Chapter 4
Annotated Corpora and Annotation Tools

Massimo Poesio, Sameer Pradhan, Marta Recasens, Kepa Rodriguez and Yannick Versley

Abstract In this Chapter we review the currently available corpora to study anaphoric interpretation, and the tools that can be used to create new ones.

4.1 Introduction

In the 1990s, the desire to use anaphora resolution in practical applications, especially in the then-nascent field of information extraction, led to a shift in focus in anaphora resolution research towards a more empirical approach to the problem. This more empirical focus also led to the creation of the first medium-size annotated corpora, which allowed for data-driven development of resolution procedures and machine learning approaches.

These changes were primarily brought about by the Message Understanding Conferences (MUC), a DARPA-funded initiative where researchers would compare the quality of their information extraction systems on an annotated corpus provided by funding agencies. MUC introduced the **coreference resolution task** already discussed in Chapters 2 and 3, and hosted two evaluations of coreference resolution systems, MUC-6 [21] and MUC-7 [13], where annotated corpora were provided to

Massimo Poesio
University of Essex, Wivenhoe, United Kingdom, e-mail: poesio@essex.ac.uk

Sameer Pradhan
Boulder Language Technologies

Marta Recasens
Google Inc.

Kepa-Joseba Rodriguez
Yad Vashem Archives

Yannick Versley
University of Heidelberg

the participants. In parallel with the development of the corpora, guidelines for the annotation of coreference were created and a common evaluation procedure for the comparative evaluation was developed. The availability of these corpora, and of common evaluation metrics, made it possible to train and test coreference resolution systems on the same datasets, and therefore to compare their results. These efforts had a tremendous influence on the field and their influence can be seen in subsequent evaluation campaigns such as the Automatic Content Extraction (ACE) initiative.[1] As a result, it is not an exaggeration to talk of a pre-MUC and post-MUC period in research on coreference and more in general on anaphora resolution.

In this Chapter we present a detailed survey of some of the proposals concerning the annotation of corpora with anaphoric and coreference information and their use for evaluation of data-driven approaches to anaphora resolution.

4.2 Annotating Anaphora: An Overview of the Options

In a data-driven perspective, the design of the annotation scheme acquires a crucial importance. This is because linguistic data annotated with anaphoric information are used both to evaluate the performance of data-driven anaphoric resolvers (cf. Chapter 5), and to train supervised systems, the most popular machine-learning approach to this problem (cf. Chapters 9, 11 and 9.5). So the annotation scheme defines what the problem of anaphora resolution is, and what is the linguistic phenomenon to be learned from the data. We begin the chapter by briefly discussing some of the decisions to be made while designing an annotation scheme, the choices made in some of the best known schemes including both initiative-oriented schemes for English such as MUC and ACE, and more general-purpose schemes. We also mention the most controversial issues.

4.2.1 Markables

One of the controversial issues in defining a coding scheme for anaphora is the definition of **markable** or **mention**—the unit of text to be chosen as mention of an entity. This definition depends on both syntactic and semantic factors.

Syntactic characterization of markables

As discussed in Chapter 2, most current work on anaphora focuses on NP anaphora, i.e., anaphoric relations expressed with noun phrases. As a result, most coding schemes for anaphoric and coreference corpora ask coders to only consider noun

[1] `http://www.nist.gov/speech/tests/ace/index.html`

phrases as markables, with a few exceptions discussed below. In fact, some of the early coding schemes focused on a subset of all NPs: e.g., only pronouns (as in the corpora created by [18, 8, 24] or in the early versions of the Prague Dependendency Treebank [38]) or only definite descriptions (as in the Vieira-Poesio corpus, [59]). Most modern schemes, however, require coders to mark all anaphoric expressions realized with noun phrases, the main restrictions being semantic (see below).

A second type of syntactic restriction concerns the boundaries of markables. Most coding schemes, including those for MUC, ACE, MATE, GNOME, ARRAU, LIVEMEMORIES and ONTONOTES, require coders to mark the entire noun phrase[2] with all postmodifiers (4.1a). The alternative is to mark noun phrases just up to the head and leave postmodifiers out of the markable, as in (4.1b).

