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] 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,*
[6]*School of Electronic Engineering and Computer Science, Queen Mary University of London*
uryupina@gmail.com, artstein@ict.usc.edu, lucanto137@libero.it,
federica.cavicchio@gmail.com, delogu@coli.uni-saarland.de
kepa.rodriguez@yadvashem.org.il

This paper presents the second release of ARRAU, a multi-genre corpus of anaphoric information created over ten year years to provide much needed 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: treating all NPs as mentions, including non-referring NPs, and annotating them according to their (non) referentiality; distinguishing between several categories of non-referentiality and annotating non-anaphoric mentions; thorough annotation of mention boundaries (minimal/maximal spans, discontinuous mentions); 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 workhop papers; and we discuss the development between the first release of ARRAU in 2008 and this second one.

Preprint submitted to Journal of Natural Language Engineering *August 3, 2017*

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] 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,*
[6]*School of Electronic Engineering and Computer Science, Queen Mary University of London*
uryupina@gmail.com, artstein@ict.usc.edu, lucanto137@libero.it,
federica.cavicchio@gmail.com, delogu@coli.uni-saarland.de
kepa.rodriguez@yadvashem.org.il

Key words: coreference, anaphora, discourse, annotation, linguistic corpora

1. Introduction

A great number of data-driven approaches to anaphora resolution have recently been proposed, considerably pushing forward the state of the art in the field (see, e.g., [1, 2, 3, 4, 5]; see also [6] 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 [7], and the success of recent evaluation campaigns using these new resources [8, 9, 6]. Most of the recently proposed approaches, however, still focus on the accurate modeling of relatively easy cases of anaphoric reference. For example, [1] 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 therefore 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—e.g., ANCORA for Catalan and Spanish [10], the Prague Dependency Treebank for Czech [11], or TÜBA-D/Z for German [12]—but not yet for English.

This paper presents the second release of the ARRAU corpus,[1] a multi-genre 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 ten 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, e.g., non-referring expressions, bridging references, and reference to abstract objects. Moreover, additional discourse-level information is available from third parties for subsets of ARRAU (e.g., the rhetorical structure annotations [13] 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 involve 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 dialogue (the TRAINS data) and fiction (the Pear Stories). Spontaneous dialogue and fiction are not covered by most commonly used coreference corpora.[2] Although several linguistic studies focus on genre-specific discourse coherence and anaphora properties [14, 15], only very few approaches aim at empirical analysis or per-genre modeling of coreference [16, 17]. In a recent work, Kunz et al [18] 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 [19, 7] only focus on *identity* anaphora—the task of identifying multiple mentions of the same discourse entity—and thus cannot support anaphoric ambiguity. By anno-

[1] `http://www.arrauproject.ororgg`

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

tating ambiguous anaphoric expressions, we make the first step toward a thorough investigation of anaphoric ambiguity.

Finally, during the ten 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 and 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: to our knowledge, the best practice in this direction doesn't go beyond ensuring high agreement between human coders, using, for example, Krippendorf's κ [20, 21]. 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 [22, 23], but this article 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 ARRAU corpus is publicly available from LDC; it will also be made available through the Anaphoric Bank.[3]

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

[3] The anaphorically annotated versions of LDC corpora such as the RST Discourse Treebank and the TRAINS-93 corpus require previous purchase of the original corpora.

ambiguous anaphoric references and reference to abstract objects. A key aspect of this work was to create large-scale annotated resources that could be used to study these types of anaphoric reference. Building on the GNOME guidelines ([24], discussed in [25, 26]) which already provided reliability-tested annotation schemes for aspects of anaphoric annotation such as bridging [27] and were used, e.g., to create the dataset in [28, 29], we developed and tested extended annotation guidelines [22] aiming specifically at abstract anaphora and ambiguity [30, 31]. These annotation guidelines, distributed with the corpus and available from the project website, also provide detailed instructions for identifying mention 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.[4]

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 [32]). 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 [33]) and for their argument structure (e.g., through the Propbank [34]) but also for rhetorical structure [32]. This dataset would therefore allow the study of the effect of these other types of linguistic information on anaphora resolution and vice versa.[5]

Apart from 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

[4] 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.

[5] This annotation took place in collaboration with, although independently from, the annotation of the same data carried out by prof. Kibrik's group at the Russian Academy of Sciences [35].

the task-oriented dialogues in the TRAINS-93 corpus;[6] 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 [36]; and the GNOME domain covers documents from the medical and art history genres covered by the GNOME corpus [37, 25] used to study both local and global salience [28, 38].

The same coding scheme was used for all domains, but separate guidelines were written for the textual domains and the spoken dialogue 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.[7] Both the RST and GNOME domains consist of carefully edited texts with complex grammatical sentences. This results in long mentions, often either multiword named entities (for example, full names of organizations) or complex NPs. Mention 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, represent spontaneous speech. The texts in these domains mostly consist of short utterances, often ungrammatical and/or with disfluencies. PEAR and TRAINS mentions therefore are on average much shorter, with a lot of one-word mentions, mostly pronouns. Discontinuous mentions (see Section 2.2 below) are present in both PEAR and TRAINS, although not very common. So for these domains, mention 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. Annotated Mentions

ARRAU is one of the 'new wave' of anaphorically annotated corpora that were created after the re-examination of annotation schemes for anaphora prompted by

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

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

	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
mentions	72013	6562	4008	16999
avg. mentions per doc	174.4	1312.4	200.4	149.1
avg. mention length (tok)	4.1	4.0	2.2	1.8
discontinuous mentions	864 (1.2%)	175 (2.7%)	3 (0%)	15 (0%)
one-word mentions	21461 (30%)	2338 (35.6%)	2164 (54.0%)	9404 (55.3%)
non-referential mentions	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%)

Table 1: Corpus statistics for the four ARRAU domains.

the Discourse Resource Initiative and the MATE and GNOME projects [39, 40, 41]. These new corpora—other examples include ANCORA [10], COREA [42], ONTONOTES [43], the anaphoric annotation of the Prague Dependency Treebank [11] and TÜBA-D/Z [12]—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 mentions. Some older corpora had imposed syntactic restrictions on mentions—for instance, in many older corpora only pronouns are annotated [44]. 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 [45]).

By contrast, according the ARRAU guidelines (which follow for text the earlier GNOME guidelines,[8] see below for the dialog guidelines) *all* NPs are considered as mentions, also when they are non-referring, like predicative *a busy place* in (1) (we discuss in Section 2.3 below 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 mentions 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, e.g., in the case of the proper name *US* in (2), and when the premodifier refers to a kind, like *exchange-rate* in (3).

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

(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.

(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 mentions, `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 mentions. One of the distinctive features of ARRAU is the support of **discontinuous mentions**—mentions built out of non-continuous material. Discontinuous chunks are problematic for many corpus annotation formats [46], and thus many guidelines developed for various linguistic phenomena allow for labeling continuous constituents exclusively.

Discontinuous mentions, however, are common in dialogue, for instance in cases of so-called **collaborative completions** [47] 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 mentions in the MMAX2 annotation tool [48], developed to support his research on anaphora resolution in dialogue [49], where spans can be arbitrary sequences of tokens. However, discontinuous mentions also provide a way to include in a mention all information provided by the text e.g., 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 mentions make it possible to include both the segments of text marked as part 1 and part 2 in the same mention. 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 mentions 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 [50] assume numbered brackets as mention boundaries that cannot encode discontinuous fragments. To make ARRAU usable for these purposes, whereas the mention can be discontinuous, *minimal spans* cannot be. This way, all the ARRAU mentions can be aligned to contiguous sequences of tokens.

