Annotating a broader 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 Computer Science and Electronic Engineering, University of Essex*
uryupina@gmail.com, artstein@ict.usc.edu, lucanto137@libero.it,
federica.cavicchio@gmail.com, delogu@coli.uni-saarland.de
kepa.rodriguez@yadvashem.org.il, poesio@essex.ac.uk

This paper presents ARRAU, a multi-genre corpus of anaphora created to provide much needed data for the next generation of coreference and 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: thorough annotation of mention boundaries (minimal/maximal spans, discontinuous mentions); full annotation of the (non) referentiality and (non) anaphoricity of mentions, distinguishing between several categories of non-referentiality and annotating non-anaphoric mentions; the annotation of a variety of mention attributes, ranging from morphosyntactic parameters to semantic category; the mark-up of genericity; the annotation of a wide range of anaphoric relations, including also bridging relations and discourse deixis; and, finally, the annotation of anaphoric ambiguity. The current version of the dataset contains 350K tokens and is publicly available from LDC.

Annotating a broader 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 Computer Science and Electronic Engineering, University of Essex*
uryupina@gmail.com, artstein@ict.usc.edu, lucanto137@libero.it,
federica.cavicchio@gmail.com, delogu@coli.uni-saarland.de
kepa.rodriguez@yadvashem.org.il, poesio@essex.ac.uk

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

1. Introduction

A great number of data-driven approaches to anaphora resolution have been proposed recently, 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 factor in this breakthrough has been the creation of larger and more theoretically motivated gold annotated corpora, such as 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 simplicity of the OntoNotes bracket-style annotation scheme and (ii) the intrinsic difficulty of the task once we aim beyond "easy victories". We believe therefore that the time is ripe for a dataset with more complex and linguistically annotation of anaphora and related discourse phenomena.

This paper presents ARRAU[1]—a multi-genre corpus of English, providing large-scale annotations of various linguistic phenomena related to anaphora. Several im-

[1] http://cswww.essex.ac.uk/Research/nle/arrau/

portant features distinguish ARRAU from similar projects. First, it supports a more complex and linguistically motivated annotation scheme for anaphora, covering 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 [10] for the `rst` domain). This enables a more thorough analysis of these phenomena, as well as creates training material for algorithms that model these tasks jointly.

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

Third, anaphoric ambiguity is annotated. Ambiguous anaphoric expressions constitute truly challenging examples that cannot be resolved with current methods for identity coreference. Moreover, the most commonly used corpora [11, 7] represent coreference chains as sets (partitions) and thus cannot support anaphoric ambiguity. By annotating ambiguous anaphoric expressions, we make the first step toward a thorough investigation of anaphoric ambiguity.

Fourth, the corpus covers, in addition to news, a variety of genres so far poorly studied, such as dialogue (TRAINS) and fiction (Pear Stories). Spontaneous dialogue and fiction are not covered by most commonly used coreference corpora.[2] We believe that anaphora, among many other discourse-related phenomena, can bring a lot of challenging genre-specific problems.

Finally, the ARRAU dataset has been under development for over ten years, during which time we have had the opportunity not only to extend the annotation and the size of the corpus, but also and crucially to improve annotation quality. In this paper we describe the second major release of the corpus, dedicated not only to increasing the corpus size, but also to implementing procedures for improving the annotation quality and consistency. 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 two versions of the ARRAU corpus have first been presented at the Language Resources and Evaluation conference Poesio and Artstein [12], Uryupina

[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.

et al. [13]. This article builds upon the two LREC papers, providing an extensive overview of the annotation guidelines and focusing more specifically on linguistically motivated 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. ARRAU Annotation Guidelines

The goal of the ARRAU project was to provide large-scale annotations of different types of linguistic information of importance in the study of anaphora. To this end, we designed annotation guidelines aiming specifically at more challenging cases of anaphora. The annotation guidelines provide extensive definitions of mention boundaries, offer a wide range of manually labeled mention attributes and finally supports complex annotations of the relevant discourse phenomena. In addition, separate guidelines were developed for each of the different genres. In this Section, we summarize our guidelines focusing on the most distinctive features of the ARRAU annotation.

