Natural Language Engineering (2020), **26**, pp. 95–128
doi:10.1017/S1351324919000056

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU Corpus

Olga Uryupina[1,*], Ron Artstein[2], Antonella Bristot, Federica Cavicchio[3], Francesca Delogu[4], Kepa J. Rodriguez[5] and Massimo Poesio[6]

[1]Department of Information Engineering and Computer Science, University of Trento, [2]Institute for Creative Technologies, University of Southern California, [3]Sign Language Lab, University of Haifa, [4]Department of Computational Linguistics & Phonetics, Saarland University, [5]Archives Division, Yad Vashem and [6]School of Electronic Engineering and Computer Science, Queen Mary University of London
*Corresponding author. Email: uryupina@gmail.com

(Received 15 January 2017; revised 17 December 2018; accepted 18 December 2018; first published online 7 May 2019)

Abstract

This paper presents the second release of ARRAU, a multigenre corpus of anaphoric information created over 10 years to provide data for the next generation of coreference/anaphora resolution systems combining different types of linguistic and world knowledge with advanced discourse modeling supporting rich linguistic annotations. The distinguishing features of ARRAU include the following: treating all NPs as markables, including non-referring NPs, and annotating their (non-) referentiality status; distinguishing between several categories of non-referentiality and annotating non-anaphoric mentions; thorough annotation of markable boundaries (minimal/maximal spans, discontinuous markables); annotating a variety of mention attributes, ranging from morphosyntactic parameters to semantic category; annotating the genericity status of mentions; annotating a wide range of anaphoric relations, including bridging relations and discourse deixis; and, finally, annotating anaphoric ambiguity. The current version of the dataset contains 350K tokens and is publicly available from LDC. In this paper, we discuss in detail all the distinguishing features of the corpus, so far only partially presented in a number of conference and workshop papers, and we also discuss the development between the first release of ARRAU in 2008 and this second one.

Keywords: coreference; anaphora; discourse; annotation; linguistic corpora

1. Introduction

A great number of data-driven approaches to anaphora resolution (also known in NLP as coreference) have recently been proposed, considerably pushing forward the state of the art in the field (see, e.g., Durrett and Klein 2013; Lee *et al.* 2013; Björkelund and Kuhn 2014; Fernandes *et al.* 2014; Martschat and Strube 2015; Clark and Manning 2016; Lee *et al.* 2017; see also Pradhan *et al.* 2012 for a comparative analysis of some of these systems). A key reason for these advances has been the creation of larger and more linguistically motivated gold annotated corpora, and in particular of OntoNotes Weischedel *et al.* (2011), and the success of recent evaluation campaigns using these new resources (Recasens *et al.* 2010; Pradhan *et al.* 2011, 2012). Most of the recently proposed approaches, however, still focus on the accurate modeling of relatively easy cases of anaphoric reference. For example, Durrett and Klein (2013) build one of the best-performing system through extensive feature engineering for "easy victories," avoiding "uphill battles" for more complex cases. This can be explained by (i) the still relative simplicity of the OntoNotes annotation scheme and (ii) the intrinsic difficulty of the task once we go beyond "easy victories." We believe that

the time is ripe for a dataset that better approximates the true complexity of the phenomenon of anaphoric reference. Such datasets now exist for languages other than English—for example, ANCORA for Catalan and Spanish (Recasens and Martí 2010), the Prague Dependency Treebank for Czech (Nedoluzhko *et al.* 2009a), or TÜBA-D/Z for German (Hinrichs *et al.* 2005)—but not yet for English.

This paper presents the second release of the ARRAU corpus,[a] a multigenre corpus of English providing large-scale annotations of a broad range of anaphoric phenomena and of linguistic information relevant to anaphora resolution. ARRAU has been under development for over 10 years, and several features distinguish it from similar projects.

First, it supports a more complex and linguistically motivated annotation scheme for anaphora than any existing corpus for English and than most corpora for other languages, covering, for example, non-referring expressions, bridging references, and discourse deixis. Moreover, additional discourse-level information is available from third parties for subsets of ARRAU (e.g., the rhetorical structure annotations Carlson *et al.* 2002 for the `rst` domain). This enables a more thorough analysis of these phenomena, as well as creates training material for algorithms that model these tasks jointly.

Second, the ARRAU guidelines specify the annotation of a number of semantic properties of mentions, most importantly of genericity. Identifying generic usages of nominal expressions is still an understudied task, and we believe that the release of a corpus annotated simultaneously for anaphora and genericity can provide much needed data.

Third, the corpus covers, in addition to news, a variety of genres so far poorly studied, such as dialog (the TRAINS data) and fiction (the Pear Stories). Spontaneous dialog and fiction are not covered by most commonly used coreference corpora.[b] Although several linguistic studies focus on genre-specific discourse coherence and anaphora properties (Neumann 2013; Kunz and Lapshinova-Koltunski 2015), only very few approaches aim at empirical analysis or per-genre modeling of coreference (Uryupina and Poesio 2012; Grishina and Stede 2015). In a recent work, Kunz *et al.* (2016) provide a comprehensive data-driven analysis of different linguistic phenomena related to anaphoricity, demonstrating considerable genre-specific differences. We believe that anaphora, among many other discourse-related phenomena, can bring a lot of challenging genre-specific problems and the ARRAU corpus opens up numerous research paths in this direction.

Fourth, anaphoric ambiguity is annotated. Ambiguous anaphoric expressions constitute truly challenging examples that cannot be tackled with current methods for coreference resolution. Moreover, the most commonly used corpora (Doddington *et al.* 2004; Weischedel *et al.* 2011) only focus on *identity* anaphora—the task of identifying multiple mentions of the same discourse entity—and thus cannot support anaphoric ambiguity. By annotating ambiguous anaphoric expressions, we make the first step toward a thorough investigation of anaphoric ambiguity.

Finally, during the 10 years in which the ARRAU dataset has been under development, we have had the opportunity not only to extend the annotation and the size of the corpus, but also crucially to continuously revise the annotation and improve its quality. In this paper, we describe the second major release of the corpus, whose development has been motivated not only by the objective of increasing the corpus size, particularly regarding spoken data, but also by improving the annotation quality and consistency in a number of ways, including via several automatic consistency checks. This is in contrast with other corpora, where subsequent releases, if any, expand the text collection and only fix occasional manually attested errors. We believe that the computational linguistic community can benefit considerably from cleaner and more curated datasets. This implies a methodology for data cleaning and maintenance that is currently in its infancy, with only very

[a] http://www.arrauproject.org.

[b] ONTONOTES contains dialog documents, with the speakers annotated manually. However, the ONTONOTES dialogs come from a curated broadcasting setting and therefore are less spontaneous and exhibit fewer dialog-specific features, such as disfluencies and incorrect/unfinished sentences, references to the visual context, and so on.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

few exceptional studies (e.g., Dickinson and Meurers 2003; Dickinson and Lee 2008) investigating possibilities for automatic error identification in manually annotated resources. Moreover, the few existing efforts are not supported by the data creation/labeling projects: to our knowledge, the common practice in annotating textual data does not go beyond ensuring high agreement between human coders, using, for example, κ or Krippendorf's α (Carletta 1996; Artstein and Poesio 2008). Corpus creators rarely make use of automatic means of data verification, such as specific consistency checks or error analysis for automatic systems trained and tested on the data. While our approach is far from being the final word on this, we think it makes a first step in the right direction.

The two versions of the ARRAU corpus were presented at the Language Resources and Evaluation conference (Poesio and Artstein 2008; Uryupina *et al.* 2016a), but this paper greatly expands upon the content of these two LREC papers, providing an extensive overview of the annotation guidelines and their motivation and a range of previously unpublished statistics about the linguistically more advanced features of ARRAU.

The second release of the ARRAU corpus, in MMAX2 format and including the original annotations of the Penn Treebank from which the markables[c] were extracted, is available from LDC, but the sub-corpora of this version of ARRAU that consist of anaphoric annotations of LDC corpora such as the RST Discourse Treebank and the TRAINS-93 corpus can only be distributed for free to groups that acquire a license for the original corpora. However, the dataset extracted from ARRAU for the CRAC 2018 Shared Task (see Section 5.3) is freely available through LDC.

The rest of the paper is organized as follows. Section 2 provides an overview of the annotation guidelines. Section 3 discusses the corpus development between the two versions. Finally, Section 4 compares ARRAU against other datasets annotated for coreference.

2. Annotation methodology

The goal of the ARRAU project was to develop methods to annotate and interpret the more challenging cases of anaphoric reference, including in particular reference to abstract objects. A key aspect of this work was to use coding schemes based on extensive reliability tests to create large-scale annotated resources that could be used to study these types of anaphoric reference. Building on the GNOME guidelines (Poesio 2000a, as discussed in Poesio 2004a,b) which already provided reliability-tested annotation schemes for aspects of anaphoric annotation such as bridging reference (Clark 1975) and were used, e.g., to create the dataset in Poesio *et al.* (2004a,b), we developed and tested extended annotation guidelines (Poesio and Artstein 2008) aiming specifically at abstract anaphora and ambiguity (Poesio and Artstein 2005a; Artstein and Poesio 2006). These annotation guidelines, distributed with the corpus and available from the project website, also provide detailed instructions for identifying markable boundaries and marking non-referentiality and non-anaphoricity, as well as a wide range of mention attributes such as genericity. In this section, we summarize these guidelines and more in general the methods adopted in the creation of the corpus, focusing on the most distinctive features of the ARRAU annotation.[d]

[c]Ever since the ACE evaluation campaigns, the term *mention* has been used to indicate the items to be classified in anaphora resolution/coreference. This terminology is appropriate for corpora such as ACE or ONTONOTES, in which indeed such items are always mentions of discourse entities. ARRAU however is designed to support the complete anaphora resolution task, also known as end-to-end coreference resolution, in which the task of discriminating non-referring from referring NPs (mentions) is not separated from the task of interpreting referring NPs, so that non-referring NPs are marked as well. For this reason, we will use in this paper the term *markable* to refer to the overall set of items annotated, reserving the term *mentions* for referring items.

[d]The term *mention* has become established in the literature on anaphoric annotation to refer to markables, whether referring or non-referring, although strictly speaking only referring expressions could be called mentions, as a markable can only be a mention *of* a discourse entity. We will stick to this terminology here even though ARRAU's markables include both referring and non-referring expressions.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 1. Corpus statistics for the four ARRAU domains

	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
Avg. markable length (tok)	4.1	4.0	2.2	1.8
Discontinuous markables	864 (1.2%)	175 (2.7%)	3 (0%)	15 (0%)
One-word markables	21461 (30%)	2338 (35.6%)	2164 (54.0%)	9404 (55.3%)
Non-referring markables	9552 (13.3%)	1047 (16.0%)	607 (15.1%)	2353 (13.8%)
Generic mentions	2793 (3.9%)	856 (13.0%)	122 (3.0%)	3077 (18.1%)

2.1 Genres

Some of the best known anaphoric corpora, particularly for English and particularly at the time when the ARRAU annotation was started, consist entirely of documents either in the news or broadcast genres. One of the objectives of the ARRAU annotation was to cover a greater variety of genres.