2.3. Mention properties

All mentions are manually annotated for a variety of properties according to the GNOME guidelines [24]: 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`.[9] The guidelines and reliability studies leading to this scheme are discussed in [37, 26]. In this Section, we will only discuss in detail two additional attributes, specifying the genericity and the referential status of the NP. `Genericity` is annotated following a scheme developed in GNOME after experiments based on the official annotation manual had shown poor reliability for this attribute. Last but not least, a `reference` attribute was annotated, specifying a combination of information about the logical form status of the NP (referring, expletive, quantificational, or predicative). We discuss each attribute in turn.

Referring and non-referring mentions. Most anaphorically annotated corpora focus on *referring* mentions—mentions that denote discourse entities and can 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* mentions—expressions that do not denote 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 [51]. In order to develop

[9] 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.

such a classifier for a corpus like ONTONOTES in which non-referring expressions are not annotated, separate classifiers are required—for example, [52] trained a pre-filtering classifier for non-anaphoric *it*, *you* and *we* on the OntoNotes data.

In ARRAU, all mentions are annotated, including non-referring 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 (13) and coordinations (14) [24, 26]. In ARRAU, coders are asked, first of all to classify mentions as referring or non-referring. If a mention is classified as referring, coders are then asked if that expression is `discourse-old` or `discourse-new` [53], and in the first case, to identify its antecedent (see Section 2.4 below). If the mention is classified as non-referring, coders have to assign it either to one of the GNOME categories of non-reference, or label it as idiomatic (11), or as an incomplete or fragmentary expressions (12).

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

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

(13) [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.

(14) 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?

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 (13) is marked as non-referring. Similarly, whereas we asked coders to mark individual noun phrases (*Charlie Rangel*, *Basil Paterson* and *David* in (14) above) as referring mentions 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 below. In

	RST	TRAINS	GNOME	PEAR
all mentions	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

Table 2: Distribution of non-referential mentions in ARRAU.

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 mentions in the entire corpus and in the four individual domains, overall and for each type of non-referring mentions. 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 later 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.

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.[10] 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. The annotator is first asked to mark cases in which the nominal is clearly in the scope of an operator such as a conditional (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)). Next, the annotator is asked to identify cases in which the nominal refers to a number of semantic objects such as substances (e.g., *gold*) whose genericity is left underspecified, as in (21) seen before or in (22). Finally, the annotator is asked whether the sentence in which the mention 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 ($\kappa = .82$) [26].

(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 pop-

[10] This is the scheme described in the most widely used version of the manual, version 4 from April 2000.

ulation 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 [22], but its annotation was only partially checked. One of the main revisions 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 mentions), 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 mentions are marked as either `discourse new` or `discourse old`. Discourse new mentions introduce new entities and thus are not marked as being coreferent with an entity already introduced (**antecedent**). For discourse-old mentions, an antecedent can be identified, either of type `phrase` (in case the antecedent was introduced using a nominal mention) or `segment` (not introduced by a nominal mention, for the cases of **discourse deixis**).[11] 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.

[11] 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.

	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%)

Table 3: Distribution of generic mentions in ARRAU.

Bridging anaphora. Annotating—indeed, identifying—bridging references in a reliable way is a difficult task [54, 55], which is one of the reasons why so few large-scale corpora for anaphora include this type of annotation [45]. The ARRAU guidelines for bridging anaphora are based on a series of experiments that started with the work of Vieira and Poesio [54, 55] and continued in the GNOME project [26]. Vieira and Poesio attempted to annotate the full range of bridging references, but only achieved very poor agreement. In GNOME, attempts were made to identify a subset of the relations that could be annotated reliably [26], 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 [26]. 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 [56];
- 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 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.

	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

Table 4: Distribution of bridging references in ARRAU.

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).[12]

(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.

[12] One exception is a very recent study [57], aimed at rule-based plural anaphora resolution for the patent domain.

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) 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 [58] to indicate the reference to abstract entities which have not been introduced in the discourse through a nominal mention,[13] 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 [31] and to resolve. Very few anaphoric annotation projects have attempted annotating discourse deixis in its entirety [31, 62]; more typical is a partial annotation, as in the work of Byron and Navarretta, who annotated pronominal reference to abstract objects [63, 64]; in ONTONOTES, where event anaphora was marked [43]; and in the work of Kolhatkar [65], that focused on so-called shell nouns. As a result, very few systems have attempted resolving this type of anaphors [66, 67, 68]

Discourse deixis was one of the ’difficult cases of anaphora’ on which the ARRAU project focused, and a number of annotation experiments were conducted [31], resulting in guidelines according to which

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

[13] For more extensive discussion of reference to abstract objects see [59, 60]; for empirical analysis of discourse deixis, see e.g., [61].

RST	TRAINS	GNOME	PEAR	TOTAL
631	862	73	67	1633

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

2. Coders choosing `segment` as the type of antecedent have to mark a sequence of (predefined) clauses

Artstein and Poesio [31] point out that measuring disagreement on this type of annotation requires making a number of assumptions, and that figures of α [69] 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 marked. It's worth noticing how the TRAINS sub-domain contains more than half the total cases of discourse deixis even though it's 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.)

Anaphoric ambiguity. A number of studies have shown that anaphoric expressions both in dialogue and text can be ambiguous [70, 71, 72, 73] A classic illustration is the example below, from the TRAINS corpus [70]. 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 [30, 71]. Other studies have shown that occurrences of *it* can be ambiguous between an expletive and a discourse deixis interpretation [74].

(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 mentions 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 [30]. 18 subjects were asked to annotate dialogues 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 mentions in the TRAINS corpus were marked as explicitly ambiguous. They also found however that a much higher percentage of mentions, up to 40%, were *implicitly* ambiguous—i.e., were annotated differently by different subjects. In [75] methods for computing agreement in a scheme allowing for ambiguity were proposed, based on developing extended distance metrics for α [69, 20]. Values of α between .58 and .67 were reported depending on the type of distance metric used and the choice of mentions.

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 mention: 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 mentions were identified, which is a very small fraction of the around 100,000 mentions in ARRAU Version 2; the results of [30] 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.

2.5. Annotation tool and markup scheme

ARRAU was annotated using the MMAX2 annotation tool [48]. 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

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 6: Distribution of ambiguity in the subdomains of ARRAU.

domain	ARRAU1			ARRAU2		
	documents	tokens	mentions	documents	tokens	mentions
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

Table 7: Corpus statistics for two releases of ARRAU

and bridging references, anaphoric information is encoded using MMAX2 **pointers** instead of set-based attributes.

Note that set-based annotation for identity anaphora (as in the ONTONOTES scheme) can be induced from such pointers in a straightforward way.

3. Two Versions of ARRAU

The first release of ARRAU [22] was made publicly available in 2008. The second release of ARRAU 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

error type	RST	GNOME	PEAR	TRAINS
missing antecedent for anaphoric mentions	332	46	49	35
non-referential mention as an antecedent	205	15	6	10
semantic type mismatch	813	64	25	34

Table 8: Enforcing annotation quality: inconsistency statistics for the first release of ARRAU, most common types of errors.

the TRAINS domain, since it provides a unique large collection of dialogues annotated with anaphoric information. More statistics for both releases of ARRAU are provided in Table 7.

