anaphoric relations marked

The ARRAU guidelines support annotation of different types of anaphoric relations. All referring markables are marked as either discourse

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

	RST	GNOME	PEAR	TRAINS
documents	413	5	20	114
tokens	228901	21458	14059	83654
avg. doc length (tok)	554.2	4291.6	703.0	733.8
markables	72013	6562	4008	16999
avg. markables per doc	174.4	1312.4	200.4	149.1
non-referring markables	9552 (13.3%)	1047 (16.0%)	607 (15.1%)	2353 (13.8%)

Table 1: Corpus statistics for the four ARRAU sub-corpora.

`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 mentions, an antecedent can be identified, either of type `phrase` (if the antecedent was introduced using a nominal markable) or `segment` (not introduced by a nominal markable, for **discourse deixis**). 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 The term **bridging reference** was introduced by Clark (1975) to refer to any reference that requires some sort of 'bridging' inference to be interpreted. Clark's very general definition covered both identity anaphora in which the description of the anaphor is different from the description of the antecedent, as in (5); and so-called **associative** anaphora (Hawkins, 1978), in which the anaphoric expression refers to an object that is associated with, but not identical to, the antecedent, as in (6). (These days, the term bridging reference is mostly used to refer to the associative cases.)

(5) I saw a black Mercedes parked outside the restaurant. [The car] belonged to Bill.

(6) I saw a black Mercedes parked outside the restaurant. [The engine] was still running.

Annotating—indeed, even identifying—bridging references in a reliable way is difficult (Vieira, 1998; Poesio and Vieira, 1998), which is one of the reasons why so few large-scale corpora for anaphora include this type of annotation (Poesio et al., 2016). The ARRAU guidelines for bridging anaphora are based on experiments that started with the work of Vieira and Poesio (Vieira, 1998; Poesio and Vieira, 1998) and continued in the GNOME project (Poesio, 2004a).

In GNOME, a subset of relations that could be annotated reliably was found (Poesio, 2004a), including three types of relations: `element-of`; `subset`; and a generalized possession relation `poss` covering both part-of relations and general possession relations. The ARRAU Release 1 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 relations identified in GNOME (indeed, the same examples were used), and in addition, if they stood in two additional relations (but without testing the reliability of this annotation):

- `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'.

The simplification was that in ARRAU Release 1, coders were not asked to specify the relation—effectively, any associative bridging reference was considered a case of 'underspecified relation'. In ARRAU Release 2, the annotation of bridging references was revised for the RST domain only and coders were now asked to mark the relations only in that domain. Some statistics about bridging references in ARRAU Release 2 are shown in Table 2. A total of 5512 bridging references were marked, but a classification of the relations was only provided for the 3777 bridging references identified in the RST domain. In the table, we write P+S+E+O+U as category for the bridging references in the other domains, currently not classified.

Discourse deixis The term **discourse deixis** was introduced by Webber (1991) to indicate the reference to abstract entities which have not been introduced in the discourse through a nominal markable, as in the following example from the TRAINS corpus, where *that* in utterance 7.6 refers to the plan of shipping boxcars to oranges to Elmira.

	RST	TRAINS	GNOME	PEAR	TOTAL
all	3777	710	692	333	5512
poss	87				$\geq$ 87
poss-inv	25				$\geq$ 25
subset	1092				$\geq$ 1092
subset-inv	368				$\geq$ 368
element	1126				$\geq$ 1126
element-inv	152				$\geq$ 152
other	332				$\geq$ 332
other-inv	7				$\geq$ 7
undersp-rel	588				$\geq$ 588
P+S+E+O+U	N/A	710	692	333	1735

Table 2: Distribution of bridging references in ARRAU.

(7)
7.3 : so we ship one
7.4 : boxcar
7.5 : of oranges to Elmira
7.6 : and that takes another 2 hours

Discourse deixis is a very complex form of reference, both to annotate (Artstein and Poesio, 2006) and to resolve. Very few anaphoric annotation projects have attempted annotating discourse deixis in its entirety (Artstein and Poesio, 2006; Dipper and Zinsmeister, 2012). More typical is a partial annotation, as in (Byron and Allen, 1998; Navarretta, 2000), who annotated pronominal reference to abstract objects; in ONTONOTES, where event anaphora was marked (Pradhan et al., 2007b); and in the work of Kolhatkar (2014), that focused on so-called shell nouns. In ARRAU,

1. A coder specifying that a referring expression is discourse old is asked whether its antecedent was introduced using a `phrase` (markable) or `segment` (discourse segment).