The corpus does include a substantial amount of news text, a sub-corpus or *domain* (we will use throughout the term domain to refer to ARRAU's sub-corpora) called RST and consisting of the entire subset of the Penn Treebank that was annotated in the RST treebank (Carlson *et al.* 2003). We annotated news data so as researchers could compare results on ARRAU with results on other news datasets, and we chose these documents 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 the Propbank; Palmer *et al.* 2005) but also for rhetorical structure (Carlson *et al.* 2003). This dataset would therefore allow the study of the effect of these other types of linguistic information on anaphora resolution and vice versa.[e]

But in addition to RST, ARRAU includes three more domains, covering genres important from the point of view of discourse analysis but not normally covered by anaphoric corpora. Specifically, the TRAINS domain of ARRAU includes all the task-oriented dialogs in the TRAINS-93 corpus[f]; the PEAR domain consists of the 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 domain covers documents from the medical and art history genres covered by the GNOME corpus (Poesio 2004a, 2000b) used to study both local and global salience (Poesio *et al.* 2004a, 2006a).

The same coding scheme was used for all domains, but separate guidelines were written for the textual domains and the spoken dialog domains; the distinct coding schemes are included in the documentation of the corpus as `man_anno_gnome` and `man_anno_trains`, respectively.

Table 1 provides basic statistics about the four ARRAU domains.[g] Both the RST and GNOME domains consist of carefully edited texts with complex grammatical sentences. This results in long markables, often either multiword named entities (for example, full names of organizations) or

[e]This annotation took place in collaboration with, although independently from, the annotation of the same data carried out by Kibrik's group at the Russian Academy of Sciences (Loukachevitch *et al.* 2011).

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

[g]All the statistics provided in this Section are for the second release of ARRAU.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

complex NPs. Markable detection for these domains requires a high-quality parser. Particularly in the GNOME domain, synonyms and bridging references abound. Successful interpretation and resolution of such expressions would require sophisticated name-matching and aliasing techniques and advanced semantic features, going beyond head-noun compatibility.

The PEAR and TRAINS domains, by contrast, consist of uses of spontaneous speech. The language in these domains mostly consists of short utterances, often ungrammatical and/or with disfluencies. PEAR and TRAINS markables therefore are on average much shorter, with a lot of one-word markables, mostly pronouns. Discontinuous markables (see Section 2.2) are present in both PEAR and TRAINS, although not very common. So for these domains, markable detection might better be implemented through a chunker robust to noisy ungrammatical input. As far as anaphora resolution is concerned, however, ambiguity and references to abstract objects (e.g., plans in TRAINS) abound, as well as demonstratives used deictically. So salience features and context modeling become key factors.

To summarize, ARRAU contains documents from four domains, representing different genres, mostly not covered by other corpora. These genres pose challenging problems for the next generation of coreference resolvers, requiring complex techniques for accurate preprocessing and resolution.

2.2 Markables in ARRAU

ARRAU belongs to the "new wave" of anaphorically annotated corpora that were created after the re-examination of annotation schemes for anaphora started with the Discourse Resource Initiative and the MATE and GNOME projects (Passonneau 1997; Poesio *et al.* 1999; van Deemter and Kibble 2000). These new corpora—other examples include ANCORA (Recasens and Martí 2010), COREA (Hendrickx *et al.* 2008), and ONTONOTES (Pradhan *et al.* 2007), the anaphoric annotation of the Prague Dependency Treebank (Nedoluzhko *et al.* 2009a) and TÜBA-D/Z (Hinrichs *et al.* 2005)—employed annotation schemes rooted in linguistic theory rather than aiming to capture domain-relevant knowledge as done in the earlier MUC and ACE corpora; for instance, the entire NP is typically marked. Not all of these corpora however consider all NPs as markables. Some older corpora had imposed syntactic restrictions on markables—for instance, in many older corpora only pronouns are annotated (Ge *et al.* 1998). Other older corpora imposed semantic restrictions: for instance, in the ACE corpora, only entities of semantic types of interest are considered. But even some of the "new generation" corpora still restrict mentions depending on their referentiality/anaphoricity properties: 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, according to the ARRAU guidelines (which follow for text the earlier GNOME guidelines,[h] see below for the dialog guidelines) *all* NPs are considered as markables, also when they are non-referring, like predicative *a busy place* in (1) (we discuss in Section 2.3 which NPs are considered non-referring in ARRAU), or when they do not corefer with any other mention and thus form a singleton coreference chain all by themselves. Moreover, non-referring markables are manually sub-classified. In addition, possessive pronouns are marked as well, and all premodifiers are marked when the entity referred to is mentioned again, for example, 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 $[US]_1$ trade deficit may worsen next year after two years of significant improvement…The statement was the $[US]_1$'s government first acknowledgment of what other groups, such as the International Monetary Fund, have been predicting for months.

[h] http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

(3) 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 …

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` corresponds to the head noun, whereas for (modified or not) named entities `min` corresponds to the proper name:

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

Discontinuous markables. One of the distinctive features of ARRAU is the support of *discontinuous markables*—markables built out of non-continuous material. Discontinuous chunks are problematic for many corpus annotation formats Amoia *et al.* (2011), and thus many guidelines developed for various linguistic phenomena allow for labeling continuous constituents exclusively.

Discontinuous markables, however, are common in dialog, for instance, in cases of so-called *collaborative completions* (Poesio and Rieser 2010) illustrated by (5), where the mention *an orange screw with a slit* is constructed out of utterances 1.2 and 1.3.

(5)

1.1	Inst	So, jetzt nimmst Du [pause]
		Well, now you take
1.2	Cnst	eine Schraube
		a screw
1.3	Inst	eine < − > orangene mit einem Schlitz.
		an < − > orange one with a slit

For this reason, Müller included the functionality for the annotation of discontinuous markables in the MMAX2 annotation tool (Müller and Strube 2006), developed to support his research on anaphora resolution in dialog (Müller 2008), where spans can be arbitrary sequences of tokens. However, discontinuous markables also provide a way to include in a markable all information provided by the text, for example, in cases of coordination where the two coordinated NPs share some information, illustrated by (6). In this example, the two names *Anna Snezak* and *Morris Snezak* are coordinated, but the last name *Snezak* is only repeated once. Discontinuous markables make it possible to include both the segments of text marked as part 1 and part 2 in the same markable. Similarly in (7).

(6) ..after owners [^{part1}Anna]^{part1} and Morris [^{part2}Snezak]^{part2}..

(7) So he doesn't have to play [^{part1}the same Mozart]^{part1} and Strauss [^{part2}concertos]^{part2} over and over again.

Discontinuous markables are typically ignored in anaphora resolution: state-of-the-art mention detection systems always output continuous chunks; the publicly available SemEval and CONLL coreference scorers (Pradhan *et al.* 2014) assume numbered brackets as mention boundaries that cannot encode discontinuous fragments. To make ARRAU usable for these purposes, whereas the markable can be discontinuous, *minimal spans* cannot be. This way, all the markables in ARRAU can be aligned to contiguous sequences of tokens.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

2.3 Markable properties

All markables are manually annotated for a variety of properties according to the GNOME guidelines (Poesio 2000a): these include morphosyntactic agreement (gender, number and person), grammatical function, and the semantic type of the entity: `person`, `animate`, `concrete`, `organization`, `space`, `time`, `plan` (for actions), `numerical`, or `abstract`.[i] The guidelines and reliability studies leading to this scheme are discussed in Poesio (2004b, 2000b). In this section, we will only discuss in detail two additional attributes, specifying the referential status of a markable and the genericity status of mentions. The `reference` attribute specifies the logical form status of the markable: referring, expletive, quantificational, or predicative. `Genericity` is annotated following a scheme developed in GNOME after experiments based on the official annotation manual had shown poor reliability for this attribute. We discuss each attribute in turn.

Referring and non-referring markables. Most anaphorically annotated corpora focus on *referring* markables, or mentions proper: markables that refer to discourse entities and participate in anaphoric relations. This decision, primarily motivated by reasons of cost, makes it however difficult to train models able to recognize and interpret *non-referring* markables—nominal expressions that do not refer a discourse entity. It has been shown, however, that filtering out at least some types of non-referring expressions can improve the performance of a coreference resolver (Uryupina *et al.* 2016b). In order to develop such a classifier for a corpus like ONTONOTES in which non-referring expressions are not annotated, separate classifiers are required—for example, Björkelund and Farkas (2012) trained a pre-filtering classifier for non-anaphoric *it*, *you* and *we* on the ONTONOTES data.

In ARRAU, all nominal expressions are treated as markables, including non-referring nominal expressions. The annotation scheme and guidelines are based on those developed for GNOME, where the `lftype` attribute ($\kappa = .73$) was used to distinguish between referring expressions proper (called `terms` in GNOME) and several types of non-referring interpretations of NPs, including expletives (8), predicatives (9,10), quantifiers (11) and coordinations (12) Poesio (2000a, 2004b). In ARRAU, coders are asked, first of all to classify markables as referring or non-referring. If a markable is classified as referring, coders are then asked if that expression is `discourse-old` or `discourse-new` (Prince 1992), and in the first case, to identify its antecedent (see Section 2.4). If the markable is classified as non-referring, coders have to either assign it to one of the GNOME categories of non-reference, or label it as idiomatic (13), or as an incomplete or fragmentary expressions (14).

(8) And [there]$^{non-referring}$'s a ladder coming out o of the tree
and [there]$^{non-referring}$'s a man at the top of the ladder

(9) It see it seems to be [a busy place]$^{non-referring}$

(10) 1 ml of the prepared solution for injection contains 0.25 mg ([8 million IU]$^{non-referring}$) of Interferon beta-1b.

(11) [Most of the analysts polled last week by Dow Jones International News Service in Frankfurt, Tokyo, London and New York]$^{non-referring}$ expect the US dollar to ease only mildly in November.

(12) Mr. Sutton recalls: "When I left, I sat down with [[Charlie Rangel], [Basil Paterson] and [David]]$^{non-referring}$, and David said, 'Who will run for borough president?

(13) so that would um if we left at six in the morning would that make [sense]$^{non-referring}$ six (mumble)

[i]The `category` attribute encoding this information is a merge of two separate attributes in the GNOME scheme: `ani` for animacy and `onto` for ontological type.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 2. Distribution of non-referring markables in ARRAU

	RST	TRAINS	GNOME	PEAR
All markables	72013	16999	6562	4008
Non-referring	9552	2353	1047	607
Expletive	444	851	75	122
Predicate	4311	145	355	79
Idiom	638	148	29	42
Coordination	2410	232	327	37
Incomplete	2	149	1	36
Quantifier	1738	818	259	132
Unknown	9	6	1	159

(14) U: okay then um okay then originally we need to have um the one boxcar go to [oranum]$^{non-referring}$ go to Corning from Elmira

The choice of marking quantifiers and coordination in ARRAU as non-referring is possibly the most controversial decision we took. The quantifier *Most of the analysts polled last week by Dow Jones international* in (11) is marked as non-referring. Similarly, whereas we asked coders to mark individual noun phrases (*Charlie Rangel*, *Basil Paterson* and *David* in (12)) as referring markables that can participate in anaphoric relations, the embedding coordinate NP is marked as non-referring. These decisions mean that any expression anaphorically related to that quantifier cannot be marked as such. However, plural anaphora to antecedents introduced by coordination can be annotated, as discussed in Section 2.4. In the case of quantifiers, the decision was motivated by the high disagreement that we observed among our coders when left free to mark a quantifier as either referring or non-referring. For the case of coordination, the reasons were more complex; we discuss them when explaining how plural anaphora is handled. Both decisions might be reconsidered in a future release of ARRAU.

Table 2 shows the distribution of various types of non-referring markables in the entire corpus and in the four individual domains, overall and for each type of non-referring markables. As could be expected, the distribution of non-referring expressions is genre-specific. Thus, the two domains with spontaneously generated no-curated texts (TRAINS and PEAR) have a large number of incomplete or fragmentary expressions, virtually non-existent in RST and GNOME documents. Idioms are common in all the genres except GNOME—a collection of medical leaflets written in a very formal language. Predicative non-referring expressions, especially appositions, are more common in news.