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. 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-referential mention can not participate 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

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:

- 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 mentions 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-referential mentions 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-referential mention 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 mentions 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 8 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) below, the annotator has erroneously assigned an incorrect semantic type (`space`) to a mention *the dollar*. In (35), the annotator marked *That* as a `discourse old` mention, but failed to provide a suitable (segment) antecedent. In (36), the annotator marked a non-referring mention as an antecedent, not distinguishing between co-reference and other anaphoric phenomena. Finding such errors manualy can be very tedious, as it requires a carefull supervision of each mention and all its attributes. The availability of multiple annotation levels, on the contrary, allows for immediate listing of such mistakes.

(34) ... thus dumping [dollar]$_1^{abstract}$ demand... Japanese institutions are comfortable with [the dollar]$_1^{space}$ 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 [any due diligence]$_1^{non-referential}$ because of competitive reasons. If we had, [it]$_1$ might have scared us.

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]$_2$ favours 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, e.g. 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 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]$_1^{abstract}$ is."

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

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, refering 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) $[\text{Kellog}]_1^{organization}$'s spokesman said . . . "As $[\text{we}]_1^{person}$ 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* [73]. 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.

(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 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 byproduct 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 [45]—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 [19] and ONTONOTES [7, 9, 6]. 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.

	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	-	±	+

Table 9: Comparison across coreferentially annotated corpora

4.1. ARRAU vs. ACE vs. OntoNotes

Table 9 provides a summary of the most distinctive features of ARRAU as opposed to ACE and ONTONOTES.

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 [76], where several anaphoric relations are encoded for the same textual material, going beyond identity anaphora.[14]

Each nominal mention 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,

[14] The only comparable large-scale approach for English we are aware of is GECCO [77]: a corpus extensively annotated for different means of discourse cohesion in a semi-automatic way.

whereas the minimal span corresponds to the head noun or to the bare named entity for complex NE-nominals. With the latest development in the 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 [78, 79]. 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 [80]. Finally, minimal spans can be intrinsically difficult to extract for non-conventional documents, such as dialogue transcripts or social media, due to the low quality of the parsing technology for such data. 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, we focus on different types of noun phrases. In particular, we label mentions that do not participate in coreference chains: singletons and non-referring NPs. The ACE guidelines restrict the annotation scope to referentials[15], whereas in ONTONOTES only co-referential mentions are marked, not singletons. Our corpus statistics show that non-referring and singleton mentions account for up to one third of all the mentions. 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 mentions annotated provides material for training mention detection systems. Mention detection for OntoNotes [80, 81] is a non-trivial problem that is further aggravated by the fact that singletons are removed and thus the direct training becomes hardly possible.

Each mention is annotated in ARRAU with its basic morphological properties: number, gender and semantic class. This allows, again, for training mention-level classifiers to assign these features automatically. Similarly to minimal span, this task can be attempted via heuristics based on parse trees, however, one can expect

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

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.[16] The OntoNotes dataset, on the contrary, provides valuable gold annotations of low-level phenomena (for example, gold part-of-speech tags or parse trees), but does not, to our knowledge, provide deep discourse-level annotations apart from coreference.

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 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 dialogue were Byron's annotation of pronouns in the TRAINS corpus [63] and the annotation of part of the SHERLOCK corpus of task-oriented instructional dialogue included in GNOME and used for the study of anaphora and discourse structure reported in [38].[17]

In the years since the situation has improved. Corpora covering textual genres other than news now exist, such NLP4EVENTS of instructional manuals [84], and the GENIA corpus of biomedical text [85]. For dialogue, Müller annotated pronominal reference in the ICSI spoken multi-partner conversation corpus [49]. 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.

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

[17] Nissim *et al.*'s annotation of the Switchboard corpus [82] only provided the information status of mentions, not their antecedents if any. Navarretta had annotated pronominal anaphora in Danish and Italian dialogue [64] and Poesio *et al.* had annotated Italian MapTask dialogues [83].

And the recently created GECCO corpus [77] 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 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 [12], 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 [10], 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.[18]

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 don't know however whether these difficulties were in fact solved as we are not aware of any results regarding the reliability of these guidelines. Herbelot and Copestake [86] developed an interesting scheme, strongly rooted in the literature on genericity in formal linguistics [87], and still attempting to capture genericity and specificity but treating them as two separate two dimensions of classification, as done in [87]. 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.

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

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 [88]. Finally regarding manual annotation, we shall mention the recent and very interesting work by Friedrich *et al.* [89] 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 $\kappa = .56$.

We should also mention that in several of the studies above, the annotation effort was carried out in order to train automatical annotators of genericity. In addition, we will mention the work by Reiter and Frank [90].

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 [55] and GNOME corpus [25],[19] 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 [39, 54, 91].

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 [92]. 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 [11]. Nedoluzhko *et al.* distinguish, apart from `part` and `subset` relations,

[19] Nissim *et al.*'s annotation for information status of parts of the Switchboard corpus [82] 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.

- A funct relation covering function-value relations, as proposed in MATE [91]

- 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 [93], who annotated 11,000 NPs in 50 texts taken from the WSJ portion of ONTONOTES for information status, building on the scheme by [82] 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.

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 [94].

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 [31].

Prior to ARRAU, we will mention the seminal work by Byron, who annotated pronominals and demonstratives in the TRAINS-93 corpus, including abstract objects [63], and used the data to develop the first resolver of references to abstract objects we are aware of [66]; Navarretta, who in parallel with Byron carried out similar studies of abstract reference in Danish and Italian [64, 95]; and by Eckert and Strube [67], who analyzed references in dialogues 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' [74, 61], that was an important influence on our own work.

The most notable efforts 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 [96], and the work by Dipper and Zinsmeister [62] on abstract anaphora in German. More recenly, a very systematic analysis and annotation of another subset of discourse deixis, so called **shell nouns**, has been carried out by Kolhatkar and collaborators [68, 65].

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 [97] as part of the work on the Potsdam Commentary Corpus [98] is the only other coreference annotation scheme we are aware of that asks coders to mark ambiguity. The guidelines produced by Chiarcos and Krasavina [99] 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 [30] 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 [100], 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 mentionss have at least two interpretations selected by at least 2 players. A first, small subset of the Phrase Detectives corpus was recently released via LDC [101]. More recently, this line of research has led us to start the DALI project,[20] in which besides carrying out the development of more Games-With-A-Purpose to study anaphora, methods to compare and analyse these interpretations will also be developed.

5. 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.

[20] `http:www.dali-ambiguity.org`

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 focus on discourse deixis and bridging, thus, providing data for joint modeling of these phenomena.

The annotation scheme developed for ARRAU has already been adopted by other projects, for example, LIVEMEMORIES corpus of anaphora in Italian [102]. 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). The SENSEI corpus consists of annotations of online forums in English (from *The Guardian* newspaper) and Italian (from *La Repubblica* newspaper) following similar guidelines.

We believe that complex ARRAU annotations provides valuable data for the next generation of anaphora resolvers.

Acknowledgments

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; in part by the ERC Project DALI, ERC-Adg-2015.

References

[1] G. Durrett, D. Klein, Easy Victories and Uphill Battles in Coreference Resolution., in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1971–1982, 2013.