2. Coders choosing `segment` have to mark a sequence of *predefined* clauses.

Statistics about discourse deixis in ARRAU Release 2 are shown in Table 3. A total of 1633 cases of discourse deixis were marked.

2.4 Markup

ARRAU was annotated using the MMAX2 annotation tool (Müller and Strube, 2006). MMAX2 is based on **token standoff** technology: the annotated anaphoric information is stored in a `phrase` level whose markables point to a base layer in which each token is represented by a separate XML element.

2.5 Two releases

There have been two releases of the corpus. The first release, in 2008, is discussed in (Poesio and Artstein, 2008). This first release was relatively small (about 100K words in total), and focused primarily on identity anaphora and on the annotation of ambiguity, but its development involved extensive experiments with the annotation of discourse deixis and of ambiguity that led to the annotation guidelines used throughout the project (Poesio and Artstein, 2005b,a; Artstein and Poesio, 2006). The second release, via LDC in 2013, is substantially larger than the first (350K) and the annotation of bridging reference, discourse deixis and genericity is much more extensive. Another key annotation effort was the annotation of minimal spans of markables (`MIN`s). Last but not least, extensive checks were run on the annotation of identity anaphora. This is the release used for the CRAC 2018 Shared Task.

3 Previous work on anaphora resolution with ARRAU

3.1 Identity anaphora

Rodriguez (2010) used BART (Versley et al., 2008) to compare the difficulty of ARRAU and the two more widely used corpora at the time, `MUC-7` and `ACE02`, and the effect of using `MIN` information to ascribe partial credit (50%) whenever a system markable overlaps with the minimal span of a gold markable, and the boundaries of the system markable do not exceed those of the gold markable, as done in MUC. He found that assigning such partial credit substantially improves the scores.

Uryupina and Poesio (2012) explored the effect of domain adaptation in anaphora resolution, comparing the results obtained by training different versions of BART separately for each domain

RST	TRAINS	GNOME	PEAR	TOTAL
631	862	73	67	1633

Table 3: Distribution of discourse deixis in the subdomains of ARRAU.

	Soon et al 2001		Extended feature set	
	Domains	Union	Domains	Union
ARRAU				
GNOME	58.06	56.92	56.38	56.11
PEAR	66.74	67.36	66.29	65.24
RST	59.51	59.36	56.88	57.97
TRAINS-93	43.17	42.9	47.55	43.31
overall	56.66	56.04	54.84	55.29
ONTONOTES				
bc	55.04	55.62	60.71	59.52
mz	59.56	60.2	61.65	62.42
wb	51.07	53.05	53.91	53.36
whole	54.17	54.5	57.74	57.05

Table 4: (Uryupina and Poesio, 2012): Running BART on different ARRAU genres and on different ONTONOTES genres. MUC score.

or the entire dataset. They did that on both ARRAU 2 and ONTONOTES, thus providing what to our knowledge is the only comparison between the two corpora in terms of system performance. Table 4 summarizes the results.

3.2 Discourse Deixis

Marasović et al. (2017) developed an approach to abstract anaphora resolution based on bi-directional LSTMs to produce representations of the anaphor and the candidate sentence, and a mention ranking component adapted from the systems by Clark and Manning (2016) and Wiseman et al. (2015). The system was tested using both the dataset by Kolhatkar et al. (2013) (for shell nouns) and the discourse deixis cases in ARRAU.

4 The Three Tasks of CRAC 2018

The CRAC 2018 Shared Task was the evaluation campaign associated with this workshop. The task was articulated in three subtasks: a first task on identity anaphora resolution, a second one on bridging reference, and a third one on discourse deixis. Researchers could participate independently, and indeed no group participated in more than one task. In this Section we discuss how the datasets for the three tasks were created using ARRAU, and the evaluation scripts that were used.