2.1. Annotation tool and markup scheme

ARRAU was annotated using the MMAX2 annotation tool [14]. MMAX2 is based on token standoff technology: the annotated anaphoric information is stored in a `phrase` level whose markables point to a base layer in which each token is represented by a separate XML element. Because of the need to encode ambiguity and bridging references, anaphoric information is encoded using MMAX2 **pointers** instead of set-based attributes.

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

2.2. Document Selection

Most anaphoric corpora consist of news or broadcast documents. To align with these resources, the corpus includes a news domain called `RST` and consisting of the entire subset of the Penn Treebank that was annotated in the RST treebank [15]).

[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.

Note that for these documents, the rhetorical structure annotation is available from [15], allowing for extensive analysis of discourse features in anaphora resolution and vice versa.[4]

Apart from news, ARRAU includes three more domains, covering genres important from the point of view of discourse analysis but not normally covered by anaphoric corpora. Thus, ARRAU includes all the task-oriented dialogues in the TRAINS-93 corpus,[5] the complete collection of spoken narratives in the Pear Stories that provided some of the early evidence on salience and anaphoric reference [17], and the medical/museum documents from the GNOME collection [18, 19] used to study both local and global salience [20, 21].

Table 1 shows some basic statistics for the four ARRAU domains.[6] Both the RST and GNOME subsets contain 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. 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 subsets, on the contrary, represent spontaneous speech. The texts contain short sentences, 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 below) are very rare in both PEAR and TRAINS. For these domains, mention detection might better be implemented through a chunker robust to noisy ungrammatical input. For anaphora resolution, salience features and context modeling become the 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.

2.3. Annotated Mentions

In ARRAU, all NPs are marked as mentions. In addition, possessive pronouns are marked as well, and all premodifiers are marked when the entity referred to

[4] 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 [16].

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

[6] All the corpus 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 different genres of ARRAU.

is mentioned again, e.g., in the case of the proper name *US* in (1a), and when the premodifier refers to a kind, like *exchange-rate* in (1b). Singletons are also marked as mentions that are part of coreference chains.

(1) ... The Treasury Department said that the [US]$_i$ trade deficit may worsen next year after two years of significant improvement. ... The statement was the [US]$_i$'s government first acknowledgment of what other groups, such as the International Monetary Fund, have been predicting for months.
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]$_i$ policies. "We believe there have continued to be indications of [exchange-rate]$_i$ manipulation .....

In contrast with other anaphoric corpora, we mark *all* nominal mentions, including singletons and non-referring NPs. Moreover, we manually subclassify non-referring mentions (see below).

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:

(2) [[$_{min}$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: spans are encoded as arbitrary sequences of tokens. This allows for correct labeling of discontinuous noun phrases:

(3) ..after owners [$_{part1}$Anna]$_{part1}$ and Morris [$_{part2}$Snezak]$_{part2}$..

(4) 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 by the anaphora resolution software: state-of-the-art mention detection system always output continuous chunks; publicly available SemEval and CoNLL coreference scorers [22] assume numbered brackets as mention boundaries that cannot encode discontinuous fragments. To make our data usable by these resources, we have decided to disallow discontinuous minimal spans. This way, all the ARRAU mentions can be aligned to contiguous sequences of tokens.

2.4. Mention properties

All mentions are manually annotated for a variety of properties including morphosyntactic agreement (gender, number and person), grammatical function (following the GNOME scheme [23]), and semantic type of the entity: `person`, `animate`, `concrete`, `organization`, `space`, `time`, `numerical`, `plan` (for actions), or `abstract`). Discourse-level properties to be marked include `reference` (non-referring, discourse-new, and discourse-old) and `genericity`, annotated following the GNOME guidelines [23]. The reliability of the coding schemes for all these properties was tested as part of the GNOME annotation and is discussed in [18]. The annotation of referentiality and genericity are discussed in detail below.