[2] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, D. Jurafsky, Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules, Computational Linguistics 39 (4) (2013) 885–916.

[3] A. Björkelund, J. Kuhn, Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 47–57, 2014.

[4] E. R. Fernandes, C. N. dos Santos, R. L. Milidiú, Latent Trees for Coreference Resolution, Computational Linguistics 40 (4) (2014) 801–835, ISSN 0891-2017.

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

[6] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, Y. Zhang, CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, in: Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012), Jeju, Korea, 2012.

[7] R. Weischedel, E. Hovy, M. Marcus, M. Palmer, R. Belvin, S. Pradhan, L. Ramshaw, N. Xue, OntoNotes: A Large Training Corpus for Enhanced Processing, in: J. Olive, C. Christianson, J. McCary (Eds.), Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation, Springer, 2011.

[8] M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, Y. Versley, SemEval-2010 Task 1: Coreference Resolution in Multiple Languages, in: Proc. SEMEVAL 2010, Uppsala, 2010.

[9] S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, N. Xue, CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011), Portland, Oregon, 2011.

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

[11] A. Nedoluzhko, J. Mirokvský, P. Pajas, The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague Dependency Treebank, in: Proceedings of the Linguistic Annotation Workshop, 108–111, 2009.

[12] E. Hinrichs, S. Kübler, K. Naumann, A Unified Representation for Morphological, Syntactic, Semantic and Referential Annotations, in: ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, Ann Arbor, 2005.

[13] L. Carlson, D. Marcu, M. E. Okurowski, RST Discourse Treebank LDC2002T07, 2002.

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

[15] S. Neumann, Contrastive Register Variation. A quantitative Approach to the Comparison of English and German, Mouton de Gruyter, 2013.

[16] Y. Grishina, M. Stede, Knowledge-lean projection of coreference chains across languages, in: Proceedings of the 8th Workshop on Building and Using Comparable Corpora, 14–22, 2015.

[17] O. Uryupina, M. Poesio, Domain-specific vs. Uniform Modeling for Coreference Resolution, in: Proceedings of the Language Resources and Evaluation Conference, 2012.

[18] K. Kunz, E. Lapshinova-Koltunski, J. M. Martínez, Beyond Identity Coreference: Contrasting Indicators of Textual Coherence in English and German, in: Proceedings of the Workshop on Coreference Resolution beyond OntoNotes (CORBON-2016), 23–31, 2016.

[19] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassell, R. Weischedel, The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation, in: Proceedings of the Language Resources and Evaluation Conference, 2004.

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

[21] J. Carletta, Assessing agreement on classification tasks: the Kappa statistic, Computational Linguistics 22 (2) (1996) 249–254.

[22] M. Poesio, R. Artstein, Anaphoric annotation in the ARRAU corpus, in: Proc. of LREC, Marrakesh, 2008.

[23] O. Uryupina, R. Artstein, A. Bristot, F. Cavicchio, K. J. Rodriguez, M. Poesio, ARRAU: Linguistically-Motivated Annotation of Anaphoric Description, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016.

[24] M. Poesio, The GNOME Annotation Scheme Manual, University of Edinburgh, HCRC and Informatics, Scotland, fourth version edn., available from `http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno\_manual\_4.htm`, 2000.

[25] M. Poesio, The MATE/GNOME Scheme for Anaphoric Annotation, Revisited, in: Proceedings of SIGDIAL, Boston, 2004.

[26] M. Poesio, Discourse Annotation and Semantic Annotation in the GNOME Corpus, in: Proceedings of the ACL Workshop on Discourse Annotation, Barcelona, 72–79, 2004.

[27] H. H. Clark, Bridging, in: Proceedings of TINLAP, 1975.

[28] M. Poesio, R. Stevenson, B. Di Eugenio, J. M. Hitzeman, Centering: A Parametric Theory and its Instantiations, Computational Linguistics 30 (3) (2004) 309–363.

[29] M. Poesio, R. Mehta, A. Maroudas, J. Hitzeman, Learning to Solve Bridging References, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, Barcelona, 143–150, 2004.

[30] M. Poesio, R. Artstein, The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account, in: A. Meyers (Ed.), Proceedings of ACL Workshop on Frontiers in Corpus Annotation, 76–83, 2005.

[31] R. Artstein, M. Poesio, Identifying reference to abstract objects in dialogue, in: D. Schlangen, R. Fernandez (Eds.), Proc. of BRANDIAL, Potsdam, 2006.

[32] L. Carlson, D. Marcu, M. E. Okurowski, Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory, in: J. Kuppevelt, R. Smith (Eds.), Current Directions in Discourse and Dialogue, Kluwer, 85–112, 2003.

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

[34] M. Palmer, D. Gildea, Kingsbury, The Proposition Bank: A Corpus Annotated with Semantic Roles, Computational Linguistics 31 (1) (2005) 71–106.

[35] N. V. Loukachevitch, G. B. Dobrov, A. A. Kibrik, M. V. Khudyajova, A. S. Linnik, Factors in Referential Choice, in: Proceedings of Dialogue, Moscow, 2011.

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

[37] M. Poesio, Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results, in: Proceedings of the 2nd Language Resources and Evaluation Conference, Athens, 211–218, 2000.

[38] M. Poesio, A. Patel, B. Di Eugenio, Discourse Structure and Anaphora in Tutorial Dialogues: an Empirical Analysis of Two Theories of the Global Focus, Research in Language and Computation 4 (2006) 229–257, special Issue on Generation and Dialogue.

[39] R. J. Passonneau, Instructions for Applying Discourse Reference Annotation for Multiple Applications (DRAMA), unpublished manuscript, 1997.

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

[41] K. van Deemter, R. Kibble, On Coreferring: Coreference in MUC and Related Annotation Schemes, Computational Linguistics 26 (4) (2000) 629–637, squib.

[42] I. Hendrickx, G. Bouma, F. Coppens, W. Daelemans, V. Hoste, G. Kloosterman, A.-M. Mineur, J. Van Der Vloet, J.-L. Verschelde, A Coreference Corpus and Resolution System for Dutch, in: Proceedings of the Language Resources and Evaluation Conference, 2008.

[43] S. Pradhan, L. Ramshaw, R. Weischedel, J. MacBride, L. Micciulla, Unrestricted Coreference: Indentifying Entities and Events in OntoNotes, in: in Proceedings of the IEEE International Conference on Semantic Computing (ICSC), 2007.

[44] N. Ge, J. Hale, E. Charniak, A statistical approach to anaphora resolution, in: Proceedings of the 6th Workshop on Very Large Corpora, 161–170, 1998.

[45] M. Poesio, S. Pradhan, M. Recasens, K. Rodriguez, Y. Versley, Annotated Corpora and Annotation Tools, in: M. Poesio, R. Stuckardt, Y. Versley (Eds.), Anaphora Resolution: Algorithms, Resources and Applications, chap. 4, Springer, 2016.

[46] M. Amoia, K. Kunz, E. Lapshinova-Koltunski, Discontinuous Constituents: a Problematic Case for Parallel Corpora Annotation and Querying, in: Proceedings of RANLP2011-Workshop on Annotation and Exploitation of Parallel Corpora (AEPC 2), 2–10, 2011.

[47] M. Poesio, H. Rieser, Completions, Coordination, and Alignment in Dialogue, Dialogue and Discourse 1 (1) (2010) 1–89.