4.1 Markable Settings

One characteristic in common to all three subtasks is that the official evaluation of systems was based on a **gold** setting, in that the markables were specified in advance.[2] This was done because the organizers of Tasks 2 and 3 felt that the state of the art in bridging anaphora and discourse deixis resolution is such that the system markable setting would be too hard, so we would need to release data in a gold setting for those tasks–and then of course it would not make sense to release them in a system markables setting for Task 1. The evaluation scripts however supported both gold and predicted markables, and the evaluations reported below carried out both.

4.2 Task 1: Identity anaphora

In this task, systems have to decide

- whether a markable is referring or not;
- if referring, whether it introduces a new entity/coreference chain (discourse new) or refers to an entity already introduced (discourse old);
- in case it is classified as discourse old, the systems have to identify the antecedent (entity, or coreference chain).

Data format For this task, the documents were exported in the format used for EVALITA-2011 (Uryupina and Poesio, 2013), derived from the tabular CONLL-style format used in the SEMEVAL 2010 shared task on multilingual anaphora (Recasens et al., 2010). The format used involves three tab-separated columns, with one line per token:

```
TOKEN    MARKABLE        MIN
```

The first column specifies the token; the second column specifies whether the token belongs to a markable in BIO format (as said above, evaluation is on gold markables, although participants could also submit runs for systems-markables evaluation); and the third column specifies which token is the minimal span (MIN) of the markable, in the sense of MUC. So for example, the first line of the

[2] Given that non-referring NPs and NPs referring to singletons are annotated in ARRAU, however, the 'gold' setting in fact resembles more the 'gold markable boundaries' setting used in the CONLL 2012 shared task (Pradhan et al., 2012) than the gold setting for that task.

document wsjarrau_2308.CONLL consists of the following three columns:

```
Ripples B-markable_45    word_1
```

where Ripples is the token (in this case, the first token of the document, i.e., word_1); the second column says the token is the beginning of markable_45; and the third column says the MIN word of the markable is token 1, i.e., this very same token (note that token indices start from 1).

The task of a system is to decide whether a markable is referring, and if so, the coreference chain it belongs to (possibly a singleton). Participation in a coreference chain is represented using the markable=set notation from EVALITA, a slight variation of the standard CONLL notation which generalizes to representations for bridging reference and discourse deixis as well, as discussed below. In the case of the example line above, the gold version of the document contains the following line:

```
Ripples B-markable_45=set_37 word_1 new
```

which states that markable_45 is referring; that the entity it refers to is discourse-new (fourth column); and that this entity is coreference chain set_37. (The EVALITA notation can easily be converted into the CONLL notation to use the standard CONLL scorer as well, as we did–see below.)

In case a token is part of distinct markables, the @ notation from EVALITA 2011 is used, derived from the | notation from SEMEVAL 2010. Consider for instance the first few lines of the same test set file, representing the NP

Ripples from the strike by 55,000 Machinists Union members against Boeing Co..

One plausible syntactic analysis of this NP can be represented using brackets as follows:

```
[Ripples from [the strike by [55,000
[Machinists Union] members] against
[Boeing Co.]]]
```

In EVALITA notation, the embedding of markables is represented as follows (to make the example more readable, coreference chain information has been omitted, and the annotation has been slightly formatted)

```
Ripples B-markable_45                    word_1
from    I-markable_45                    word_1
the     I-markable_45@B-markable_47      word_1@word_4
strike  I-markable_45@I-markable_47      word_1@word_4
by      I-markable_45@I-markable_47      word_1@word_4
55,000  I-markable_45@I-markable_47@B-markable_49
                                         word_1@word_4@word_6
Machinists  I-markable_45@I-markable_47@I-markable_49@
     B-markable_609              word_1@word_4@word_6@word_8
union   I-markable_45@I-markable_47@I-markable_49
     @I-markable_609             word_1@word_4@word_6@word_8
members I-markable_45@I-markable_47@I-markable_49
                                         word_1@word_4@word_6
against I-markable_45@I-markable_47      word_1@word_4
Boeing  I-markable_45@I-markable_47@B-markable_50
                                word_1@word_4@word_11..word_12
Co.     I-markable_45@I-markable_47@I-markable_50
                                word_1@word_4@word_11..word_12
```

This states that, for instance, the token Machinists is the Beginning of markable_609, which in turn is Inside markable_49, in turn markable_47, and then of markable_45. For each of these markables, the coreference chain to which it belongs is specified using the The third column specifies the MINs of each of these markables, again using the @ notation.