Genericity. The guidelines for genericity adopted for the GNOME corpus were developed to distinguish generic uses of nominal expressions (as in *Dogs bark*) from non-generic cases (as in *I saw dogs in the street*). Developing reliable guidelines for this type of annotation proved quite a challenge, and two schemes were conceived before developing one achieving sufficient reliability. The first scheme attempted to capture the type / token distinction—a similar distinction to that between generic and specific entities made in the ACE-2 coding scheme—but this type of judgment proved difficult to agree on in particular with mentions referring to substances such as *oil* or chemical components of medicines such as *oestradiol*, as illustrated in (15). The result was that this simple scheme only achieved a very modest level of reliability ($\kappa = .33$).

(15) Not that [oil]^{generic} suddenly is a sure thing again.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

A second scheme was then developed in which a new value, `undersp-generic`, was introduced as the value to be used for all references to substances.[j] The new scheme achieved better reliability, but still only $\kappa = 0.55$. The biggest remaining problem were quantifiers (including definites and indefinites). Our annotators found it very hard to agree on whether a quantified NP used (non-generically) to quantify over a specific set of individuals at a particular spatio-temporal location, as in *Many lecturers went on strike (on March 16th, 2004)*, should be marked as generic or not. A third and last scheme was therefore developed, in which separate values were introduced for each type of quantifier, as well as new guidelines, according to which the annotation of the genericity attribute is carried out following a decision tree going from the easiest cases to the more complex ones. Coders are first asked to check whether the nominal is in the scope of an *explicit* operator such as a conditional like *if* (as in (16)) or an individual quantifier such as *every* or *most* (`iquant`) (as in (17)) or a temporal quantifier like *always* or *once* (as in (18)) a modal (as in (19)) or an instruction (as in (20)). 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 (21) seen before or in (22). 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 (23), or `generic-no` otherwise. With these instructions, reasonable intercoder agreement was finally achieved ($\kappa = .82$) (Poesio 2004b).

(16) 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]^{generic}.
(`operator-conditional`)

(17) Mr. Uhr said that Mr. Petrie or his company have been accumulating Deb Shops stock for several years, each time issuing [a similar regulatory statement]^{generic}.
(`operator-iquant`)

(18) In addition , once [money]^{generic} is raised , [investors]^{generic} usually have no way of knowing how [it]^{generic} is spent.
(`operator-tquant`)

(19) They argue that their own languages should have [equal weight]^{generic}, although recent surveys indicate that the majority of the country's population understands Filipino more than any other language.
(`operator-modal`)

(20) Use [alcohol wipes]^{generic} to clean the tops of the vials move in one direction and use one wipe per vial.
(`operator-instruction`)

(21) Not that [oil]^{generic} suddenly is a sure thing again .
(`underspecified-substance`, `RST`)

(22) 1 ml of [the prepared solution for injection]^{generic} contains 0.25 mg (8 million IU) of [Interferon beta-1b]^{generic}.
`underspecified-substance`, `GNOME`)

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

Genericity was already marked according to these guidelines in the first release of ARRAU (Poesio and Artstein 2008), but its annotation was only partially checked. One of the main revisions

[j] This is the scheme described in the best-known version of the GNOME manual, version 4 from April 2000.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 3. Distribution of generic mentions in ARRAU

	RST	TRAINS	GNOME	PEAR	OVERALL
All	72013	16999	6562	4008	99582
Generic-yes	1438	728	12	74	2252 (2%)
Operator-conditional	90	231	201	2	524
Operator-instruction	15	163	211	–	389
Operator-iquant	7	6	–	–	13
Operator-modal	443	1080	147	16	1686
Operator-question	54	432	39	10	535
Operator-tquant	16	4	–	–	20
Total operator bound					3167 (3%)
Underspecified-disease	–	–	84	–	84
Underspecified-replicable	37	1	2	21	61
Underspecified-substance	692	431	160	–	1283
Underspecified-generic	1	3	–	–	4
Total underspecified					1432 (1.4%)

carried out for the second release of the corpus was a systematic check that the annotation of this attribute was consistent with the guidelines. The distribution of generics and quantifiers in the separate ARRAU domains resulting from this verification is shown in Table 3. In total 2252 mentions were annotated as generic (2% of the total number of markables), 3167 as being bound by some other operator (3%), and 1.4% as underspecified.

2.4 Range of relations

The ARRAU guidelines support annotation of different types of anaphoric relations. All referring markables are marked as either `discourse new` or `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` (in case the antecedent was introduced using a nominal expression) or `segment` (not introduced by a nominal expression, for the cases of *discourse deixis*).[k] In addition, referring NPs can be marked as *related* to a previously mentioned discourse entity in order to identify them as examples of associative or *bridging* anaphora. We discuss the three most distinctive types of annotation in ARRAU—bridging anaphora, plural anaphora, and discourse deixis—in turn.

Bridging anaphora. Annotating—indeed, identifying—bridging anaphora in a reliable way is a difficult task (Poesio and Vieira 1998; Vieira 1998), which is one of the reasons why so few large-scale corpora for anaphora include this type of annotation (apart from our own work, we are only aware of few attempts to do so; see Section 4.4 for a discussion of this work and (Poesio *et al.* 2016) for additional discussion of larger corpora some of which also include anaphora). The ARRAU guidelines for bridging anaphora are based on a series of experiments that started with the work of

[k]Identity anaphora also includes plural anaphoric reference to entities introduced via plural mentions, as in *We need to put the pizzas in the oven else they will get cold*, as opposed to plural reference to antecedents introduced by distinct singular mentions, which is annotated as a form of bridging reference, as discussed above.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Vieira (1998) and Poesio and Vieira (1998) and continued in the GNOME project (Poesio 2004b). Vieira and Poesio attempted to annotate the full range of bridging references as discussed, for example, in Passonneau (1997) and Poesio *et al.* (1999), but only achieved very poor agreement. In GNOME, attempts were made to identify a subset of the relations that could be annotated reliably (Poesio 2004b), finding that by limiting the annotation to three types of relations: `element-of` as in (24), where *the middle* is a bridging reference to the middle of the three horizontal zones; `subset` as in (25), where *Polygonal openwork rings incorporating an inscription* in (u2) is a bridging reference to *two gold finger rings* in (u1) based on an inverse subset relation; and a generalized possession relation `poss` covering both part-of relations as in (26) and general possession relations, as in (27). The `element` relation was also used to annotate certain types of *other* anaphora, as in (28).

(24) The sixteen panels are each divided into [three horizontal zones]$_1$, [the middle]$_{\rightarrow 1}$ containing a letter

(25) (u1) [Two gold finger-rings from Roman Britain (2nd–3rd century AD)]$_1$.
(u2) [Polygonal openwork rings incorporating an inscription]$_{\rightarrow 1}$ are a distinctive type found throughout the Empire.

(26) (u1) [These "egg vases"]$_1$ are of exceptional quality
(u2) basketwork bases support [egg-shaped bodies]$_{\rightarrow 1}$
(u3) and bundles of straw form [the handles]$_{\rightarrow 1}$

(27) (u1) [The Getty museums microscope]$_1$ still works,
(u2) and [the case]$_{\rightarrow 1}$ is fitted with a drawer filled with the necessary attachments.

(28) (u39) [The two stands]$_1$ are of the same date as the coffers, but were originally designed to hold rectangular cabinets.
(u42) [One stand]$_{\rightarrow 1}$ was adapted in the late 1700s or early 1800s century to make it the same height as [the other]]$_{\rightarrow 1}$.

Poesio *et al.* found that coders following the GNOME guidelines achieved good precision but low recall on identifying bridging references (Poesio 2004b). When asked to mark mentions as either discourse-new, discourse-old, or bridging according to the GNOME definition of bridging, coders agreed on the type of relation for bridging references in 95.2% of the cases, but each of them only spotted about 1/3 of bridging references on average, and typically different bridging references, so that only 22% of bridging references were marked as such by all annotators.

The ARRAU Release 1 guidelines followed the GNOME guidelines, but with an extension and a simplification. Annotators were asked to mark a mention 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 Modieska (2003);
- an `undersp-rel` relation for "obvious cases of bridging that didn't fit any other category."

In ARRAU Release 1, however, 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. The resulting statistics about bridging references in ARRAU Version 2 are shown in Table 4. 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

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 4. Distribution of bridging references in ARRAU

	RST	TRAINS	GNOME	PEAR	TOTAL
All	3777	710	692	333	5512
Poss	87				≥ 87
Poss-inv	25				≥ 25
Subset	1092				≥ 1092
Subset-inv	368				≥ 368
Element	1126				≥ 1126
Element-inv	152				≥ 152
Other	332				≥ 332
Other-inv	7				≥ 7
Undersp-rel	588				≥ 588
P+S+E+O+U	N/A	710	692	333	1735

the other domains, currently not classified. We intend to provide a classification of these bridging references, as well as re-checking the existing classifications, in Release 3 of the corpus, currently planned for 2018.

Plural anaphora. Till recently, no data-driven studies were attempting to model plural anaphora specifically, except for the simplest cases of plural reference to a plural antecedent, as in (29).[1]

(29) (u1) from Avon going to Dansville pick up [the three boxcars]$_1$
(u2) go to Corning load [them]$_1$ and ...

This is because some types of plural reference are intrinsically difficult, both for annotation and resolution. We believe therefore that a dataset annotated for plural anaphora in a principled way will open several challenging research possibilities.

One example of the more complex forms of plural anaphora is plural reference to sets of objects introduced by listing their elements, as in the following toy examples.

(30) a. [*Mr. Luzon* and *his team*]$_1$, however, say [they]$_1$ aren't interested in a merger.
b. *Mr. Luzon* agreed with *his team* that [they]$_?$ aren't interested in a merger.

Anaphoric annotation schemes that do require coders to mark plural reference to antecedents introduced by coordination do so by assuming that the coordination *Mr. Luzon and his team* in (30a) (an actual example from the RST portion of ARRAU) introduces a discourse entity, and asking coders to link *they* to that entity. Indeed, this is the approach that was followed in GNOME. This approach will not however work for the very similar (30b) (our own), since in this example there is no longer a constituent for *Mr. Luzon and his team*—so *they* becomes a discourse new mention with no antecedent. The approach to annotating plurals adopted in ARRAU was based on the belief that these two very similar cases of plural reference should be treated in the same way. In ARRAU, we annotate plural anaphors to sets of individually introduced entities as bridging references to each member of the corresponding set encoding an (element-of) bridging relation. Thus, in (30a)

[1]One exception is a very recent study (Burga *et al.* 2016), which aimed at rule-based plural anaphora resolution for the patent domain.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 5. Distribution of discourse deixis in the subdomains of ARRAU

RST	TRAINS	GNOME	PEAR	TOTAL
631	862	73	67	1633

as well as in (30b) "They" is linked to both "Mr. Luzon" and "his team" individually. Note that such annotation allows for a more uniform interpretation of plural reference to individually introduced entities.