[48] C. Müller, M. Strube, Multi-Level Annotation of Linguistic Data with MMAX2, in: S. Braun, K. Kohn, J. Mukherjee (Eds.), Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods, vol. 3 of *English Corpus Linguistics*, Peter Lang, 197–214, 2006.

[49] M.-C. Müller, Fully Automatic Resolution of It, This And That in Unrestricted Multy-Party Dialog, Ph.D. thesis, Universität Tübingen, 2008.

[50] S. Pradhan, X. Luo, M. Recasens, E. Hovy, V. Ng, M. Strube, Scoring coreference partitions of predicted mentions: A reference implementation .

[51] O. Uryupina, M. Kabadjov, M. Poesio, Detecting non-reference and non-anaphoricity, in: M. Poesio, R. Stuckardt, Y. Versley (Eds.), Anaphora Resolution: Algorithms, Resources and Applications, chap. 13, Springer, 2016.

[52] A. Björkelund, R. Farkas, Data-driven Multilingual Coreference Resolution using Resolver Stacking, in: Joint Conference on EMNLP and CoNLL - Shared Task, Association for Computational Linguistics, Jeju Island, Korea, 49–55, URL `http://www.aclweb.org/anthology/W12-4503`, 2012.

[53] E. F. Prince, The ZPG letter: subjects, definiteness, and information status, in: S. Thompson, W. Mann (Eds.), Discourse description: diverse analyses of a fund-raising text, John Benjamins, 295–325, 1992.

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

[55] M. Poesio, R. Vieira, A Corpus-Based Investigation of Definite Description Use, Computational Linguistics 24 (2) (1998) 183–216, also available as Research Paper CCS-RP-71, Centre for Cognitive Science, University of Edinburgh.

[56] N. N. Modieska, Resolving other anaphors, Ph.D. thesis, University of Edinburgh, 2003.

[57] A. Burga, S. Cajal, J. Codina-Filba, L. Wanner, Towards Multiple Antecedent Coreference Resolution in Specialized Discourse, in: N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings

of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France, ISBN 978-2-9517408-9-1, 2016.

[58] B. L. Webber, Structure and Ostension in the Interpretation of Discourse Deixis, Language and Cognitive Processes 6 (2) (1991) 107–135.

[59] E. Schuster, Pronominal reference to events and actions: Evidence from naturally-occurring data, LINC LAB 100, University of Pennsylvania, Dept. of Computer and Information Science, Philadelphia, 1988.

[60] N. Asher, Reference to Abstract Objects in English, D. Reidel, Dordrecht, 1993.

[61] J. K. Gundel, M. Hegarty, K. Borthen, Cognitive Status, Information Structure, and Pronominal Reference to Clausally Introduced Entities, Journal of Logic, Language and Information 12 (3) (2003) 281–299.

[62] S. Dipper, H. Zinsmeister, Annotating Abstract Anaphora, Language Resources and Evaluation 46 (1) (2012) 37–52.

[63] D. Byron, J. Allen, Resolving Demonstrative Anaphora in the TRAINS-93 Corpus, in: Proceedings of the Second Colloquium on Discourse, Anaphora and Reference Resolution, University of Lancaster, 1998.

[64] C. Navarretta, Abstract Anaphora Resolution in Danish, in: L. Dybkjaer, K. Hasida, D. Traum (Eds.), Proc. of the 1st SIGdial Workshop on Discourse and Dialogue, ACL, 56–65, 2000.

[65] V. Kolhatkar, Resolving Shell Nouns, Ph.D. thesis, University of Toronto, 2014.

[66] D. Byron, Resolving pronominal references to abstract entities, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 80–87, 2002.

[67] M. Eckert, M. Strube, Dialogue Acts, Synchronising Units and Anaphora Resolution, Journal of Semantics 17 (1) (2000) 51–89.

[68] V. Kolhatkar, G. Hirst, Resolving “this-issue” anaphora, in: Proceedings of EMNLP-CONLL, Jeju Island, Korea, 1255–1265, 2012.

[69] K. Krippendorf, Content Analysis: An Introduction to Its Methodology, second edition, chap. 11, Sage, Thousand Oaks, CA, 2004.

[70] M. Poesio, U. Reyle, Underspecification in Anaphoric Reference, in: E. T. H. Bunt, I. van der Sluis (Ed.), Proceedings of the Fourth International Workshop on Computational Semantics, Tilburg, 286–300, 2001.

[71] M. Poesio, P. Sturt, R. Arstein, R. Filik, Underspecification and anaphora: Theoretical Issues and Preliminary Evidence, Discourse Processes 42 (2) (2006) 157–175.

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

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

[74] J. K. Gundel, N. Hedberg, R. Zacharski, Pronouns Without Explicit Antecedents: How do We Know When a Pronoun is Referential?, in: Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium, 2002.

[75] M. Poesio, R. Artstein, Annotating (Anaphoric) Ambiguity, in: Proceedings of the Corpus Linguistics Conference, Birmingham, 2005.

[76] A. Nedoluzhko, J. Mírovskỳ, R. Ocelák, J. Pergler, Extended coreferential relations and bridging anaphora in the Prague Dependency Treebank, in: Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India, 1–16, 2009.

[77] E. Lapshinova-Koltunski, K. A. Kunz, Annotating Cohesion for Multilingual Analysis, in: Proc. of the LREC ISO Workshop on Interoperable semantic resources, 57–64, 2014.

[78] D. Zhekova, S. Kübler, Machine Learning for Mention Head Detection in Multilingual Coreference Resolution, in: RANLP, 747–754, 2013.

[79] H. Peng, K.-W. Chang, D. Roth, A Joint Framework for Coreference Resolution and Mention Head Detection, in: Proceedings of the Conference on Natural Language Learning, 2015.

[80] J. K. Kummerfeld, M. Bansal, D. Burkett, D. Klein, Mention Detection: Heuristics for the OntoNotes annotations, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, Association for Computational Linguistics, Portland, Oregon, USA, 102–106, 2011.

[81] O. Uryupina, A. Moschitti, Multilingual Mention Detection for Coreference Resolution, in: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP'13), 2013.

[82] M. Nissim, S. Dingare, J. Carletta, M. Steedman, An annotation scheme for information status in dialogue, in: Proceedings of the Language Resources and Evaluation Conference, 2004.

[83] M. Poesio, R. Delmonte, A. Bristot, L. Chiran, S. Tonelli, The VENEX corpus of anaphoric information in spoken and written Italian, available online at `http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf`, 2004.

[84] L. Hasler, C. Orasan, K. Naumann, NPs for Events: Experiments in coreference annotation, in: Proceedings of the Language Resources and Evaluation Conference, 2006.

[85] N. L. T. Nguyen, J.-D. Kim, J. Tsujii, Challenges in Pronoun Resolution System for Biomedical Text, in: Proceedings of the Language Resources and Evaluation Conference, 2008.

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

[87] G. N. Carlson, F. J. Pelletier (Eds.), The Generic Book, University of Chicago Press, 1995.

[88] A. Nedoluzhko, Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank, in: Proceedings of the Linguistic Annotation Workshop, 103–111, 2013.

[89] A. Friedrich, A. Palmer, M. P. Sorensen, M. Pinkal, Annotating genericity: a survey, a scheme, and a corpus, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2015.