A system correctly interpreting these markables should output for every markable its coreference chain and information status (non referring, discourse new, or discourse old).

Evaluation script The coreference evaluation script developed by Moosavi and Strube was modified to produce the scorer for Task 1. We will refer to this script as 'the extended coreference scorer' below.[3] The extended scorer, when run excluding non-referring expressions and singletons and ignoring MIN information, evaluates a system's response using the same metrics (indeed, a reimplementation of the same code) as the standard CONLL evaluation script, v8 (Pradhan et al., 2014).[4] When required to use MIN information, the extended scorer follows the MUC convention, and considers a mention boundary correct if it contains the MIN and doesn't go beyond the annotated maximum boundary. When singletons are to be considered, singletons are also included in the scores (all metrics apart from MUC can deal with singletons). Finally, when run in all-markables mode, the script scores referring and non-referring expressions separately. Referring expressions are scored using the CONLL metrics; for non-referring expressions, the script evaluates P, R and F1 at non-referring expression identification. The extended coreference scorer is available from Moosavi's github at https://github.com/ns-moosavi/coval.

4.3 Task 2: Bridging Anaphora

Data format For the bridging task, the documents were exported in a similar format to that

[3] Discussions are under way to incorporate some of the aspects of this scorer in the official CONLL scorer.

[4] In addition to MELA and related metrics, the extended scorer also computes Moosavi and Strube's LEA metric (Moosavi and Strube, 2016).

of Task 1. Again, the test set already specifies the gold markables (in this case, only the bridging references). The test set provides four tab-separated columns, with one line for each token:

```
TOKEN MARKABLE MIN BRIDGE
```

The meaning of the first three columns is as in Task 1. The fourth column specifies whether the markable is a bridging reference. For example, the following lines

```
a          B-markable_311 word_695 B-markable_311
speedy     I-markable_311 word_695 I-markable_311
resolution I-markable_311 word_695 I-markable_311
```

state that tokens `a`, `speedy`, and `resolution` are part of `markable_311`, with head token `word_695`, and that this markable is a bridging reference. The objective of participating systems is to identify which **anchor entity** and **anchor markable** referring to that entity the bridging reference refers to, using the notation

```
bridg_ref=bridg_rel=_anchor_mark=_anchor_ent
```

For example, in the case of `markable_311` above, the correct answer would be:

```
a       B-markable_311=set_148  word_695
            B-markable_311=undersp-rel=markable_308=set_3
speedy  I-markable_311=set_148  word_695
            I-markable_311=undersp-rel=markable_308=set_3
resolution      I-markable_311=set_148  word_695
            I-markable_311=undersp-rel=markable_308=set_3
```

stating that `markable_311` has been identified as belonging to entity `set_148` as well as being an associative reference to entity `set_3` through the `undersp-rel` relation.

Evaluation script The evaluation script for Task 2 is based on the evaluation method proposed in (Hou et al., 2013). The script separately measures precision and recall at anchor entity recognition (e.g., whether `set_3` is the right coreference chain) and at anchor markable detection (i.e., whether `markable_308` is the appropriate markable of `set_3`). Note that whereas the identification of the anchoring entity is considered correct whenever the right coreference chain is identified, irrespective of the particular anchor markable chosen, the identification of the anchor markable is strict, i.e., it is only considered correct if the same markable as annotated is found.

4.4 Task 3: Discourse deixis

Finally, in this task (discourse deixis) systems have to identify the **unit**–clausal text segment–that evokes the abstract entity the discourse deixis refers to.