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

(31)
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 in its full form is a very complex form of reference, both to annotate (Artstein and Poesio 2006; Dipper and Zinsmeister 2012) and to resolve (Marasović *et al.* 2017). Very few anaphoric annotation projects have attempted annotating discourse deixis in its entirety (Artstein and Poesio 2006; Dipper and Zinsmeister 2012; Kolhatkar 2014); more typical is a partial annotation, as in the work of Byron and Navarretta, who annotated pronominal reference to abstract objects (Byron and Allen 1998; Navarretta 2000); in ONTONOTES, where event anaphora was marked (Pradhan *et al.* 2007); and in the work of Kolhatkar Kolhatkar (2014), which focused on so-called shell nouns. As a result, very few systems have attempted resolving this type of anaphors (Eckert and Strube 2000; Byron 2002; Kolhatkar and Hirst 2012; Marasović *et al.* 2017)

Discourse deixis was one of the "difficult cases of anaphora" on which the ARRAU project focused, and a number of annotation experiments were conducted (Artstein and Poesio 2006), resulting in guidelines according to which

1. A coder specifying that a referring expression is discourse old is asked whether its antecedent was introduced using a `phrase` (mention) or `segment` (discourse segment)
2. Coders choosing `segment` as the type of antecedent have to mark a sequence of (predefined) clauses

Artstein and Poesio (2006) point out that measuring disagreement on this type of annotation requires making a number of assumptions, and that figures of α Krippendorf (2004) ranging from 0.45 to 0.9 can be achieved depending on which assumptions are made.

The statistics about discourse deixis in ARRAU Version 2 are shown in Table 5. A total of 1633 cases of discourse deixis were identified. It is worth noting how the TRAINS sub-domain contains more than half the total cases of discourse deixis even though it is less than half the size of the RST sub-domain. (We intend to re-check the annotation in Release 3 of the corpus, currently planned for 2018.)

[m] For more extensive discussion of reference to abstract objects, see Schuster (1988) and Asher (1993); for empirical analysis of discourse deixis, see, for example, Gundel *et al.* (2003).

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Anaphoric ambiguity. A number of studies have shown that anaphoric expressions both in dialog and text can be ambiguous (Poesio and Reyle 2001; Poesio *et al.* 2006b; Versley 2008; Recasens *et al.* 2011). A classic illustration is Example (32), from the TRAINS corpus (Poesio and Reyle 2001). The pronoun *it* in (u2) could refer equally well to *engine E2* or *the boxcar at Elmira*. Studies carried out as part of ARRAU showed that such examples were fairly common in the TRAINS corpus, and that different coders would interpret them differently (Poesio and Artstein 2005a; Poesio *et al.* 2006b). Other studies have shown that occurrences of *it* can be ambiguous between an expletive and a discourse deixis interpretation (Gundel *et al.* 2002).

(32) (u1) M: can we .. kindly hook up ... uh ... [engine E2]$_1$ to [the boxcar at .. Elmira]$_2$
(u2) M: +and+ send [it]$_{1,2}$ to Corning as soon as possible please

The ARRAU coding scheme accommodates this. Referring markables can be marked as ambiguous between a discourse-new and a discourse-old interpretation; discourse-old mentions can be marked as ambiguous between a discourse-deictic and a `phrase` reading; and both `phrase` and `segment` mentions can be marked as ambiguous between two distinct interpretations. The annotated corpus contains examples of ambiguous anaphoric expressions from text as well, as in the following example.

(33) Criticism of [the Abbie Hoffman segment]$_1$ is particularly scathing among people who knew and loved the man. <...> Both women say they also find it distasteful that [CBS News is apparently concentrating on Mr. Hoffman's problems as a manic-depressive]$_2$. "[This]$_{1,2}$ is dangerous and misrepresents Abbie's life," says Ms. Lawrenson, who has had an advance look at the 36-page script .

In (33), the anaphoric mention "This" is ambiguous between "the Abbie Hoffman segment" (identity anaphora) and "CBS News is apparently concentrating on Mr. Hoffman's problems as a manic-depressive" (discourse deixis).

The extent of ambiguity in anaphoric interpretation found using the ARRAU scheme was analyzed in a study reported in Poesio and Artstein (2005a). A total of 18 subjects were asked to annotate dialogs from the TRAINS subdomain of ARRAU with a scheme allowing them to mark for ambiguity. Poesio and Artstein reported that a minimum of 10% of markables in the TRAINS corpus were marked as explicitly ambiguous. They also found however that a much higher percentage of markables, up to 40%, were *implicitly* ambiguous—i.e., were annotated differently by different subjects. In Poesio and Artstein (2005b) methods for computing agreement in a scheme allowing for ambiguity were proposed, based on developing extended distance metrics for α (Krippendorf 2004; Artstein and Poesio 2008). Values of α between .58 and .67 were reported depending on the type of distance metric used and the choice of markables.

Statistics about anaphoric ambiguity in ARRAU Version 2 can be found in Table 6. The first column of the table shows the category of the first interpretation of the ambiguous markable: discourse old (either `phrase` or `segment`), discourse new, or non-referring. The second column shows the second interpretation indicated by the coder: again discourse old (`phrase`) but with a different antecedent, discourse new, discourse deixis, or non-referring. A total of 234 cases of ambiguous markables were identified, which is a very small fraction of the around 100,000 markables in ARRAU Version 2; the results of Poesio and Artstein (2005a) suggest however that this figure substantially underestimates the actual extent of ambiguity, at least by a factor of 4. The majority of these ambiguities (75%) are between two discourse old interpretations with different antecedents, but there are also several cases of DN/DO ambiguity and DO/DD ambiguity. We also note how in all cases of ambiguity the first interpretation chosen is discourse old; this is because the instructions explicitly require coders to choose `DO` as first interpretation if the ambiguity is between a discourse-old interpretation and some other interpretation.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 6. Distribution of ambiguity in the subdomains of ARRAU

1st int	2nd int	RST	TRAINS	GNOME	PEAR	TOTAL
DO	DO	31	112	4	28	175
	DN	37	4	2	1	44
	DD	8	1	0	2	11
	NR	0	4	0	0	4
DN	DO	0	0	0	0	0
	DN	0	0	0	0	0
	DD	0	0	0	0	0
	NR	0	0	0	0	0
NR	DO	0	0	0	0	0
	DN	0	0	0	0	0
	DD	0	0	0	0	0
	NR	0	0	0	0	0
Total		76	121	6	31	234

Table 7. Reliability of the several aspects of the ARRAU coding scheme

Attribute	Reliability measure	Reference
Markable attributes		
Identification of non-referring markables	$\kappa = .73$	Poesio (2004b)
Genericity	$\kappa = .82$	Poesio (2004b)
Complex anaphoric relations		
Bridging references	95% agreement	Poesio (2004b)
Discourse deixis	$.45 \leq \alpha \leq .9$	Artstein and Poesio (2006)
Ambiguity		
Anaphoric ambiguity	$\alpha = .67$	Poesio and Artstein (2005b)

2.5 Reliability of the coding scheme, summarized

Table 7 summarizes the reliability of the different aspects of the ARRAU coding scheme presented in this section.

2.6 Annotation tool and markup scheme

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. Because of the need to encode ambiguity and bridging references, anaphoric information is encoded using MMAX2 *pointers*, linking together pairs of mentions and specifying discourse relations between them. This is in contrast with commonly used (e.g., in the ONTONOTES scheme) set-based annotations, where each mention is only labeled with the id of the corresponding discourse entity and no relations are annotated. Note that set-based annotation for identity anaphora can be induced from such pointers in a straightforward way.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 8. Corpus statistics for two releases of ARRAU

	ARRAU1			ARRAU2		
Domain	Documents	Tokens	Markables	Documents	Tokens	Markables
RST	204	146512	45590	413	228901	72013
PEAR	20	14059	3881	20	14059	4008
GNOME	5	21599	6215	5	21458	6562
TRAINS	35	25783	5198	114	83654	16999
Total	264	184748	60884	552	348072	99582

3. From ARRAU 1 to ARRAU 2: Checking annotation consistency

The first release of ARRAU (Poesio and Artstein 2008) was made publicly available in 2008. The second release of ARRAU has augmented the corpus annotating all the documents available within the TRAINS and RST datasets. This has resulted in a significant increase in the data size. This quantitative improvement is extremely important for the TRAINS domain, since it provides a unique large collection of dialogs annotated with anaphoric information. More statistics for both releases of ARRAU are provided in Table 8.

Most importantly, between the two releases we have invested a considerable effort in enforcing the annotation consistency. We believe that a large and complex annotation project, such as ARRAU, undergoing several rounds of manual adjudication and revision, should implement specific measures for preserving and improving the data quality. Unfortunately, the NLP community does not pay enough attention to the data consistency issue beyond the inter-annotator agreement. A notable exception is a series of studies by Dickinson and Meurers (2003) and Boyd *et al.* (2008) on enforcing consistency in syntactic treebanks, as well as more recent approaches (Dickinson and Lee 2008; Hollenstein *et al.* 2016) on identifying errors in basic semantic annotation (predicate-argument structure, multi-word expressions, and super-sense tagging). These studies rely on corpus statistics (e.g., n-gram or production rule frequencies) to identify annotation anomalies. Differently from these studies, we assess the interaction between multiple annotation layers and derive constraints to identify inconsistencies and thus improve the overall labeling. A similar approach, albeit on a much smaller scale, has been adopted in Frank *et al.* (2012) for improving the labeling quality for automatic annotation of multiple NLP phenomena in a domain adaptation experiment.

In what follows, we describe our effort aimed at enforcing the formal consistency of the ARRAU data, in a hope to raise a discussion and make first steps in the direction of establishing good practice in this respect. The ARRAU scheme assumes simultaneous labeling of a variety of closely related phenomena, and therefore different parts of the mark-up can be used for deriving constraints for semi-automatic clean-up. For example, we can ensure that a non-referring markable is not marked as participating in a coreference chain. All the violating cases can be extracted automatically and then further checked and re-annotated manually. In a few cases, these constraints revealed intriguing cases of anaphoric expressions. Mostly, however, they have helped us identify and eliminate clear annotation errors.

3.1 Enforcing annotation consistency in ARRAU

A significant effort has been devoted to improving not only the quantity, but also the quality of the material annotated within the ARRAU project. To this end, we have implemented the following measures for the second release of the dataset:

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 9. Enforcing annotation quality: inconsistency statistics for the first release of ARRAU, most common types of errors

Error type	RST	GNOME	PEAR	TRAINS
Missing antecedent for anaphoric mentions	332	46	49	35
non-referring markables as an antecedent	205	15	6	10
semantic type mismatch	813	64	25	34

- Minimal and maximal spans, genericity and referentiality have been (re) annotated for all the documents. This enforces consistency across domains and allows for more principled cross-domain studies of the relevant phenomena. We have expanded our annotation of reference and genericity to all the domains, adopting a more principled approach. This has resulted in a more consistent annotation of reference: more than 10% of non-referring markables have been added to the documents already covered in ARRAU-1. For genericity, the first release only attempted a pilot annotation for the RST domain.
- All the unspecified attributes have been re-annotated.
- Morphological attributes have been checked across coreference chains. For example, a typical chain should not include two mentions of different gender. All the violating cases have been assessed manually.
- Semantic type has been checked for consistency across coreference chains.
- All the non-referring markables have been checked to exclude their participation in coreference chains. While the annotation scheme does not allow non-referentials to be anaphors, no MMAX2 functionality prevents a non-referring markable from being selected as an antecedent.
- All the mentions labeled as discourse-old have been assigned an antecedent.
- Basic bracketing constraints have been enforced: no nominal markables should intersect each other or sentence boundaries.

The result of this effort has been two-fold. On the one hand, we have identified and removed various typos and inconsistencies that inevitably arise as a result of manual annotation. Table 9 shows the number of problematic cases for the three most common types of errors. Most of these cases are plain annotation mistakes: sometimes an incorrect labeling is introduced at the initial annotation stage; more often, however, the errors are by-products of post-corrections, either by the supervisor or by the annotators themselves.