[90] N. Reiter, A. Frank, Identifying Generic Noun Phrases, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 40–49, 2010.

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

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

[93] K. Markert, Y. Hou, M. Strube, Collective Classification for Fine-grained Information Status, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, Jeju island, Korea, 2012.

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

[95] C. Navarretta, Pronominal types and abstract reference in the Danish and Italian DAD Corpora, in: C. Johansson (Ed.), Proc. of the Second Workshop on Anaphora Resolution (WAR II), 63–71, 2008.

[96] M. Recasens, Discourse deixis and coreference: evidence from AnCoRa, in: C. Johansson (Ed.), Proc. of the 2nd Workshop on Anaphora Resolution (WAR-2), 2008.

[97] O. Krasavina, C. Chiarcos, The Potsdam Coreference Scheme, in: Proc. of the 1st Linguistic Annotation Workshop, 156–163, 2007.

[98] M. Stede, The Potsdam Commentary Corpus, in: Proc. of the ACL Workshop on Discourse Annotation, 2004.

[99] C. Chiarcos, O. Krasavina, Annotation Guidelines PoCoS—Potsdam Coreference Scheme: Core Scheme, Draft 0.912, University of Potsdam and Humboldt University, Potsdam and Berlin, 2005.

[100] M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, L. Ducceschi, Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation, ACM Transactions on Intelligent Interactive Systems 3 (1) (2013) 1–44.

[101] J. Chamberlain, M. Poesio, U. Kruschwitz, Phrase Detectives Corpus 1.0: Crowdsourced Anaphoric Coreference, in: Proceedings of the Language Resources and Evaluation Conference, Portoroz, Slovenia, 2016.

[102] K.-J. Rodriguez, F. Delogu, Y. Versley, E. Stemle, M. Poesio, Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus, in: Proceedings of the Language Resources and Evaluation Conference, 2010.

6. Response to reviewers

We are very grateful to the reviewers for their helpful suggestions. The heavily revised version of the paper we are submitting addresses, we hope, all the issues raised by the reviewers. In summary, we

- Substantially expanded the description of the annotation guidelines in Section 2, providing reliability results for all novel forms of annotation;
- Substantially expanded the discussion of alternative projects, comparing in some detail our approach with theirs;
- Substantially expenaded the discussion of our measures to improve the annotation consistency;
- Generally revised and clarified various parts of the text that the reviewers found unclear or sloppily written.

We detail these changes below. We have slightly reformatted the original reviews to allow for the issue-response section. No comments/concerns were omitted or modified through this reformatting.

6.1. Reviewer 1

We wish to thank the reviewer for the many helpful suggestions and the many useful references.

0) The overall comment is related to the issue of inter-annotator agreement and general information on the annotators. The authors do not mention if there were several annotators and what his/her/their background was. If there is another paper describing this, then a reference should be provided. The same problem is the report on the inter-annotator agreement, which is missing for this paper.

Response: *Virtually all the reliability studies leading to the* ARRAU *guidelines were reported in publications. We included in the paper the inter-annotator agreement results, and added references to the original studies for the details.*

1. Introduction

1) Genre-dependent variation of coreference-related phenomena has been analysed also in other studies on coreference, see for instance Grishina & Stede (2015) or those with deeper linguistic analyses, such as Kunz et al. (2016), Kunz & Lapshinova-Koltunski (2015), Neumann (2013, p. 254-255) and Taboada & Gomez-Gonzalez (2012).

Response: *We have greatly expanded the discussion to cover several relevant studies, such as the work by Kunz et al on the GECCO corpus.*

2) There are further resources that are annotated for anaphora beyond identity relations (e.g. briging, reference to abstract entities, events, etc.). For instance, Prague Dependency Treebank (PDT, Zirkanova et al., 2015) ? a corpus of Czech, GECCo (Lapshinova & Kunz, 2014; Martinez et al. 2016).

Response: *We have included an extensive discussion of all these projects and the relation between their coding schemes and* ARRAU*'s in the Related Work section (page 25).*

3) Related phenomena: in the very beginning of the paper, the authors mention that the annotation schemes includes non-referring expressions without clarifying what exactly is meant here. Later on, in the paper (page 7), the definition follows -'nominal expressions that do not denote any specific entity'. It would be better to include this definition already in the beginning to make it clear for the reader. The authors also use the notion of 'identity coreference'. According to some frameworks, coreference, as opposed to the other means of cohesion, always express identity (see e.g. Halliday & Hasan, 1976). So, a reference to the framework the authors base on (probably OntoNotes) would be an advantage.

Response: *(i) The* ARRAU *guidelines cover a lot of phenomena, discussed in detail in Section 2. Introducing each of them properly in the Introduction would result in an extremely lengthy Section, so we tried to keep definitions there to a minimum. (ii) We have however rephrased the discussion of identity coreference as requested by Reviewer 1, linking the terminology to that used in ACE/OntoNotes more explicitly (Page 4)*

4) Ambiguity (as presented on page 3): It is not clear why coreference chains as sets cannot support ambiguities ? every element could be a member of two (or more) different sets?

Response: *This was done in part because of the semantics of markup schemes, in part because of the annotation tool, MMAX2. In set-based markup schemes (e.g. OntoNotes), it is impossible to assign two sets to a single mention, as each mention is mapped to a unique set id. The only possibility would be to duplicate a mention and assign different sets to these two mentions with the same span. This, however, is not a reasonable solution for two reasons: (i) it conveys the wrong semantics, creating two same-span mentions instead of a single ambiguous one and (ii) more practically, the scoring algorithm would misinterpret such gold annotation. We think that this, however, is a fairly technical point, so we haven't included it in the* *paper.*

5) The authors claim 'This second release of ARRAU, therefore, contains cleaner annotations. This is in contrast with other corpora, where subsequent releases, if any, expand the text collection and only fix occasional manually attested errors.' The problem of obtaining 'cleaner' annotations with systematic techniques is an important issue for corpus annotation tasks. The authors suggest techniques to enforce annotation consistency. It is interesting to know if there are further studies in the area of automatic annotation consistency check that the authors could cite as related work?

Response: *We are not aware of any systematic NLP approach to improving the annotation consistency apart from controlling for the inter-annotator agreement, and we are almost certain that there are no published discussions of such methods in the literature on coreference annotation. We believe, however, that this is a very important issue and expect our study to bring attention to this problem–in fact, we think that our methodology in this regard is an important contribution and that the consistency checks we applied are a key part of the improvements between ARRAU 1 and ARRAU 2. We have included this discussion on Page 4.*

6) Reference: is there a reference to the ARRAU guidelines? There is a problem with the format of the references at the end of page 3 (The two versions of the ARRAU corpus have first been presented at the Language Resources and Evaluation conference Poesio and Artstein [12] ?)

Response: *The two annotation manuals used in ARRAU - one for text, the other for spoken dialogue - are included with the corpus; we added to the paper this information. We also fixed the problem with the citations identified by the Reviewer, thanks for spotting it.*

Section 2

1) Page 5: The authors mention that 'ARRAU contains documents from four domains, representing different genres, mostly not covered by other corpora'. What is the difference between domains and genres? In some approaches, these notions are used for the same concept, in others ? they refer to different concepts. Do the authors differentiate between domains and genres, if yes, how? It would be also useful to already mention the genres included into the corpus directly here, or even earlier in the paper.