For this task, the documents have been exported in a format again consisting of three columns, again with one line for each token:

```
TOKEN UNIT MARKABLE
```

The second column specifies which unit (= utterance in the case of dialogue data, clause in the case of textual data) the token belongs to. (All units have already been marked, so systems do not need to recognize them.) The third column specifies whether the token belongs to a discourse deixis - and if so, which unit (utterance) evoked the antecedent.

For example, consider the following fragment:

```
TOKEN      UNIT                              MARKABLE
But        B-markable_565
some       I-markable_565
investors  I-markable_565
might      I-markable_565
prefer     I-markable_565
a          I-markable_565
simpler    I-markable_565
strategy   I-markable_565
then       I-markable_565
hedging    I-markable_565@B-markable_106
their      I-markable_565@I-markable_106
individual I-markable_565@I-markable_106
holdings   I-markable_565@I-markable_106
.          I-markable_565
They       B-markable_566
can        I-markable_566
do         I-markable_566
this       I-markable_566                    B-markable_322
...
```

The first 14 lines contain tokens belonging to unit `markable_565`. The following 4 lines contain tokens belonging to unit `markable_566`. The last of these is marked as a discourse deixis:

```
this    I-markable_566  B-markable_322
```

This line states that token `this` belongs to unit `markable_566`[5], and it is the beginning of a discourse deixis, `B-markable_322`. The systems' task is to identify which unit the discourse deixis refers to. The gold interpretation, using the `=unit:<markable_ID> format` would be as follows:[6]

```
this    I-markable_566
       B-markable_322=unit:markable_565
```

Evaluation script The evaluation script for Task 3 computes the **Success@N** metric proposed by Kolhatkar (e.g., (Kolhatkar and Hirst, 2014)) and also used by Marasović et al. (2017). SUCCESS@N is the proportion of instances where the gold answer–the unit label–occurs within a systems first n choices. (S@1 is standard precision.)

[5] All levels of annotation have markables named `markable_N` where N is an integer, but those names are independent: so unit `markable_566` is different from coreference `markable_566`.

[6] It is actually not entirely clear from the example whether demonstrative *this* refers to 'preferring a simpler strategy' or 'hedging their individual holdings' or, more likely, a more complex abstract object.

Configuration	P	R	F1
ONTONOTES			
CoreNLP CoNLL predicted	40.38	89.46	55.65
CoreNLP Rule-based	43.68	83.56	49.02
CoreNLP Hybrid	33.3	84.9	47.84
CoreNLP Dep	32.23	82.20	46.30
Our LSTM Best F1	73.53	74.01	73.77
Our LSTM High Recall	51.53	87.53	64.87
ARRAU RST			
CoreNLP Rule-based	70.95	62.74	66.59
CoreNLP Hybrid	71.55	67.28	69.35
CoreNLP Dep	70.27	66.08	68.11
Our LSTM	79.33	86.16	82.60

Table 5: Markable extraction in ARRAU and ONTONOTES.

5 Anaphoric Resolution with The Three New Datasets: Results

No system participated in Task 1 and Task 3 of the shared task. In this Section we discuss the results obtained with Task 2, as well as the baseline results for markable extraction and Task 1.

5.1 Markable extraction

One of the important differences between corpora for anaphora / coreference is the definition of mentions (or markables, in this case). In order to compare the difficulty of markable extraction in ARRAU with that of mention extraction ONTONOTES, we ran two markable extractors on both corpora: a few versions of a mention extractor based on the Stanford CORE pipeline, and our own implementation of an LSTM architecture for markable extraction. Our markable extractor is a modified version of the neural named entity recognition system proposed by Lample et al. (2016). Two versions of this markable extractor were run on the ONTONOTES dataset, one optimized for F1, one for recall. The results are shown in Table 5.

The results suggest that markable extraction in ARRAU is considerably easier than mention extraction in ONTONOTES. This might be due to the differences in markable definition, since singletons and non-referring NPs have to be excluded in ONTONOTES. But the accuracy gaps might also be a result of the domain differences between ONTONOTES and ARRAU. To test this we tested the Stanford pipeline on the WSJ portion of the ONTONOTES test set. The highest scores on the WSJ portion is obtained by the rule-based version of the pipeline, and is lower (43.1% F1) than that for the entire set. This suggests the difference in performance are due to the more releaxed notion of markable used in ARRAU.