For example, in (34), the annotator has erroneously assigned an incorrect semantic type (`space`) to a mention *the dollar*. In (35), the annotator marked *That* as `discourse old`, but failed to provide a suitable (segment) antecedent. In (36), the annotator marked a non-referring markable as an antecedent, not distinguishing between co-reference and other anaphoric phenomena. Finding such errors manually can be very tedious, as it requires a careful supervision of each markable and all its attributes. The availability of multiple annotation levels, on the contrary, allows for immediate listing of such mistakes.

(34) …thus dumping $[\text{dollar}]_1^{\text{abstract}}$ demand…Japanese institutions are comfortable with $[\text{the dollar}]_1^{\text{space}}$ anywhere between current levels and 135 yen.

(35) …production could increase to 23 millions or 24 millions barrels a day …[That] would send prices plummeting…

(36) We weren't allowed to do $[\text{any due diligence}]_1^{\text{non-referential}}$ because of competitive reasons. If we had, $[\text{it}]_1$ might have scared us.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

The following example illustrates a rather common problem with annotation projects that undergo several rounds of manual correction and adjudication. While each revision may fix some errors locally, the state-of-the-art annotation tools do not provide functionalities for ensuring the global data consistency.

(37) [Mr Dinkins]$_1$' position papers have more consistently reflected anti-development sentiment. [He]$_2$ favors a form of commercial rent control.

Here, the (rather large) coreference chain for *Mr Dinkins* underwent several revisions, with individual mentions being deleted and re-annotated. As a result, some other annotations, for example, the one for *He*, became corrupt. Note that the mention *He* was not re-annotated per se, it merely contained a link to a mention that underwent deletion and re-annotation.

On the other hand, our quality control procedures have revealed, through identifying conflicting attributes within coreference chains, cases of coreference that are problematic for annotators and therefore lead to inconsistent labeling. We have identified two types of difficulties. First, some examples require a practical approach that could have been discussed in the guidelines. Consider the following snippets:

(38) [Mr. Wathen]$_1$ says. "Their approach didn't work, [mine]$_1^{\text{abstract}}$ is."

(39) Currency analysts around the world have toned down their assessment of [the dollar]$_1^{\text{concrete}}$'s near-term performance…He said he expects U.S. interest rates to decline, dragging [the dollar]$_1^{\text{abstract}}$ to [around 1.80 marks]$_2^{\text{abstract}}$…I can't really see it dropping far below [1.80 marks]$_2^{\text{num}}$.

In (38), the annotators had difficulties labeling the mention *mine*, since the guidelines have no specific instructions on how to label this type of possessives. This resulted in an inconsistent labeling of *mine* as an abstract object (thus, referring to Mr. Wathen's approach) coreferent with the person entity (Mr. Wathen). For (39), the guidelines provide no explicit instructions for assigning semantic class to currencies, resulting in a very inconsistent labeling, with four different values within the same document. Clearly, no annotation guidelines are perfectly complete, so, we believe that semi-automatic consistency checks can help identify and clarify such issues and consequently lead to better schemes with higher inter-annotator agreement.

Second, some semantic and discourse level phenomena are intrinsically difficult to annotate. In particular, we have seen a lot of inconsistent semantic class labelings. These cover cases where annotators cannot decide reliably on a unique semantic class for the whole chain, for example, cases of regular metonymy:

(40) [Kellog]$_1^{\text{organization}}$'s spokesman said …"As [we]$_1^{\text{person}}$ regain our leadership…."

Analyzing the data consistency logs, we have identified a number of truly challenging cases of coreference, both in terms of annotation and automatic resolution. These cases often fall in the category of *near-identity coreference* Recasens *et al.* (2011). For example, in (41) *survey*, *data* and *figures* are very closely related mentions. It can be argued that they are all referring to the same entity—at the same time, it can also be argued, that *figures*, representing *data*, are part of *survey*, making the case for bridging relations. Example (42) shows another tricky case, posing a challenge, especially for automatic coreference resolution algorithms. Here, the same entity is described from two very different angles, using two mentions that are semantically rather dissimilar.

(41) [The Confederation of British Industry's latest survey]$_1$ shows…But despite mounting recession fears, [government data]$_1$ don't yet show the economy grinding to a halt…[The latest government figures]$_1$ said retail prices in September were up 7.6% from a year earlier.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 10. Comparison across anaphoricaly annotated corpora

	ACE-05	ARRAU	OntoNotes
Corpus size (# tokens)	220K	350K	1.5M
Different genres	−	+	+
Min and max mention boundaries	+	+	−
Discontinuous mentions	−	+	−
Mention type annotated	+	−	−
Mention attributes annotated	±	+	−
Singletons annotated	+	+	−
All (co)referential mentions annotated	−	+	+
Non-referentials	−	+	−
Explicit annotation for generics	−	+	−
Discourse deixis/events	−	+	+
Anaphoric ambiguity	−	+	−
Rich gold linguistic annotations of text	−	±	+

Information marked + is annotated, − is not annotated, and ± is partially annotated.

(42) Nearby Pasadena, Texas, police reported that [104 people]$_1$ had been taken to area hospitals, but a spokeswoman said [that toll]$_1$ could rise.

The near-identity coreference presents a true challenge for the community, yet, it is essential for the correct interpretation of textual inputs, especially in more complex domains (e.g., fiction) with evolving entities. For example, a Machine Reading system equipped with a strong coreference resolver, can suggest an informative answer (*The Confederation of British Industry's latest survey*) to such queries as *Which source is optimistic about the current economic situation?* or *Where can I find the data on the recent retail price trends?*—whereas without coreference, the answer would be rather superficial and not helpful for the user (*government data* or *the latest government figures*.

The detailed analysis of such examples constitutes a part of our ongoing work. Note that producing a non-negligible amount of challenging examples has only been made possible as a by-product of our thorough linguistically motivated annotation, for example, through a conflict between coreference and non-referentiality annotations.

4. Related work: ARRAU vs. other anaphoric corpora

A number of anaphorically annotated corpora have appeared in the past two decades—an extensive overview can be found in Poesio *et al.* (2016)—but very few of these cover the range of genres and the types of anaphoric relations annotated in ARRAU, such as bridging reference and discourse deixis, and much of this work started after the ARRAU annotation began. In this section, we discuss first of all the main differences between ARRAU and the two most commonly used corpora annotated for coreference in English, ACE (Doddington *et al.* 2004) and ONTONOTES (Pradhan *et al.* 2011; Weischedel *et al.* 2011; Pradhan *et al.* 2012). We then discuss related work on annotating genres other than news, semantic properties of mentions such as referentiality and genericity, bridging references, and discourse deixis.

4.1 ARRAU vs. ACE vs. OntoNotes

Table 10 provides a summary of the most distinctive features of ARRAU as opposed to ACE and ONTONOTES.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

The most prominent feature of ARRAU is its rich linguistically motivated annotation of mentions and relations between them. Thus, unlike ACE and ONTONOTES, ARRAU combines identity coreference with a number of related phenomena, such as referentiality, genericity, discourse deixis, and bridging. Moreover, we allow for ambiguity between different relations. The other datasets focus mainly on identity anaphora, with references to events being annotated in ONTONOTES. We believe that it is very important to have the same corpus annotated for different anaphora-related phenomena to allow for deeper linguistic analysis and joint modeling. In this respect, ARRAU follows the line adopted by the Prague Dependency Tree Bank (Nedoluzhko *et al.* 2009b), where several anaphoric relations are encoded for the same textual material, going beyond identity anaphora.[n]

In ARRAU each markable is shown with its minimal and maximal span. This solution is in line with the ACE annotation guidelines and it is unfortunate it was not been adopted in ONTONOTES in order to decrease the annotation price and thus augment the corpus size. The maximal span corresponds to the full noun phrase, whereas the minimal span corresponds to the head noun or to the bare named entity for complex NE-nominals. With the latest development in parsing technology, it might seem redundant to include minimal spans in the manual annotation directly: using dependencies or constituents with head-finding rules, one might expect to extract the minimal span for each NP rather reliably. It has been shown, however, that naive parsing-based heuristics do not lead to the best performance and a coreference resolver might benefit considerably from explicit or latent identification of minimal spans or heads (Zhekova and Kübler 2013; Peng *et al.* 2015). Moreover, explicitly annotated minimal spans allow for better lenient matching that has been shown to improve the training procedure of coreference resolvers through better alignment of automatically extracted and gold mentions (Kummerfeld *et al.* 2011). Finally, minimal spans can be intrinsically difficult to extract for non-conventional documents, such as dialog transcripts or social media, due to the low quality of parsing technology for such data [cf. an overview of parsing technology across domains/genres (Versley 2005), as well as a recent study discussing numerous problems related to syntactic parsing specific for conversational data (Nasr *et al.* 2016)]. We believe therefore that the combination of minimal and maximal spans is the most reliable way of annotating mention boundaries for coreference. The second release of ARRAU provides minimal and maximal spans for all the domains.

Consistently with linguistic views on nominal expressions, ARRAU supports discontinuous mentions. The ACE mark-up could potentially allow for discontinuous mentions, but the guidelines explicitly instruct the annotators to always select contiguous chunks. The CONLL mark-up is not expressive enough to support discontinuous mentions.

In ARRAU, all types of markables are annotated. In particular, we label singletons (mentions that do not participate in coreference chains) and non-referring NPs. The ACE guidelines restrict the annotation scope to referentials,[o] whereas in ONTONOTES only co-referential mentions are marked, not singletons. Our corpus statistics show that non-referring markables and singleton mentions account for up to one third of all the markables. Again, restricting the annotation scope allows for reducing the manual effort per document and thus for increasing the corpus size. However, a dataset with all the nominal expressions annotated provides material for training mention detection systems. Mention detection for ONTONOTES (Kummerfeld *et al.* 2011; Uryupina and Moschitti 2013) is a non-trivial problem that is further aggravated by the fact that singletons are removed and thus direct training becomes hardly possible.

Each markable is annotated in ARRAU with its basic morphological properties: number, gender and semantic class. This allows, again, for training markable-level classifiers to assign these

[n]The only comparable large-scale approach for English we are aware of is GECCO (Lapshinova-Koltunski and Kunz 2014): a corpus extensively annotated for different means of discourse cohesion in a semi-automatic way. Unfortunately, this dataset is not publicly available and we are not aware of any plans for releasing it to the scientific community.

[o]Moreover, the ACE guidelines focus on specific semantic types of referential mentions, motivated from an Information Extraction perspective: person, organization, location, and so on.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

features automatically. Similarly to minimal span, this task can be attempted via heuristics based on parse trees, however, one can expect a higher performance if such tasks are attempted in a data-driven way.

The text collections used in ARRAU have been annotated for a variety of relevant discourse-level properties by other projects. For example, our news documents are taken from the RST treebank and thus further annotations can be induced from RST to investigate possible interactions between coreference and rhetorical structure.[p] The ONTONOTES dataset, on the contrary, provides valuable gold annotations of low-level phenomena (e.g., gold part-of-speech tags or parse trees), but does not, to our knowledge, provide deep discourse-level annotations apart from coreference.[q] We believe that a careful analysis of the overlapping documents, annotated within both the ARRAU and ONTONOTES schemes, will provide valuable insights for computational modeling of coreference/anaphora.

To summarize, the ARRAU dataset provides a high-quality refined annotation of anaphora and related phenomena. It relies on much more detailed and specific annotation guidelines than other commonly used corpora. We believe therefore that while the ONTONOTES corpus, being much larger, is of crucial importance for data-intensive modeling of linguistically easier cases of coreference, ARRAU can be valuable, on the one hand, for deeper linguistically oriented analysis of complex cases and, on the other hand, for learning models for related phenomena (genericity, referentiality, etc.).