Response: *Our intention is to use the term 'genre' to refer to a literary genre (fiction, news, task-oriented dialogue), and to use the term 'domain' to refer to one of the sub-corpora of ARRAU. We revised the paper to ensure this distinction is* *made consistently. We also moved the presentation of the genres covered in ARRAU, and relative domains, at the beginning of Section 2, as suggested by Reviewer 1.*

2) Page 6: References to the examples are different from the example numeration (1a and 1b ? missing in the numeration). The reference to example 3 is missing. There is a colon instead. It would be better to use a consistent reference style in the paper, so use: This allows for correct labeling of discontinuous noun phrases, see (3) instead of ?This allows for correct labeling of discontinuous noun phrases:?.

Response: *Thanks to the Reviewer for spotting this. The problem has been fixed, and the notation made more consistent throughout.*

3) Discontinuous constituents: there is another study on discontinuous constituents which might be interesting for the authors ? Amoia et al. (2011). For the reference [22], the year is missing.

Response: *Thanks a lot for bringing this paper to our attention. It is very relevant for our study, since only very few projects allow dor discontinuous chunks at all. We have included the reference.*

4) Page 7: a linguistic illustration (example) of non-referring, discourse-new and discourse-old reference would be an advantage here.

Response: *The terms discourse-new and discourse-old from Prince 1992 are universally adopted in the literature on anaphora and coreference, but we did add a reference to Prince and an example of non-referring* NP *earlier on.*

5) Table 2: It is not clear directly from the design of the paper that non-bold categories are subtypes of non-referential markables. So, if possible, the authors should consider a different way of presentation here. The same problem is observed for Table 3 representing generic subtypes.

Response: *We revised the Tables and added some explanation trying to make them clearer.*

6) ?Table 2 shows the distribution of various types of non-referential mentions in the whole corpus and in the four individual domains.? - do the authors mean subcorpora here?

Response: *As discussed earlier, yes, we do use the term 'domain' to mean 'sub-corpus', and we revised the text to make this clearer.*

In the same paragraph, the authors discuss differences in the distribution of categories across subcorpora representing domains (or genres?). However, these are absolute numbers for frequencies. How comparable are they? Are all the parts of the corpus balanced? This paragrapoh also contains information on one of the subcorpora (GNOME) ? it is the first time the authors provide an explanation on this part. As mentioned above, it would be more useful and readable to have information on all the genres included already earlier in the paper.

Response: *As suggested by the reviewer, we moved the discussion of genres and domains in ARRAU to the beginning of Section 2.*

7) Page 9: References to examples 16 and 18 are missing in the text. Format: ?whose genericity is left underspecified, as in 17? ? ? as in (17).

Response: *All examples are now referenced in the text, and the format of the references to the examples has been fixed.*

Table 3 contains information on the distribution of various generic markables in the corpus. However, a reference and any explanation in the text are missing. In this context,a study by Friedrich and Pinkal (2015) on discouse-sensitive behaviour of generics might be interesting. There is another study on generics (however, on Czech) by Nedoluzhko (2013), which might be of relevance for the authors.

Response: *The table is now discussed in the text. The study by Friedrich and Pinkal and the work by Nedoluzhko are now discussed in Section 4.*

8) Range of relations: can a mention be ambiguous in the range of relations? Do the authors consider such cases (if any available in the corpus)?

Response: *Potentially yes - a bridging reference could be marked as being marked by two different relations to the same anchor. In practice no such case was found by our coders in the corpus.*

9) Anaphoric ambiguity: the possibility to mark something as ambiguous probably solves some problems in a low inter-annotator agreement. Did the authors measure how annotators agree on ? if something is ambiguous or not??

Response: *Yes, this is indeed the case. We carried out two studies of agreement on ambiguity marking as part of the ARRAU project, they are now both discussed in the text.*

Section 4

1) Related Work: ARRAU vs. Other Anaphoric Corpora ? ? Other...

Response: *This has been fixed.*

2) page 15: in the footnote section, something went wrong in the Latex code, as there is a cut sentence.

Response: *One line of the foonote 'spilled over' in the next page. This has been fixed.*

6.2. Reviewer: 2

The worse "near redundancy" relates to the presentation of conjoined NPs. At the bottom of p.6, they are presented as "discontinuous mentions" in the discussion of ANNOTATED MENTIONS (Section 2.3). It is then asserted that they have decided to DISALLOW discontinuous minimal spans. But it is never said whether this is done by making the spans in Exs. 3 & 4 CONTINUOUS, by including the "and", or by just discarding all such discontinuous tokens.

The next section (Section 2.4) presents coordinations, as in Ex. 11, as being annotatable "non-referential" mentions, where "non-referential mentions" don't "denote any specific entity". How these differ from the "discontinuous mentions" in Ex. 3&4 is not discussed (nor is WHY coordinated NPs don't denote any specific entity).

Finally, the next section (Section 2.5) discusses "plural anaphors", saying that the use of "they" to refer to the set John, Mary should be the same, whether the set derives from the coordinated NP in (21a) or the separate arguments to "met" in (21b).

Response: *We revised the discussion of discontinuous mentions, non-referring NPs, and plural anaphora, to clarify that these are orthogonal issues. Discontinuous mentions are an annotation device introduced by Mueller and Strube in MMAX2 to handle mentions in dialogue expressed through non-continuous sequences of tokens. They are not used in ARRAU for all coordinated NPS, but only for those in which the two conjuncts share part of the tokens (as in THE FRIENDS AND ENEMIES OF THIS COUNTRY). The new text also explains why coordinated NPs are marked in ARRAU as non-referring.*

- Claim in Section 1 that procedures have been implemented for improving annotation quality and consistency. I didn't see anything that addressed improving annotation quality, and with respect to consistency, a few "consistency checks" are listed in Section 3.1, but a reader is left to figure out for themselves how something like "minimal and maximal spans, genericity and referentialitydo" begin annotated for all documents "enforces consistency across domains". I, for one, haven't a clue. Several of the items on the list in Section 3.1 seem to have nothing to with "enforcing consistency" (e.g., the re-annotation of unspecified attributes). If consistency CAN be claimed for the ARRAU corpus, then the authors need to take it seriously and tell us more about HOW, including the extent to which each type of consistency check has caught inconsistencies. One wants othe corpus developers to take heed of this.

Response: *We have considerably expanded Section 3, including a considerable number of illustrative examples and providing statistics on the number of problematic cases identified via the consistency checking procedure. To our knowledge, very little, if any, effort is dedicated to discussing formal data/annotation consistency in the NLP community. We are far from suggesting any final solution to this problem, however, we hope to bring it to the community's attention and suggest at least some useful steps in the right direction. Therefore, while we agree with the general concern expressed by the reviewer that our contribution can hardly qualify as a "methodology", it surely suggests some "good practice". We have therefore adjusted our terminology.*

- Claim in Section 2 that "separate guidelines were developed for each of the different genres". If there are separate guidelines for each genre, then readers (i.e., other researchers) need to know how they differ. For example, do the separate guidelines all achieve the same ends, only differing in terms of the examples they use? Or do some genres have types of referring expressions or anaphors that are absent from other genres.

Response: *The two coding manuals are distributed with the corpus, one for the dialogue genres, the other for the text genres - we now say that in the text. We also say that the coding scheme used is exactly the same, the manuals only differ in that different types of examples are used in each case.*

- Claim in Section 3 that "we have identified a number of truly challenging cases of coreference", with no further explanation or demonstration of what they are or why they are challenging.