Configuration	P	R	F1
Excluding singletons and non-referring			
MUC	72.32	58.88	64.91
B^3	67.85	48.45	56.53
$CEAF_e$	54.24	52.95	53.59
CONLL score			58.34
LEA	43.20	61.61	50.79
CoNLL official scorer			
MUC	72.12	59.02	64.92
B^3	67.56	48.55	56.50
$CEAF_e$	53.99	53.01	53.49
CONLL score	64.56	53.53	58.30
Including singletons but excluding non-referring			
MUC	72.08	58.88	64.81
B^3	77.46	77.12	77.29
$CEAF_e$	64.18	88.13	74.27
CONLL score			72.13
LEA	60.10	64.26	62.11
Results on non-referring			
Non-referring	0	0	0

Table 6: Baseline results on Task 1. Gold markables.

5.2 Task 1

The results from (Uryupina and Poesio, 2012) suggest that the resolution of identity anaphoric reference in ARRAU is no harder than in ONTONOTES, but to further test this the Stanford CORE deterministic coreference resolver (Lee et al., 2013) was run on the RST subset of the dataset for Task 1 as a baseline, using the division into training, development and test built-in the shared task for this subdomain. The system was run both on gold and on predicted mentions, and evaluated first using both the CONLL official scorer and the extended coreference scorer ignoring singletons and non-referring markables, then including those.

On gold markables The first 10 lines of Table 6 show the results obtained using the extended coreference scorer and the CONLL official scorer excluding both singletons (4161 markables) and non-referring markables (1391)–i.e., the same conditions as in the standard CONLL evaluations. In these conditions, the extended coreference scorer and the CONLL official scorer obtain the same scores modulo rounding. The following lines in Table 6 show the results when including in the assessment singletons; for this evaluation, the Stanford deterministic coreference resolver was made to output singletons instead of removing them prior to evaluation. When non-referring markables are included as well, the results for referring expressions remain identical, but in addition, the scorer outputs the results on those separately. (The Stanford deterministic coreference resolver does not attempt to identify non-referring markables, hence all values are 0.)

The first conclusion that can be obtained from this Table is that the results achieved by the Stan-

Configuration	P	R	F1
Exclude singletons and non-referring			
MUC	58.65	42.33	49.17
B^3	53.20	32.40	40.27
$CEAF_e$	42.77	37.88	40.18
CONLL score			43.21
LEA	27.61	46.17	34.55
CoNLL official scorer			
MUC	58.47	42.44	49.18
B^3	53.00	32.53	40.32
$CEAF_e$	42.64	37.98	40.18
CONLL score	51.37	37.65	43.23

Table 7: Baseline results on Task 1 with predicted mentions, without MIN information.

Configuration	P	R	F1
Exclude singleton and non-referring			
MUC	67.83	46.93	55.48
B^3	62.93	36.90	46.52
$CEAF_e$	47.48	42.05	44.60
CONLL score			48.87
LEA	56.71	32.27	41.13

Table 8: Baseline results on Task 1 with predicted mentions, using MIN information.

ford resolver on gold markables on this dataset are broadly comparable to the results the system achieved on gold markables at CONLL 2011, where it achieved a CONLL score of 60.7. The second observation is that the system appears quite good at identifying singletons, as its CONLL score in that case is over ten percentage points higher– in other words, the system is very much penalized when running on the CONLL dataset.

On Predicted Markables Table 7 shows the results obtained by the Stanford deterministic coreference resolver when evaluated on predicted markables instead of gold markables. These are the results that are more directly comparable with those obtained by this system in the CONLL 2011 shared task. We can see a substantial drop in CONLL score, from 58.3 on predicted markables in the CONLL 2011 shared task to 43.2 on predicted markables with the Task 1 dataset. Most likely, that indicates that some degree of optimization to the characteristics of CONLL dataset was carried out in the system even though the system is not trained.

Using the MIN information Finally, Table 8 shows the effect of using the MIN information. As can be seen from the Table, this results in five extra percentage points.