4.2 Genres: beyond news

When the ARRAU annotation started in 2004, the main available corpora for studying coreference/anaphora resolution in English, such as MUC and ACE, focused on news content; there were a few resources covering other types of text, such as the GNOME corpus already mentioned; but the only corpora covering English spoken dialog were Byron and Allen's annotation of pronouns in the TRAINS corpus (Byron and Allen 1998) and the annotation of part of the SHERLOCK corpus of task-oriented instructional dialog included in GNOME and used for the study of anaphora and discourse structure reported in Poesio *et al.* (2006a).[r]

In the years since the situation has improved. Corpora covering textual genres other than news now exist, such NLP4EVENTS of instructional manuals (Hasler *et al.* 2006), the GENIA corpus of biomedical text (Nguyen *et al.* 2008) or the TED talks dataset annotated for coreference within the ParCor project (Guillou *et al.* 2014). For dialog, Müller annotated pronominal reference in the ICSI spoken multi-partner conversation corpus (Müller 2008). Most importantly, the latest releases of ONTONOTES and of the Prague Dependency Treebank include substantial amounts of spoken text—for instance, release 5 of ONTONOTES contains, apart from newswire and broadcast news, a substantial amount of broadcast conversation and telephone conversation, as well as web data. And the recently created GECCO corpus (Lapshinova-Koltunski and Kunz 2014) covers a variety of genres including spoken language used both formally and informally (but as far as we know this corpus is not yet available).

4.3 Mention attributes

Referentiality. As mentioned above, the annotation schemes for coreference used in the best-known resources for English (the MUC and ACE corpora, ONTONOTES) do not require the

[p] We do not provide RST annotations with the ARRAU distributions. The relevant information can be extracted through straightforward corpora alignment.

[q] However, a portion of ONTONOTES builds upon the material from Penn TreeBank and thus can be aligned with RST as well.

[r] Nissim *et al.*'s annotation of the Switchboard corpus (Nissim *et al.* 2004) only provided the information status of mentions, not their antecedents if any. Navarretta had annotated pronominal anaphora in Danish and Italian dialog (Navarretta 2000) and Poesio *et al.* had annotated Italian MapTask dialogs (Poesio *et al.* 2004*c*).

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

annotation of non-referring expressions, or even singletons. But several of what we have called the "new wave" of linguistically motivated corpora, particularly those with a syntactic definition of mention, do, although typically (non-)referentiality is indicated in a more indirect way than in ARRAU. In TÜBA/DZ (Hinrichs *et al.* 2005), for instance, an `expletive` attribute is used to mark pleonastic instances of the impersonal third person singular pronoun *es* (*it*). No other types of non-referentiality seem to be marked. In ANCORA (Recasens and Martí 2010), mentions are automatically extracted from the syntactic tree, and an `entityref` attribute is used to mark referring NPs, so non-referring mentions can be identified although again the representation is not quite so explicit as in ARRAU.[s]

Genericity. When the ARRAU annotation started, we were only aware of one attempt at marking the genericity status of mentions apart from our own efforts in GNOME—the annotation of the `entity-class` attribute in ACE-2, with values `generic` and `specific`— but there has been some more work in this area since.

The ACE-2 Entity Detection and Tracking Guidelines do provide instructions for distinguishing generic from specific mentions, relying heavily on examples to address the problems we encountered in GNOME. We do not know however whether these difficulties were in fact solved as we are not aware of any results regarding the reliability of these guidelines. This annotation was nevertheless used to train one of the first models of automatic genericity classification (Reiter and Frank 2010). Herbelot and Copestake (2008) developed an interesting scheme, strongly rooted in the literature on genericity in formal linguistics (Carlson and Pelletier 1995), and still attempting to capture genericity and specificity but treating them as two separate two dimensions of classification, as done in Carlson and Pelletier (1995). The first version of the scheme provides a label for generic entities (`gen`), which would be the equivalent of the ARRAU label `generic-yes`; one for non-generic, specific entities (`spec`); and one for non-generic, non-specific entities (`non-spec`). In addition, the label `amb` is used for references ambiguous between generic and non-generic reading (like `undersp-gen` in the ARRAU scheme), and a label `group` for references to subgroups of a generic entity. This version of the scheme achieves a similar reliability to that of the second version of the GNOME scheme for genericity. The authors then developed a second set of guidelines, based on the same scheme but providing detailed instructions for a number of special cases; with this second set of guidelines they manage to achieve a reliability of $\kappa = 0.74$. In the Prague Dependency Treebank, all nominals are marked as generic or specific, and coreference relations are only marked between nominals with the same category (generic or specific). Recently, a systematic analysis of coreference with generic NPs was carried out by Nedoluzhko (2013). Finally regarding manual annotation, we shall mention the recent and very interesting work by Friedrich *et al.* (2015) who annotated *clauses* and *subjects* for genericity, a type of annotation that would be a very useful preliminary step towards the annotation of genericity of mentions in ARRAU. Friedrich and colleagues reported an average agreement of around $\kappa = .56$.

4.4 Other corpora annotated for more complex forms of anaphoric reference

Corpora annotated for bridging references. At the time the ARRAU annotation started, there had not been many other attempts to annotate bridging reference apart from our own work as part of the Viera / Poesio corpus (Poesio and Vieira 1998) and GNOME corpus (Poesio 2004a),[t] but there have been a number of efforts, since many of which have attempted to annotate a broader range of the bridging relations identified in the early literature (Passonneau 1997; Davies *et al.* 1998; Vieira 1998).

[s]Expletives are rare in Catalan and Spanish, so the `entityref` attribute is primarily missing from predicative NPs.

[t]Nissim *et al.*'s annotation for information status of parts of the Switchboard corpus (Nissim *et al.* 2004) did include identification of associative bridging references ("mediated" references) according to a scheme that covered, apart from the part and set relations covered in GNOME, *situation*, and *event* relations, but the actual anchor was not marked.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

One of the most ambitious such efforts in terms of coverage of relations is the work by Gardent and Manuélian as part of the annotation of the DeDe corpus (Gardent and Manuélian 2005). Gardent and Manuélian annotate a range of bridging relations including, apart from the part relations encoded in ARRAU, a more general *circumstantial* relation covering a variety of relations. The annotation was also carried out using MMAX2 and the markup scheme is very compatible with that used in ARRAU. No agreement results were however reported as far as we're aware.

Possibly the most extensive effort towards annotating bridging carried out in parallel with the annotation of ARRAU is the annotation of bridging coreference in the Prague Dependency Treebank (Nedoluzhko *et al.* 2009a). Nedoluzhko *et al.* distinguish, apart from `part` and `subset` relations,

- A `funct` relation covering function-value relations, as proposed in MATE (Davies *et al.* 1998)
- A new `contrast` relation covering relations between opposites (*People don't chew, it's cows who chew*)
- A more general underspecified group `rest`, which is used for capturing other types of bridging references such as event argument.

Nedoluzhko *et al.* measured interannotator agreement using a combination of F1 values (for the antecedent) and κ (for the relation), achieving F1=.59 for the antecedent, and $\kappa = .88$ for the relation.

Another substantial annotation effort was carried out by Hou, Markert and Strube (Markert *et al.* 2012), who annotated HSNOTES, a corpus of 11,000 NPs in 50 texts taken from the WSJ portion of ONTONOTES for information status, building on the scheme by Nissim *et al.* (2004) but also annotating the anchors of 663 bridging NPs. Their scheme also expands on the definition of `mediated` from Nissim *et al.* by also including *other* anaphora among the bridging references, as well as the `funct` relation. An interesting analysis of the differences between the notion of bridging reference annotated in ARRAU and that annotated in HSNOTE can be found in Roesiger (2018).

Finally, we should mention that there has been quite a lot of research on bridging reference in corpus linguistics, which, while not producing usable corpora, did involve developing annotation guidelines—a notable example being the work by Botley (2006).

Discourse Deixis. Annotating discourse deixis is another task not tackled in all large-style anaphoric annotations, but there have been a number of efforts both preceding, parallel with, and subsequent to the effort in ARRAU and the already discussed studies of agreement in discourse deixis annotation (Artstein and Poesio 2006).

Prior to ARRAU, we will mention the seminal work by Byron, who annotated pronominals and demonstratives in the TRAINS-93 corpus, including abstract objects (Byron and Allen 1998), and used the data to develop the first resolver of references to abstract objects we are aware of Byron (2002); Navarretta, who in parallel with Byron carried out similar studies of abstract reference in Danish and Italian (Navarretta 2000, 2008); and by Eckert and Strube (2000), who analyzed references in dialogs to both concrete and abstract objects. We will also mention the illuminating work by the corpus linguists Gundel, Hedberg, Hegarty, and associates on reference to "clausally introduced entities" (Gundel *et al.* 2003, 2002), that was an important influence on our own work.

The most notable effort carried out in parallel with the work on ARRAU is the work in ONTONOTES on annotating event anaphora, an important category of reference to abstract objects.

Following the first work in ARRAU on annotating discourse deixis, there appeared a number of studies that attempted to annotate a comparable subset of the phenomenon in other languages. Most notably among these efforts are the work in ANCORA on discourse deixis in Catalan and Spanish (Recasens 2008), and the work by Dipper and Zinsmeister (2012) on abstract anaphora in

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

German. More recently, a very systematic analysis and annotation of another subset of discourse deixis, so called *shell nouns*, has been carried out by Kolhatkar (2014) and Kolhatkar and Hirst (2012).

Ambiguity. Anaphoric ambiguity is a very understudied phenomenon and there have been hardly any other attempts to create a corpus in which the ambiguity of expressions is marked, with two exceptions. The coreference annotation carried out by Krasavina and Chiarcos (2007) as part of the work on the Potsdam Commentary Corpus (Stede 2004) is the only other coreference annotation scheme we are aware of that asks coders to mark ambiguity. The guidelines produced by Chiarcos and Krasavina (2005) require coders to use the `ambiguity` mention attribute to indicate ambiguous mentions, and the type of ambiguity: for instance, `ambig-ante` if the mention is clearly discourse old but it's not clear what the antecedent is, or `ambig-expl` for instances of *es* ("it") that could be interpreted either as anaphoric or as expletives.

We must admit however that the results of Poesio and Artstein (2005a) convinced us that this aspect of the ARRAU guidelines clearly needs rethinking, and in our research since we have taken a completely different direction, aiming to capture *implicit* ambiguity rather than explicit, as in ARRAU. Indeed, this aim was one of the primary motivations behind the *Phrase Detectives* project Poesio *et al.* (2013), which has developed a Game-With-A-Purpose to elicit from players multiple interpretations for anaphoric expressions (25 on average and as many as 32 in some cases). The *Phrase Detectives* game uses a simplified version of the ARRAU coding scheme, but all interpretations are stored, and preliminary analysis suggests that around 40% of mentions have at least two interpretations selected by at least two players. A first, small subset of the Phrase Detectives corpus was recently released via LDC (Chamberlain *et al.* 2016). More recently, this line of research has led us to start the DALI project,[u] in which besides carrying out the development of more Games-With-A-Purpose to study anaphora, methods to compare and analyze these interpretations will also be developed.

5. Anaphora resolution with ARRAU

In this Section we briefly discuss anaphora resolution work that used ARRAU.

5.1 Identity anaphora

Rodriguez (2010) used a preliminary release of ARRAU 2—about half the size of the final release, but already annotated with `MIN` information–to carry out a comparative analysis of anaphora resolution in English and Italian. Using BART (Versley *et al.* 2008), he compared the difficulty of anaphora resolution in ARRAU and in the two more widely used corpora at the time, `MUC-7` and `ACE02`. He also studied 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 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 11 summarizes the results.