Referentiality. Most anaphoric corpora focus on *referring* mentions—expressions that denote specific entities and can participate in anaphoric relations. It makes it difficult to analyze and model *non-referring* mentions—nominal expressions that do not denote any specific 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. For example, [24] train a pre-filtering classifier for non-anaphoric *it*, *you* and *we* on the OntoNotes data.

In ARRAU, we take a more principled approach, annotating all the mentions, including non-referentials. Moreover, we further sub-classify non-referential mentions into expletives (5), predicatives (6,7), idiomatic (8), incomplete or fragmentary expressions (9), quantifiers (10) and coordinations (11).

(5) And [there]'s a ladder coming out o of the tree
and [there]'s a man at the top of the ladder

(6) It see it seems to be [a busy place]

(7) 1 ml of the prepared solution for injection contains 0.25 mg ([8 million IU]) of Interferon beta-1b.

(8) so that would um if we left at six in the morning would that make [sense] six (mumble)

	RST	TRAINS	GNOME	PEAR
all	72013	16999	6562	4008
non-referential	9552	2353	1047	607
coordination	2410	232	327	37
expletive	444	851	75	122
idiom	638	148	29	42
incomplete	2	149	1	36
predicate	4311	145	355	79
quantifier	1738	818	259	132
unknown	9	6	1	159

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

(9) U: okay then um okay then originally we need to have um the one boxcar go to [oran- um] go to Corning from Elmira

(10) [Half of those polled] see the currency trending lower over the next three months, while the others forecast a modest rebound after the New Year.

(11) Mr. Sutton recalls: " When I left, I sat down with [Charlie Rangel, Basil Paterson and David], and David said, 'Who will run for borough president?

The treatment of coordination in ARRAU might sound controversial: we mark individual noun phrases ("Charlie Rangel", "Basil Paterson" and "David" in (11) above) as referring mentions that can participate in anaphoric relations and the embedding coordinate NP as non-referring. This allows for a more principled annotation of plural coreference (cf. below).

Table 2 shows the distribution of various types of non-referential mentions in the whole corpus and in the four individual domains. As expected, the distribution of non-referentials 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 int RST and GNOME documents. Idioms are common for all the genres except GNOME—a collection of medical leaflets written in a very formal language. Predicative non-referentials, especially appositions, are more common for news.

Genericity. The guidelines for genericity developed for GNOME were designed to distinguish generic uses of nominal expressions proper (as in *Dogs bark*) from non-generic cases (as in *I saw dogs in the street*) and cases in which the nominal expression is simply bound by a quantifier, conditional or other semantic operator

(as in *When dogs see a cat, the dogs bark*). In order to achieve reliability, the annotation of the genericity attribute of a nominal 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 12) or an individual quantifier (iquant) (as in 13) or a temporal quantifier (as in 14) or a modal (as in 15). 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 17. Finally, the annotator is asked to mark the nominal as `generic-yes` if it refers generically, or `generic-no` if it is non-generic.

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

(13) 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] . [operator-iquant]

(14) In addition , once [money] is raised , [investors] usually have no way of knowing how [it] is spent . [operator-tquant]

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

(16) Use [alcohol wipes] to clean the tops of the vials move in one direction and use one wipe per vial. [operator-instruction]

(17) Not that [oil] suddenly is a sure thing again . [underspecified-substance, RST]

(18) 1 ml of [the prepared solution for injection] contains 0.25 mg (8 million IU) of [Interferon beta-1b] . [underspecified-substance, GNOME]

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

2.5. Range of relations

The ARRAU guidelines support annotation of different types of anaphora: (identity) coreference, discourse deixis and bridging. These relations are marked on the same set of documents, allowing for deeper analysis and joint modeling of the three phenomena.