5.3 Task 2

One aspect of anaphoric interpretation for which there were no previous results with ARRAU is bridging reference. One group from the University of Stuttgart participated in this subtask (Roesiger, 2018). We summarize here the results; for further detail, see the paper.

Roesiger developed two systems, one rule-based, one ML-based. The results obtained by these systems on all three subdomains are summarized in Table 9 in the Appendix. The three columns present the result of the two systems at the tasks of (i) attempting to resolve all gold bridging references; (ii) only producing results when the system is reasonably convinced; and (iii) identifying and resolving bridging references. These results appear broadly comparable to those obtained by Hou et al. (2013) over the ISNotes corpus as far as the RST and TRAINS domain are concerned, but much lower for the PEAR domain– although given the small number of bridging references in this domain (354) not too much should be read into this. See Roesiger (2018) for some interesting hypotheses regarding the differences between the two corpora.

6 Conclusions

In this paper we discuss a dataset based on the ARRAU corpus that supports three fundamental anaphora resolution tasks: identity anaphora resolution, bridging reference resolution, and discourse deixis. We are not aware of any other dataset supporting all three tasks, which makes the resource fairly unique. In this paper we have discussed preliminary experiments with the data that can give other groups an idea of how to use them and what results have been achieved so far.

Acknowledgments

The original work on the ARRAU corpus was supported by EPSRC project ARRAU, GR/S76434/01.[7] This research was supported in part by the ERC project DALI.[8] We wish to thank LDC for their support with the organization and the running of the shared task.

[7] https://arrauproject.wordpress.com/
[8] http://www.dali-ambiguity.org

References

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

D. Byron and J. Allen. 1998. Resolving demonstrative anaphora in the trains-93 corpus. In *Proceedings of the Second Colloquium on Discourse, Anaphora and Reference Resolution*. University of Lancaster.

L. Carlson, D. Marcu, and M. E. 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.

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

H. H. Clark. 1975. Bridging. In *Proceedings of TINLAP*.

K. Clark and C. D. Manning. 2016. Improving coreference resolution by learning entity- level distributed representations. In *Proc. of ACL*, Berlin.

S. Dipper and H. Zinsmeister. 2012. Annotating abstract anaphora. *Language Resources and Evaluation*, 46(1):37–52.

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

Y. Hou, K. Markert, and M. Strube. 2013. Global inference for bridging anaphora resolution. In *Proc. of the NAACL*, pages 907–917, Atlanta, Georgia.

V. Kolhatkar. 2014. *Resolving Shell Nouns*. Ph.D. thesis, University of Toronto.

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

V. Kolhatkar, H. Zinsmeister, and G. Hirst. 2013. Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In *Proc. of EMNLP*, Seattle.

G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL*, pages 260–270. Association for Computational Linguistics.

H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

A. Marasović, L. Born, J. Opitz, and A. Frank. 2017. A mention-ranking model for abstract anaphora resolution. In *Proc. of EMNLP*, pages 221–232, Copenhagen.

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

K. Markert, Y. Hou, and M. Strube. 2012. Collective classification for fine-grained information status. In *Proc. of the ACL*, Jeju island, Korea.

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

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

C. Müller and M. 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.

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

M. Palmer, D. Gildea, and Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–106.

M. Poesio. 2000a. Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In *Proc. of LREC*, pages 211–218, Athens.

M. Poesio. 2000b. *The GNOME Annotation Scheme Manual*, fourth version edition. University of Edinburgh, HCRC and Informatics, Scotland. Available from `http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm`.

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

M. Poesio. 2004b. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proceedings of SIGDIAL*, Boston.

M. Poesio and R. Artstein. 2005a. Annotating (anaphoric) ambiguity. In *Proceedings of the Corpus Linguistics Conference*, Birmingham.

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

M. Poesio and R. Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proc. of LREC*, Marrakesh.

M. Poesio, A. Patel, and B. Di Eugenio. 2006. Discourse structure and anaphora in tutorial dialogues: an empirical analysis of two theories of the global focus. *Research in Language and Computation*, 4:229–257. Special Issue on Generation and Dialogue.