5.2 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

[u] http://www.dali-ambiguity.org.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 11. (Uryupina and Poesio 2012): Running BART on different ARRAU genres and on different ONTONOTES genres. MUC score

	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

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.

5.3 The CRAC 2018 shared task

The first evaluation campaign based on ARRAU (Poesio *et al.* 2018) was organized in connection with the 2018 NAACL Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC).[v] The shared task was composed of three subtasks: Task 1 on identity anaphora resolution, Task 2 on bridging reference, and Task 3 on discourse deixis.

Datasets. Three separate datasets were made available for the three distinct tasks. All three datasets were in the format used for the EVALITA-2011 evaluation campaign (Uryupina and Poesio 2013), which in turn was derived from the tabular CONLL-style format used in the SEMEVAL 2010 shared task on multilingual anaphora (Recasens *et al.* 2010). (See Poesio *et al.* 2018 or the shared task page for further details on the format.) Three of the ARRAU sub-corpora were used: PEAR, RST and TRAINS.

Evaluation Scripts. Three evaluation scripts were developed for the three tasks.

The coreference evaluation script developed by Moosavi and Strube, which, in turn, builds upon the official CONLL implementation (Pradhan *et al.* 2014), was modified to produce the scorer for Task 1. We will refer to this script as 'the extended coreference scorer'.[w] 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).[x] When required to use MIN information, the extended scorer follows the MUC convention, and considers a mention

[v] http://crac18.dali-ambiguity.org.

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

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

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 12. Markable detection in ARRAU and ONTONOTES

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

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.

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.

Finally, the evaluation script for Task 3 computes the *Success@N* metric proposed by 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 system's first n choices. (S@1 is standard precision.)

Task 1: Markable detection. 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 detection 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 detection (see Poesio *et al.* 2018 for details). Two versions of this markable detection were run on the ONTONOTES dataset, one optimized for F1, one for recall. The results are shown in Table 12.

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

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 13. Baseline results on Task 1. Gold markables

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

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 straightforward notion of markable used in ARRAU.

Task 1. 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.

The first 10 lines of Table 13 show the results by the Stanford Deterministic Coreference Solver when run over gold markables, scored using both 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 13 show the results when including singletons in the assessment; 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.)

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 14. Baseline results on Task 1 with predicted mentions, without MIN information

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 15. Baseline results on Task 1 with predicted mentions, using 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

The first conclusion that can be obtained from this table is that the results achieved by the Stanford resolver on gold markables on this dataset are broadly comparable to the results the system achieved on gold markables (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.

Table 14 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.

Finally, Table 15 shows the effect of using the MIN information. As can be seen from the table, this results in five extra percentage points.

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.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Table 16. Roesiger's results on Task 2 for all domains

	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

Roesiger developed two systems, one rule-based, one ML-based. The results obtained by these systems on all three subdomains are summarized in Table 16. 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. Conclusion

This paper presents ARRAU—a publicly available corpus of anaphora, annotated according to linguistically motivated guidelines. The dataset contains documents from four different genres for a total of 350K tokens.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

ARRAU supports rich annotation of individual mentions: apart from morphosyntactic properties, we mark semantic type, genericity and referentiality. For the latter two properties, we also provide fine-grained subclassification. Apart from identity coreference, ARRAU guidelines cover bridging references and discourse deixis, thus providing data for joint modeling of these phenomena. We believe that the resulting resource provides valuable data for the next generation of anaphora resolvers. A few interesting studies in this direction were carried out as a result of the CRAC 2018 Shared Task (Poesio *et al.* 2018).

The annotation scheme developed for ARRAU, as well, could be useful for future research. It has already been employed in other projects, for example, the creation of the LIVEMEMORIES corpus of anaphora in Italian (Rodriguez *et al.* 2010). containing texts from Wikipedia and blogs. The main distinguishing feature of the LIVEMEMORIES coding scheme with respect to that of ARRAU is the incorporation of the MATE / VENEX proposals concerning incorporated clitics and zeros in standoff schemes whose base layer is words (instead of an annotation of morphologically decomposed argument structure, as in the Prague Dependency Treebank). A second project using the ARRAU guidelines is the creation of the SENSEI corpus,[y] consisting of annotations of online forums in English (from *The Guardian* newspaper) and Italian (from *La Repubblica* newspaper) following similar guidelines.

Acknowledgements. The ARRAU corpus has been under development over several years and we are grateful to the many funding agencies that contributed to its development. Initial work was in part supported by the EPSRC-funded ARRAU Project (GR/S76434/01). Subsequent work was funded in part by the LiveMemories project, funded by the Provincia of Trento; in part by the EU Project H2020 5G-CogNet; and in part by the ERC Project DALI, ERC-Adg-2015.

References

Amoia M., Kunz K. and Lapshinova-Koltunski E. (2011). Discontinuous constituents: A problematic case for parallel corpora annotation and querying. In *Proceedings of RANLP2011 Workshop on Annotation and Exploitation of Parallel Corpora*, pp. 2–10.

Artstein R. and Poesio M. (2006). Identifying reference to abstract objects in dialogue. In Schlangen D. and Fernandez R. (eds.), *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue*.

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

Asher N. (1993). *Reference to Abstract Objects in English*. Dordrecht: D. Reidel.

Björkelund A. and Farkas R. (2012). Data-driven multilingual coreference Resolution using resolver stacking. In *Proceedings of the Conference on Computational Natural Language Learning: Shared Task*, 49–55.

Björkelund A. and Kuhn J. (2014). Learning structured perceptions for coreference resolution with latent antecedents and non-local features. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 47–57.

Botley S.P. (2006). Indirect anaphora: Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics* **11**(1), 73–112.

Boyd A., Dickinson M. and Meurers D. (2008). On detecting errors in dependency treebanks. *Research on Language and Computation* **6**(2), 113–137.

Burga A., Cajal S., Codina-Filba J. and Wanner L. (2016). Towards multiple antecedent coreference resolution in specialized discourse. In *Proceedings of the Language Resources and Evaluation Conference*.

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

Byron D. (2002). Resolving pronominal references to abstract entities. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 80–87.

Carletta J. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics* **22**(2), 249–254.

Carlson G.N. and Pelletier F.J. (eds.) (1995). *The Generic Book*. Chicago, IL: University of Chicago Press.

Carlson L., Marcu D. and Okurowski M.E. (2002). *RST Discourse Treebank LDC2002T07*.

Carlson L., Marcu D. and Okurowski M.E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Kuppevelt J. and Smith R. (eds.), *Current Directions in Discourse and Dialogue*. Dordrecht: Kluwer, pp. 85–112.

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

Chamberlain J., Poesio M. and Kruschwitz U. (2016). Phrase detectives corpus 1.0: Crowdsourced anaphoric coreference. In *Proceedings of the Language Resources and Evaluation Conference*.

[y] http://www.sensei-conversation.eu.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

Chiarcos C. and Krasavina O. (2005). *Annotation Guidelines PoCoS—Potsdam Coreference Scheme: Core Scheme, Draft 0.912*. Potsdam and Berlin: University of Potsdam and Humboldt University.

Clark H.H. (1975). Bridging. In *Proceedings of the Conference on Theoretical Issues in Natural Language Processing*.

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

Davies S., Poesio M., Bruneseaux F. and Romary L. (1998). *Annotating Coreference in Dialogues: A Proposal for a Scheme for MATE*. Available at http://www.cogsci.ed.ac.uk/ poesio/MATE/anno_manual.html.

Dickinson M. and Lee C.M. (2008). Detecting errors in semantic annotation. In *Proceedings of the Language Resources and Evaluation Conference*.

Dickinson M. and Meurers D. (2003). Detecting errors in part-of-speech annotation. In *Proceedings of the 10th Conference of European chapter of the Association for Computational Linguistics*, pp. 107–114.

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

Doddington G., Mitchell A., Przybocki M., Ramshaw L., Strassell S. and Weischedel R. (2004). The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proceedings of the Language Resources and Evaluation Conference*.

Durrett G. and Klein D. (2013). Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1971–1982.

Eckert M. and Strube M. (2000). Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics* **17**(1), 51–89.

Fernandes E.R., dos Santos C.N. and Milidiú R.L. (2014). Latent trees for coreference resolution. *Computational Linguistics* **40**(4), 801–835.

Frank A., Bögel T., Hellwig O. and Reiter N. (2012). Semantic annotation for the digital humanities using Markov logic networks for annotation consistency control. *Linguistic Issues in Language Technology* 1–21.

Friedrich A., Palmer A., Sorensen M.P. and Pinkal M. (2015). Annotating genericity: A survey, a scheme, and a corpus. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Gardent C. and Manuélian H. (2005). Creation d'un corpus annotée pour le traitment des descriptions définies. *Traitement Automatique des Langues* **46**(1), 115–139.

Ge N., Hale J. and Charniak E. (1998). A statistical approach to anaphora resolution. In *Proceedings of the 6th Workshop on Very Large Corpora*, pp. 161–170.

Grishina Y. and Stede M. (2015). Knowledge-lean projection of coreference chains across languages. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora*, pp. 14–22.

Guillou L., Hardmeier C., Smith A., Tiedemann J. and Webber B. (2014). ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 3191–3198.

Gundel J.K., Hedberg N. and Zacharski R. (2002). Pronouns without explicit antecedents: How do we know when a pronoun is referential?. In *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium*.

Gundel J.K. , Hegarty M. and Borthen K. (2003). Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information* **12**(3), 281–299.

Hasler L., Orasan C. and Naumann K. (2006). NPs for events: Experiments in coreference annotation. In *Proceedings of the Language Resources and Evaluation Conference*.

Hendrickx I., Bouma G., Coppens F., Daelemans W., Hoste V., Kloosterman G., Mineur A.-M. , Van Der Vloet J. and Verschelde J.-L. (2008). A coreference corpus and resolution system for Dutch. In *Proceedings of the Language Resources and Evaluation Conference*.

Herbelot A. and Copestake A. (2008). Annotating genericity: How do humans decide? (a case study in ontology extraction). In Featherston S. and Winkler S. (eds.), The Fruits of Empirical Linguistics. Berlin: de Gruyter.

Hinrichs E., Kübler S. and Naumann K. (2005). A unified representation for morphological, syntactic, semantic and referential annotations. In *ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*.

Hou Y., Markert K. and Strube M. (2013). Global inference for bridging anaphora resolution. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Hollenstein N., Schneider N. and Webber B. (2016). Inconsistency detection in semantic annotation. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 3986–3990.

Kolhatkar V. and Hirst G. (2012). Resolving "this-issue" anaphora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1255–1265.

Kolhatkar V. and Hirst G. (2014). Resolving shell nouns. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Kolhatkar V., Zinsmeister H. and Hirst G. (2013). Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Kolhatkar V. (2014). *Resolving Shell Nouns*. PhD Thesis, University of Toronto.

Krasavina O. and Chiarcos C. (2007). The potsdam coreference scheme. In *Proceedings of the 1st Linguistic Annotation Workshop*, 156–163.

Krippendorf K. (2004). *Content Analysis: An Introduction to Its Methodology*, 2nd Edn., chapter 11. Thousand Oaks, CA: Sage.

Kummerfeld J.K., Bansal M., Burkett D. and Klein D. (2011). Mention detection: Heuristics for the OntoNotes annotations. In *Proceedings of the Conference on Computational Natural Language Learning: Shared Task*, 102–106.