All referring mentions are marked as either `discourse new` or `discourse old`. Discourse new mentions introduce new entities and thus are not annotated for

	RST	TRAINS	GNOME	PEAR
all	72013	16999	6562	4008
generic	2793	3077	856	122
generic-yes	1438	728	12	74
episodic-no	-	4	-	-
operator-conditional	90	231	201	2
operator-instruction	15	163	211	-
operator-iquant	7	6	-	-
operator-modal	443	1080	147	16
operator-question	54	432	39	10
operator-tquant	16	4	-	-
underspecified-decease	-	-	84	-
underspecified-generic	1	3	-	-
underspecified-replicable	37	1	2	21
underspecified-substance	692	431	160	-

Table 3: Distribution of generic markables in ARRAU.

antecedents. For discourse-old mentions, an antecedent can be identified, either of type `phrase` (an already labeled preceding mention) or `segment` (not a nominal mention, in cases of discourse deixis).

Anaphoric ambiguity. 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` markables can be marked as ambiguous between two distinct interpretations.

(20) Criticism of [the Abbie Hoffman segment]$_1$ is particularly scathing among people who knew and loved the man. $<\ldots>$ 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] is dangerous and misrepresents Abbie's life," says Ms. Lawrenson, who has had an advance look at the 36-page script .

In (20), the anaphoric mention "This" is ambiguous between "the Abbie Hoffman segment" (identity coreference) and "CBS News is apparently concentrating on Mr. Hoffman's problems as a manic-depressive" (discourse deixis). To our knowledge, no other anaphoric corpora provides annotation for ambiguity. Moreover, no state-of-the-art anaphora resolution algorithms aim at modeling such cases.

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 4: Corpus statistics for two releases of ARRAU

Bridging anaphora. 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 anaphors. The ARRAU guidelines for bridging anaphora are a much simplified version of the guidelines developed for GNOME, that concentrated on a subset of the range of bridging relations, covering the cases in which the underlying relation is part-of anaphora or one of a range of set-based relations such as subset-of and element-of.

Plural anaphors. Anaphoric mentions referring to sets of objects are intrinsically difficult both for annotation and resolution. Consider the following toy example:

(21) a. Mary and John went to the bar. [They] had a great time.
b. Mary met John at the bar. [They] had a great time.

Most annotation schemes introduce an entity for "Mary and John" in (21a) and link "They" to this entity. For (21b), however, the same annotation trick is not applicable, since there is no longer a constituent for "Mary and John" —and "They" becomes a discourse new mention with no antecedent. We believe, however, that these two very similar cases should be treated along the same lines. In ARRAU, we annotate plural anaphors as pointers to each member of the corresponding set encoding an (element-of) bridging relation. Thus, in (21a) as well as in (21b) "They" is linked to both "Mary" and "John" individually. Note that such annotation allows for a more straightforward interpretation of plurals.

Till recently, no data-driven studies were attempting to model plural anaphora specifically. A very recent work [25] aimed at rule-based plural anaphora resolution for the patent domain. We believe that a dataset annotated for plural anaphora in a principled way will open several challenging research possibilities.

3. Two Versions of ARRAU

The first release of ARRAU [12] 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 the TRAINS domain, since it provides a unique large collection of dialogues annotated with anaphoric information. More statistics for both releases of ARRAU are available in Table 4.

Most importantly, between the two releases we have designed a methodology for enforcing the annotation consistency. 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 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 MMAX functionality prevents a non-referential mention from being selected as an antecedent.

	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 5: Comparison across coreferentially annotated corpora

- 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. On the other hand, we have identified a number of truly challenging cases of coreference. The linguistic analysis of such examples constitutes a part of our ongoing work. Note that producing a non-negligible amount of challenging example has only been made possible as a byproduct of our thorough linguistically motivated annotation, for example, through a clash between coreference and non-referentiality.