(4.1) a. It is more important to preserve high inter-annotator agreement than to capture [every possible phenomenon that could fall under the heading of ”coreference”].

b. It is more important to preserve high inter-annotator agreement than to capture [every possible phenomenon] that could fall under the heading of ”coreference”.

However, this tendency to mark the noun phrase in its entirety raises markable identification problems for systems: because of pre-processing errors such as parsing inaccuracies, the phrases annotated in the gold standard and those automatically identified by a system can be partially misaligned, e.g., they may differ on which postmodifiers of a noun are included in the markable. In order not to penalize anaphora resolution systems on the incorrect identification of the markable boundaries, the decision was taken in MUC to instruct coders to mark the maximal span of a noun phrase, and, in addition, to identify its head in a separate attribute called MIN. In this way, systems in MUC could also be evaluated in a relaxed evaluation setting where they received credit for markable identification based only on the matching of heads and minimal spans – the rationale being that the full set of modifiers can be optionally recovered later with the help of separate syntactic information. In ACE, the head and the minimal extent required to guarantee correct identification were marked separately, in the HEAD and EXTENT attributes, respectively. In subsequent proposals, annotators have also been generally required to annotate the NP with all its modifiers [51, 53, 63], but heads / minimal spans are not always annotated (e.g., ONTONOTES), and in some annotation projects only parts of the NP are annotated.

Most schemes include some exceptions to the rule of annotating only NPs. One type of constituent treated as markable in many schemes are noun premodifiers. In linguistics, it is generally thought that such modifiers do not add discourse referents to the discourse model, i.e., are **anaphoric islands** [60], on the grounds of contrasts such as that between (4.2a), which is generally considered acceptable, and (4.2b), which is generally considered ungrammatical.

[2] As discussed in Chapter 2, many types of expressions in language are anaphoric to a degree, but the type of anaphoric reference most studied in computational linguistics, by far, is anaphoric reference via noun phrases, so in this Chapter, as in the rest of the book, we will focus on coding schemes and corpora for NP anaphoric reference.

(4.2) a. Hunters of [animals]$_i$ tend to like [them]$_i$.
b. *[Animal]$_i$ hunters tend to like [them]$_i$.

However, [76] proposed a rather different account of these data, pointing out first of all that such positions not only do not block anaphoric reference in general–see (4.3a)–but that also nominal modification is possible at least in certain cases, as shown in (4.3b). They proposed that whereas in a subset of these examples anaphoric reference is indeed blocked, in general the possibility to refer depends on pragmatic factors.

(4.3) a. Millions of [Oprah Winfrey]$_i$ fans were thoroughly confused last week when, during [her]$_i$ show, [she]$_i$ emotionally denied and denounced a vile rumour about [herself]$_i$.
b. I had a [paper]$_j$ route once but my boss told me I took too long to deliver [them]$_j$.

Many, if not most, coding schemes for anaphoric reference require coders to annotate at least some cases of reference to antecedents introduced by prenominal modifiers. For instance, the MUC guidelines state that prenominal modifiers are markables only if the coreference chain contains one element that is not a modifier. Thus, *drug* is a markable in (4.4a), but *contract drilling* is not in (4.4b).

(4.4) a. He was accused of money laundering and [drug]$_i$ trafficking. However, the trade in [drugs]$_i$...
b. Ocean Drilling & Exploration Co. will sell its [contract drilling] business. ... Ocean Drilling said it will offer 15% to 20% of the [contract drilling] business through an initial public offering in the near future.

Similar instructions are found in the ARRAU and GNOME guidelines, where coders are also required to annotate *drug* and *drugs* in (4.4a) as generic. It should be noted, however, that this 'on-demand' annotation makes mention detection difficult for systems as they cannot simply rely on syntactic structure, and not many systems are good at identifying generic cases.

Another class of markables not associated with (realized) NPs are **incorporated anaphors** in Romance languages (see Chapter 2). As a reminder, incorporated anaphors are cases of anaphoric reference in which the anaphoric expression is expressed by an affix to another expression, e.g., a verb, as in the following example from Italian, where clitic suffix *lo* refers back to Giovanni.