Kunz K. and Lapshinova-Koltunski E. (2015). Cross-linguistic analysis of discourse variation across genres. *Nordic Journal of English Studies* **14**(1), 258–288.

Kunz K., Lapshinova-Koltunski E. and Martínez J.M. (2016). Beyond identity coreference: Contrasting indicators of textual coherence in English and German. In *Proceedings of the Workshop on Coreference Resolution beyond OntoNotes*, pp. 23–31.

Lapshinova-Koltunski E. and Kunz K.A. (2014). Annotating cohesion for multilingual analysis. In *Proceedings of the LREC ISO Workshop on Interoperable semantic resources*, pp. 57–64.

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

Lee K., He L., Lewis M. and Zettlemoyer L. (2017). End-to-end neural coreference resolution. In *Proceedings of EMNLP*.

Loukachevitch N.V., Dobrov G.B. , Kibrik A.A. , Khudyajova M.V. and Linnik A.S. (2011). Factors in referential choice. In *Proceedings of Dialogue*, Moscow.

Marasović A., Born L., Opitz J. and Frank A. (2017). A mention-ranking model for abstract anaphora resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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

Martschat S. and Strube M. (2015). Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics* **3**, 405–418.

Moosavi N.S. and Strube M. (2016). A proposal for a link-based entity aware metric. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Markert K., Hou Y. and Strube M. (2012). Collective classification for fine-grained information status. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Modieska N.N. (2003). *Resolving Other Anaphors*. PhD Thesis, University of Edinburgh.

Müller C. and Strube M. (2006). Multi-level annotation of linguistic data with MMAX2. In Braun S., Kohn K. and Mukherjee J. (eds.), *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, vol. **3** of English Corpus Linguistics. New York: Peter Lang, pp. 197–214.

Müller M.-C. (2008). *Fully Automatic Resolution of it, This and That in Unrestricted Multy-Party Dialog*. PhD Thesis, Universität Tübingen.

Nasr A., Damnati G., Guerraz A. and Bechet F. (2016). Syntactic parsing of chat language in contact center conversation corpus. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, pp. 175–184.

Navarretta C. (2000). Abstract anaphora resolution in Danish. In *Proceedings of SIGdial Workshop on Discourse and Dialogue*, pp. 56–65.

Navarretta C. (2008). Pronominal types and abstract reference in the Danish and Italian DAD Corpora. In Johansson C. (ed.), *Proceedings of the Second Workshop on Anaphora Resolution NEALT Proceedings Series*, Bergen. pp. 63–71.

Nedoluzhko A. (2013). Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank. In *Proceedings of the Linguistic Annotation Workshop*, pp. 103–111.

Nedoluzhko A., Mirokvský J. and Pajas P. (2009a). The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague Dependency Treebank. In *Proceedings of the Linguistic Annotation Workshop*, pp. 108–111.

Nedoluzhko A., Mírovskỳ J., Ocelák R. and Pergler J. (2009b). Extended coreferential relations and bridging anaphora in the Prague Dependency Treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium*, 1–16.

Neumann S. (2013). *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. Berlin: Mouton de Gruyter.

Nguyen N.L.T. , Kim J.-D. and Tsujii J. (2008). Challenges in pronoun resolution system for biomedical text. In *Proceedings of the Language Resources and Evaluation Conference*.

Nissim M., Dingare S., Carletta J. and Steedman M. (2004). An annotation scheme for information status in dialogue. In *Proceedings of the Language Resources and Evaluation Conference*.

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

Passonneau R.J. (1997). Instructions for applying discourse reference annotation for multiple applications (DRAMA). unpublished manuscript.

Peng H., Chang K.-W. and Roth D. (2015). A joint framework for coreference resolution and mention head detection. In *Proceedings of the Conference on Computational Natural Language Learning*.

Poesio M. (2000a). *The GNOME Annotation Scheme Manual*, 4th Edn. Scotland: University of Edinburgh, HCRC and Informatics. Available at http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm.

Poesio M. (2000b). Annotating a corpus to develop and evaluate discourse entity realization algorithms: Issues and preliminary results. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 211–218.

Poesio M. (2004a). The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*.

Poesio M. (2004b). Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pp. 72–79.

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

Poesio M., Mehta R., Maroudas A. and Hitzeman J. (2004b). Learning to solve bridging references. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 143–150.

Poesio M., Pradhan S., Recasens M., Rodriguez K. and Versley Y. (2016). Annotated corpora and annotation tools. In Poesio M., Stuckardt R. and Versley Y. (eds.), *Anaphora Resolution: Algorithms, Resources and Applications*, chapter 4. Berlin and Heidelberg: Springer.

Poesio M. and Artstein R. (2005a). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In Meyers A. (ed.), *Proceedings of ACL Workshop on Frontiers in Corpus Annotation*. pp. 76–83.

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

Poesio M., Delmonte R., Bristot A., Chiran L. and Tonelli S. (2004c). *The VENEX Corpus of Anaphoric Information in Spoken and Written Italian*. Available at http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf.

Poesio M., Patel A. and Di Eugenio B. (2006a). 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).

Poesio M., Chamberlain J., Kruschwitz U., Robaldo L. and Ducceschi L. (2013). Phrase detectives: Utilizing collective intelligence for Internet-scale language resource creation. *ACM Transactions on Intelligent Interactive Systems* **3**(1), 1–44.

Poesio M., Grishina Y., Kolhatkar V., Moosavi N., Roesiger I., Roussel A., Simonjetz F., Uma A., Uryupina O., Yu J. and Zinsmeister H. (2018). Anaphora resolution with the ARRAU corpus. In *Proceedings of the NAACL Workshop on Computational Models of Reference, Anaphora and Coreference*.

Poesio M., Bruneseaux F. and Romary L. (1999). The MATE meta-scheme for coreference in dialogues in multiple languages. In Walker M. (ed.), *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging*, 65–74.

Poesio M., Sturt P., Arstein R. and Filik R. (2006b). Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes* **42**(2), 157–175.

Poesio M. and Reyle U. (2001). Underspecification in anaphoric reference. In Bunt E.T.H. and van der Sluis I. (eds), *Proceedings of the Fourth International Workshop on Computational Semantics*, Tilburg: Tilburg University, pp. 286–300.

Poesio M. and Vieira R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics* **24**(2), 183–216. Also available as Research Paper CCS-RP-71, Centre for Cognitive Science, University of Edinburgh.

Poesio M. and Artstein R. (2008). Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Language Resources and Evaluation Conference*.

Poesio M. and Rieser H. (2010). Completions, coordination, and alignment in dialogue. *Dialogue and Discourse* **1**(1) 1–89.

Pradhan S., Ramshaw L., Weischedel R., MacBride J. and Micciulla L. (2007). Unrestricted coreference: Identifying entities and events in OntoNotes. In *in Proceedings of the IEEE International Conference on Semantic Computing*.

Pradhan S., Ramshaw L., Marcus M., Palmer M., Weischedel R. and Xue N. (2011). CoNLL-2011 Shared Task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Conference on Computational Natural Language Learning*.

Pradhan S., Moschitti A., Xue N., Uryupina O. and Zhang Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning*.

Pradhan S., Luo X., Recasens M., Hovy E., Ng V. and Strube M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the Annual Meeting of The Association for Computational Linguistics*.

Prince E.F. (1992). The ZPG letter: Subjects, definiteness, and information status. In Thompson S. and Mann W. (eds.), *Discourse Description: Diverse Analyses of a Fund-Raising Text*. Amsterdam, The Netherlands: John Benjamins, pp. 295–325.

Reiter N. and Frank A. (2010). Identifying generic noun phrases. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 40–49.

Roesiger I. (2018). Rule- and learning-based methods for bridging resolution in the ARRAU corpus. In *Proceedings of the NAACL Worskhop on Computational Models of Reference, Anaphora and Coreference*.

Recasens M. (2008). Discourse deixis and coreference: Evidence from AnCoRa. In Johansson C. (ed.), *Proceedings of the 2nd Workshop on Anaphora Resolution*.

Recasens M., àrquez L.M., Sapena E., Mart M.A.í, Taulé M., Hoste V., Poesio M. and Versley Y. (2010). SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the International Workshop on Semantic Evaluation*.

https://doi.org/10.1017/S1351324919000056 Published online by Cambridge University Press

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

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

Rodriguez K. (2010). *Resources for Linguistically Motivated Multilingual Anaphora Resolution*. PhD Thesis, Universitá di Trento.

Rodriguez K.-J., Delogu F., Versley Y., Stemle E. and Poesio M. (2010). Anaphoric annotation of Wikipedia and blogs in the live memories Corpus. In *Proceedings of the Language Resources and Evaluation Conference*.

Schuster E. (1988). *Pronominal Reference to Events and Actions: Evidence from Naturally-Occurring Data* (LINC LAB 100). Philadelphia, PA: Department of Computer and Information Science, University of Pennsylvania.

Stede M. (2004). The potsdam commentary corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*.

Uryupina O. and Poesio M. (2012). Domain-specific vs. uniform modeling for coreference resolution. In *Proceedings of the Language Resources and Evaluation Conference*.

Uryupina O. and Poesio M. (2013). Evalita 2011: Anaphora resolution task. In Magnini B., Cutugno F., Falcone M. and Pianta E. (eds.), *Evaluation of Natural Language and Speech Tools for Italian*, vol. 7689 of Lecture Notes in Computer Science. Springer, pp. 146–155.

Uryupina O., Artstein R., Bristot A., Cavicchio F., Rodriguez K.J. and Poesio M. (2016a). ARRAU: Linguistically-motivated annotation of anaphoric description. In *Proceedings of the Language Resources and Evaluation Conference*.

Uryupina O. and Moschitti A. (2013). Multilingual mention detection for coreference resolution. In *Proceedings of the International Joint Conference on Natural Language Processing*.

Uryupina O., Kabadjov M. and Poesio M. (2016b). Detecting non-reference and non-anaphoricity. In Poesio M., Stuckardt R. and Versley Y. (eds.), *Anaphora Resolution: Algorithms, Resources and Applications*, chapter 13. Berlin and Heidelberg: Springer.

van Deemter K. and Kibble R. (2000). On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics* **26**(4), 629–637.

Versley Y. (2005). Parser evaluation across text types. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*.

Versley Y. (2008). Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation* **6**, 333–353.

Versley Y., Ponzetto S., Poesio M., Eidelman V., Jern A., Smith J., Yang X. and Moschitti A. (2008). BART: A modular toolkit for coreference resolution. In *Proceedings of the Annual Meeting of The Association for Computational linguistics, demo session*.

Vieira R. (1998). *Definite Description Resolution in Unrestricted Texts*. PhD Thesis, University of Edinburgh, Centre for Cognitive Science.

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

Weischedel R., Hovy E., Marcus M., Palmer M., Belvin R., Pradhan S., Ramshaw L. and Xue N. (2011). OntoNotes: A large training corpus for enhanced processing. In Olive J., Christianson C. and McCary J. (eds.), *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Berlin and Heidelberg: Springer.

Wiseman S.J., Rush A.M. , Shieber S.M. and Weston J. (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the Annual Meeting of The Association for Computational linguistics*.

Zhekova D. and Kübler S. (2013). Machine learning for mention head detection in multilingual coreference resolution. In *Proceedings of the Recent Advance in Natural Language Processing Conference*, 747–754.

Cite this article: Uryupina O, Artstein R, Bristot A, Cavicchio F, Delogu F, Rodriguez KJ and Poesio M (2020). Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU Corpus. *Natural Language Engineering* **26**, 95–128. https://doi.org/10.1017/S1351324919000056