4. Related Work: ARRAU vs. Othe Anaphoric Corpora

A number of coreferentially annotated corpora have been created and released in the past two decades, bringing forward data-driven research on anaphora. An extensive overview of such resources can be found in [26]. In this section, we highlight the main differences between ARRAU and two other commonly used corpora annotated for coreference in English, ACE [11] and OntoNotes [7, 9, 6]. Table 5 provides a summary of the most distinctive features of ARRAUas 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 coreference, 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.

Each nominal mention is shown with its minimal and maximal span. This solution is in line with the ACE annotation guidelines and has unfortunately been discarded for the OntoNotes dataset in order to decrease the annotation price and thus augment the corpus size. The maximal span corresponds to the full noun phrase, whereas the minimal span corresponds to the head noun or to the bare named entity for complex NE-nominals. With the latest development in 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 [27, 28]. 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 [29]. 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.

To stay aligned 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 OntoNotes/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-referentials. The ACE guidelines restrict the annotation scope to referentials[7],

[7]Moreover, the ACE guidelines focus on specific semantic types of referential mentions, moti-

whereas OntoNotes only marks co-referential (no singletons) mentions. Our corpus statistics show that non-referentials and singletons 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 [30, 31] 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 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 a higher performance if such tasks are attempted in a data-driven way.

The text collections used in ARRAUhave 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.[8] 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, ARRAUcan be valuable, on one hand, for deeper linguistically oriented analysis of complex cases and, on the other hand, for learning models for related phenomena (genericity, referentiality etc).

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.

ARRAU supports rich annotation of individual mentions: apart from morphosyntactic properties, we mark semantic type, genericity and referentiality. For the latter

vated from the Information Extraction perspective: person, organization, location and so on.

[8] We do not include RST annotations in the ARRAU distributions. The relevant information can be extracted through straightforward corpora alignment.

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 [32]. 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. The work was in part supported by the EPSRC-funded ARRAU Project (GR/S76434/01), in part by the LiveMemories project, funded by the Provincia of Trento, in part by the EU Project H2020 5G-CogNet.

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 ACL, 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, doi:\bibinfo{doi}{10.1162/COLI_a_00200}, URL http://dx.doi.org/10.1162/COLI_a_00200.

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

[11] 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.

[12] M. Poesio, R. Artstein, Anaphoric annotation in the ARRAU corpus, in: Proceedings of the Language Resources and Evaluation Conference, 2008.

[13] 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.

[14] C. Müller, M. Strube, Multi-level annotation of linguistic data with MMAX2, in: Corpus Technology and Language Pedagogy, S. Braun and K. Kohn and J. Mukherjee, 197–214, 2006.

[15] 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.

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

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

[18] M. Poesio, Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results, in: Proc. of the 2nd LREC, Athens, 211–218, 2000.

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

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

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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, Springer, 2016.

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

[28] H. Peng, K.-W. Chang, D. Roth, A Joint Framework for Coreference Resolution and Mention Head Detection, CoNLL 2015 51 (2015) 12.

[29] J. K. Kummerfeld, M. Bansal, D. Burkett, D. Klein, Mention Detection: Heuristics for the OntoNotes annotations, in: Conference on Natural Language Learning, Association for Computational Linguistics, 102–106, 2011.

[30] 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, URL http://www.aclweb.org/anthology/W11-1916, 2011.

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

[32] K.-J. Rodriguez, F. Delogu, Y. Versley, E. Stemle, M. Poesio, Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus, in: Proc. LREC (poster), 2010.

6. Reviews

6.1. Reviewer: 1

This is an interesting piece of work, important for the area of coreference resolution. I especially like the idea of including texts belonging to several genres, as there were already studies showing that coreference-related phenomena may vary across genres (and registers). However, there are some issues in the paper that should be clarified. I also include reference to some studies that are relevant for the discussed issues.

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.

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 & Gmez-Gonzlez (2012).

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).

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.

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?

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?

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] ?)

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.

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:?.

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.

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

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.

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? 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.

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).

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.

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)?

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??