Response: *We have now added examples and a discussion of why they are challenging.*

Section 1 (p.3) The phrase "rst domain" doesn't make sense. The authors are talking about the RST Corpus, which is annotated (as they later note) over a small subset of the Penn Wall Street Journal corpus.

Response: *We added to the paper a discussion of the terminology we used - using the term 'genre' to refer to genres in the traditional sense (e.g., fiction, scientific non-fiction) and the term 'domain' to indicate the sub-corpora of ARRAU (RST, GNOME, PEAR, and TRAINS)*

Section 2.3 (p.6)

- The authors refer to (1a) and (1b), but there is only example (1).

Response: *This has been fixed*

- I don't know what is meant by saying "Singletons are also marked as mentions that are part of coreference chains". The authors haven't defined what THEY mean by "singleton", but it is usually used to refer to an entity that is introduced but never mentioned again. So in what way are they part of "coreference chains"? Are these single link chains?

Response: *This has been clarified*

The sentence after Ex (1) says that "... we mark all nominal mentions, including singletons and non-referring NPs". What does this say about singletons that wasn't said in the previous sentence?

Response: *We have expanded the discussion of the guidelines and also removed the redundant sentence.*

Section 2.4 (p.8)

- In what way is the NP in Ex 10 ("half of those polled") non-referential? One can certainly use an coreferential anaphor to refer to this subset, and the non-coreferential anaphor "the others" in the same sentence certainly refers to that set in taking its complement.

Response: *The decision to mark all quantifiers as non-referring is indeed controversial, but was taken as we were observing a high rate of disagreement among annotators on whether to mark quantifiers as non-referring or referring. We added an explanation of this point to Section 2.*

Section 2.4 (p.9)

The authors should indicate what they take to be the "individual quantifier" in (13), the "temporal quantifier" in (14), and the modal in (15). It's certainly not clear to this reader what the authors are referring to.

Response: *We added explanations for all the labels*

Section 2.5 (Table 3, p.10)

What is the label "episodic-no"? There is no other mention of "episodic" anywhere in the paper. What are the labels "underspecified-decease" and "underspecified-replicable"? Again, there is not other mention of "decease" or "replicable" anywhere in the paper.

Response: `undersp-replicable` *is used for products such as books or medicines whose proper name can refer either to an abstract object ('Moby Dick') or to the specific book purchased by somebody.* `undersp-disease` *is to be* *used for illnesses. The category `episodic-no` was introduced in GNOME but virtually unused in ARRAU.*

Section 3 (p.12)

What is presented is NOT a methodology for enforcing consistency but rather a set of consistency checks whose coverage is unclear. I commend the authors for even considering the importance of ”consistency”, but they haven’t presented a methodology that is usable by others, not have they indicated what aspects of consistency they have NOT been able to address yet. (If what they are claiming to have eliminated all inconsistencies from the corpus, then one would like to be shown some external verification of such a claim.)

Response *This is a very important point related to one of the ’main issues’ raised by the same reviewer. We have discussed it above.*

6.3. Reviewer: 3

The paper describes an effort to create a relatively large scale corpus annotated with several linguistic phenomena related to anaphora and coreference. Such a resource is desperately needed by the community as most recent research on coreference resolution tends to overfit on the OntoNotes data (this definitely started with Durrett et al. (EMNLP 2013) who gained most by using so-called lexical features). So, a corpus with a variety of genres is definitely needed. A corpus which does not omit more difficult cases is also welcome.

I have reservations about the paper, however:

1. NLE does not seem to be a good fit for the description of a corpus annotation effort. If the authors would report experiments on the data and establish baselines, then this may be different. As it is, I’d recommend to submit this paper to LRE instead.

Response: *While we agree that LRE would in general be a better fit for a paper summarizing a data annotation project, we believe that ARRAU is of very high relevance for the audience of the special issue on Knowledge-Rich Coreference Resolution, e.g., the coding of bridging references and of genericity. We decided to avoid reporting experimental results in order to keep the paper within a reasonable page count.*

2. A journal paper is not the right place to omit details. Many important details, however, are missing. E.g. reliability. The authors give hints that the annotation has been performed reliably. But they do not give evidence that the data to be released is annotated reliably in fact. On p.7 they say that the coding scheme for mention properties has been tested for GNOME (referring to a paper published in 2000, many years before two of the authors of the current paper came up with a

clarification for metrics to compute reliability ...). GNOME encompasses, however, only one of the four genres in ARRAU. How about the other three genres? Did the authors check reliability here? On p.9 reliability is mentioned in the context of annotating genericity. No results provided. – The authors also do not mention whether the annotation of coreference, bridging, etc. has been performed reliably. Do I have to check the LREC papers to get this information? I believe that a journal paper should be self-contained.

Response *A great number of reliability studies were carried out both in GNOME and in ARRAU for all aspects of the annotation. We have now added reliability results for all the categories discussed as well as references to the original papers for additional details.*

3. The authors claim that they annotated a wide range of phenomena. But they fail to compare their annotation decisions with related efforts. E.g. genericty. How does the authors' annotation of genericity compare with the ACE annotation of genericity? E.g. bridging. Hou et al. (ACL 2012) describe annotating bridging on top of OntoNotes. How does the authors' annotation of bridging compare with Hou et al.? For once, Hou et al. found out that the bridging phenomenon is much broader then claimed previously, while the authors of the current paper severerly restrict the phenomenon. Don't misunderstand me. I don't want you to annotate like Hou et al., because one can disagree with their decisions (Nedoluzhko and colleagues do it differently as are Grishina and colleagues and Lapshinova-Koltunski and colleagues, see CORBON 2016). I want you, however, to discuss differences and justify your decisions.

Response: *We have greatly extended Section 4 to include a discussion of all related work we are aware of on genericity, bridging, discourse deixis, and ambiguity*

I did not find the annotation guidelines. The webpage mentioned in footnote 1 has a lot of information, but I did not find the guidelines.

Response: *The guidelines are included with the corpus. We added this information to the paper.*

Smaller issues:

The paper uses frequently forward references to unspecified points in the paper, eg. "see below". Avoid this.

Quite a few typos for a journal submission.

The references are a mess.

Response: *We have improved the presentation, proofreading the paper and* *cleaning up the references.*

6.4. Reviewer: 4

Comments to the Author

Just a few typos, and queries:

Bottom p. 3: author names occur together with bracketed references.

P. 4: if possible, direct links to the corpus and to the guidelines would have been helpful.

Response: *We now added to the text the information that the guidelines are distributed with the corpus.*

P. 8: The bracketing in (11) suggests the opposite of what you say in text right below.

p. 9: In the enumeration (12-19), would you say that (16) is an instance of being in the scope of a verb? The use of 'in the scope of' in this section seems a bit casual, lingering between what in logic would be 'in the scope of' and something like 'string adjacent to'.

Response: *The intended interpretation of 'in the scope of' is in the logical sense.*

p. 10, 2 lines from bottom: provides -¿ provide.

p. 13: 3 lines from end of section 3, example -¿ examples. 'clash' should perhaps rather be 'contrast'?

p. 15. end of section 4. 'on the one hand'

Response: *We have addressed all the minor issues raised by the reviewer.*