M. Poesio, S. Pradhan, M. Recasens, K. Rodriguez, and Y. 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.

M. Poesio, R. Stevenson, B. Di Eugenio, and J. M. Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

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

S. Pradhan, X. Luo, M. Recasens, E. Hovy, V. Ng, and M. Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proc. of the ACL*, pages 30–35, Baltimore.

S. S. Pradhan, E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2007a. Ontonotes: A unified relational semantic representation. *International Journal on Semantic Computing*, 1(4):405–419.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007b. Unrestricted Coreference: Indentifying Entities and Events in OntoNotes. In *in Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*.

M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proc. SEMEVAL 2010*, Uppsala.

K. Rodriguez. 2010. *Resources for linguistically motivated multilingual anaphora resolution*. Ph.D. thesis, Universitá di Trento.

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

W. M. Soon, D. C. Y. Lim, and H. T. Ng. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4).

O. Uryupina, R. Artstein, A. Bristot, F. Cavicchio, F. Delogu, K. Rodriguez, and M. Poesio. In press. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Journal of Natural Language Engineering*.

O. Uryupina and M. Poesio. 2012. Domain-specific vs. uniform modeling for coreference resolution. In *Proc. of LREC*, pages 187–191, Istanbul. ELRA.

O. Uryupina and M. Poesio. 2013. Evalita 2011: Anaphora resolution task. In *Evaluation of Natural Language and Speech Tools for Italian*, number 7689 in Lecture Notes in Computer Science, pages 146–155. Springer.

Y. Versley, S. Ponzetto, M. Poesio, V. Eidelman, A. Jern, J. Smith, X. Yang, and A. Moschitti. 2008. Bart: A modular toolkit for coreference resolution. In *Proc. of ACL, demo session*, Columbus, OH.

R. Vieira. 1998. *Definite Description Resolution in Unrestricted Texts*. Ph.D. thesis, University of Edinburgh, Centre for Cognitive Science.

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

S. J. Wiseman, A. M. Rush, S. M. Shieber, and J. Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proc. of the ACL*, Bejing.

A Appendix

	Gold bridges-all			Gold bridges-partial			Full bridging resolution		
	P	R	F1	P	R	F1	P	R	F1
RST									
Rule (IR, entity)	39.8	39.8	39.8	63.6	22.0	32.7	18.5	20.6	19.5
Rule (official, phrase)	32.2	32.9	32.5	54.0	19.1	28.2	16.2	12.7	14.2
Rule (official, entity)	36.5	35.7	36.1	58.4	20.6	30.5	16.8	13.2	14.8
ML (IR, entity)	-	-	-	47.0	22.8	30.7	17.7	20.3	18.6
ML (official, phrase)	-	-	-	41.4	13.0	19.8	10.8	12.0	11.4
ML (official, entity)	-	-	-	51.7	16.2	24.7	12.6	15.0	13.7
PEAR									
Rule (IR, entity)	28.2	28.2	28.2	69.2	13.7	22.9	57.1	12.2	20.1
Rule (official, phrase)	22.0	23.8	22.9	40.6	7.3	12.4	43.8	4.0	7.3
Rule (official, entity)	30.5	28.2	29.3	62.5	11.3	19.1	53.1	4.8	8.8
ML (IR, entity)	-	-	-	26.6	5.7	9.4	5.47	12.5	7.61
ML (official, phrase)	-	-	-	15.0	1.7	3.1	15.5	4.8	7.3
ML (official, entity)	-	-	-	37.5	4.2	7.6	23.6	7.3	11.2
TRAINS									
Rule (IR, entity)	48.9	48.9	48.9	66.7	36.0	46.8	27.1	21.8	24.2
Rule (official, phrase)	41.7	47.8	41.7	58.0	32.4	41.6	28.4	11.3	16.2
Rule (official, entity)	47.5	47.3	47.4	64.4	36.0	46.2	28.4	11.3	16.2
ML (IR, entity)	-	-	-	56.6	23.6	33.3	10.3	14.6	12.1
ML (official, phrase)	-	-	-	58.8	11.9	19.8	17.4	10.1	12.8
ML (official, entity)	-	-	-	63.2	12.8	21.3	19.0	11.0	13.9

Table 9: Roesiger's results on Task 2 for all domains.