Section 4

1) Related Work: ARRAU vs. Othe Anaphoric Corpora ? ? Other... 2) page 15: in the footnote section, something went wrong in the Latex code, as there is a cut sentence.

REFERENCE

Amoia, M., K. Kunz and E. Lapshinova-Koltunski (2011). 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). 15th September, Hissar, Bulgaria, pp. 2-10.

Friedrich, A.-M. and M Pinkal (2012). Discourse-sensitive Automatic Identification of Generic Expressions. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL). Beijing, China.

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

Halliday, Michael A. K., and Ruqaiya Hasan. 1976. Cohesion in English. London, New York: Longman.

Kunz, K. and E. Lapshinova-Koltunski (2015). Cross-linguistic analysis of discourse variation across registers. In K. Ajmer and H. Hasselgrd (eds). Cross-linguistic Studies at the Interface between Lexis and Grammar. Nordic Journal of English Studies. Vol 14 (1). pp. 258-288.

Kunz, K., E. Lapshinova-Koltunski and J.M. Martinez-Martinez (2016). Beyond Identity Coreference: Contrasting Indicators of Textual Coherence in English and German. In Proceedings of CORBON at NAACL-HLT2016, San Diego, pp. 23-31.

Lapshinova-Koltunski, E. and K. Kunz (2014). Annotating Cohesion for Multillingual Analysis. In Opens external link in new windowProceedings of the 10th

Joint ACL - ISO Workshop on Interoperable Semantic Annotation, Reykjavik, May 26, 2014.

Martnez Martnez, J. M.., Lapshinova-Koltunski, E. and K. A. Kunz (2016). Annotation of Lexical Cohesion in English and German: Automatic and Manual Procedures. In Proceedings of the Conference on Natural Language Processing (Konferenz zur Verarbeitung natrlicher Sprache) - KONVENS-2016, September 19-21, Bochum, Germany, pp. 165-176.

Nedoluzhko, A. (2013). Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW). SIGANN, Sofia, Bulgaria, Association for Computational Linguistics, pp. 103?111.

Neumann, Stella (2013). Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German. Berlin, Boston: De Gruyter Mouton.

Taboada, M., Gmez-Gonzlez, M. (2012). Discourse markers and coherence relations: Comparison across markers, languages and modalities. Linguistics and the Human Sciences, 6 (1-3), pp. 17-41.

Ziknov rka, Haji?ov Eva, Hladk Barbora, Jnov Pavlna, Mrovsk Ji?, Nedoluzhko Anna, Polkov Lucie, Rysov Kate?ina, Rysov Magdalna, Vcl Jan: Discourse and Coherence. From the Sentence Structure to Relations in Text. FAL, Praha, Czech Republic.

6.2. Reviewer: 2

The paper is based on two previously published LREC conference papers (Poesio & Artstein, 2008; Uryupina et al, 2016). While I completely support the practice of synthesizing a set of previously published conference papers into a well-organized, non-redundant journal paper that can be a reader's "port of call" into the work, it can also be a source of weaknesses that have to be addressed:

- The result can contain redundancies, as well as "near redundancies" that actually present conflicting assertions;

- Issues that aren't addressed in the conference papers – perhaps for lack of room – become glaring omissions when space isn't at a premium.

The current paper suffers from both these weaknesses, as well as being confusing in other, more minor, ways. I view all these problems as being addressable with MINOR REVISIONS, especially given that there are SEVEN authors, there's not that much for each to do.

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).

In a paper for a journal, all this needs to be brought together and made sense of in a consistent way. The paper should not be accepted until it reads like a single coherent text.

Among issues that are raised but not addressed in the paper:

- 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.

- 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.

- 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.

This is just really sloppy, and the paper shouldn't be accepted until such claims are either addressed or omitted.

In addition to the significant problems noted above, I would ask for the following to be addressed:

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.

Section 2.3 (p.6)

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

- 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?

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?

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.

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.

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.

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.)

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.

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.

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.

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.

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.

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.

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'.

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 sectin 4. 'on the one hand'