(4.5) a. [IT] Giovanni$_i$ e' in ritardo così mi ha chiesto se posso incontrar[lo]$_i$ al cinema.
b. [EN] John$_i$ is late so he$_i$ asked me if I can meet him$_i$ at the movies.

A second class of anaphors that may cause problems from the point of view of markable identification are **zero anaphors**–cases of anaphoric reference in which one argument is unrealized, as in the following examples from Italian and Japanese.

(4.6) a. [IT] [Giovanni]$_i$ andò a far visita a degli amici. Per via, ϕ_i comprò del vino.
b. [JA] [John]$_i$-wa yujin-o houmon-sita. Tochu-de ϕ_i wain-o ka-tta.
c. [EN] [John]$_i$ went to visit some friends. On the way, [he]$_i$ bought some wine.

Such markables can be a problem for markup-based annotation (i.e., annotation in which markables are chunks of text), depending on the limitations of the annotation tool (see Section 4.4.1). They are not a problem when anaphoric annotation piggybacks on a syntactically and morphologically annotated layer which serves as a base layer, as in the case of ANCORA [66], the Prague Dependency Treebank [23], or ONTONOTES [77, 63]. This ideal situation is however rather uncommon among existing annotated corpora. Even when the base layer is text, as it is often the case, these expressions are not particularly problematic when standoff is based on character offset, as done in the NAIST corpus of anaphora in Japanese [31], annotated using Tagrin,[3] or in annotations using CALLISTO.[4] This is because with standoff, markables can point to a subset of the verbal expression (i.e., *-lo* in (4.5a)) or to a zero-length string before the markable (4.5b). However, with token standoff, some convention has to be introduced to associate those anaphors with other markables. A common approach is to mark the nearest verbal constituent, as proposed in the MATE guidelines and done in the Italian LIVEMEMORIES corpus [67]. In (4.7), the verbal form *dargli*, which includes the incorporated clitic *-gli* referring to Giovanni, would be treated as a markable of type `verbal`, and it would be annotated as anaphoric to Giovanni.

(4.7) [Giovanni]$_i$ è un seccatore. Non [dargli]$_i$ retta.
[John]$_i$ is a nuisance. Do not pay any attention to [him]$_i$.

The last syntactic (but also semantic) restriction on markables that we will discuss are cases of anaphoric reference in which the antecedent is not introduced by an NP, as in cases of so-called **event reference** and **discourse deixis**, discussed in Chapter 2. In the example of event anaphora in (4.8), the pronoun *it* refers to the event of John breaking his leg, not introduced by a nominal; in the example of discourse deixis in (4.9), the demonstrative pronoun *that* in B's statement refers to the proposition asserted by A in her previous utterance. These types of anaphora were not annotated in the MUC or ACE corpora (see, e.g., [28]), or in most existing corpora, but event anaphora is annotated in ONTONOTES, and discourse deixis in the ARRAU corpus.

(4.8) John broke his leg yesterday.
It happened while he was skiing.
(4.9) A: John broke his leg yesterday.
B: That's not true - I saw him this morning and he seemed fine to me.

A particularly intricate issue with defining markables is what to do with coordination, which we discuss in Section 4.2.3.

[3] `http://kagonma.org/tagrin/`

[4] `http://mitre.github.io/callisto/index.html`

Semantic restrictions on markables

From a semantic perspective, a coding scheme may either require coders to annotate mentions of all types of entities, or of a subset of them only. In the context of information extraction applications, coreference resolution is most important for members of a small number of **semantic classes** that are relevant for the domain at hand. Many early machine-learning approaches such as [41] and [2], only concerned themselves with organizations and persons. As a result, the guidelines for the ACE coreference annotation, for instance, identified seven types of entities as most relevant (PERSON, ORGANIZATION, GEO/SOCIAL/POLITICAL ENTITY, LOCATION, FACILITY, VEHICLE, WEAPON) and only asked annotators to annotate mentions of those types [39].

One benefit of narrowly focusing on a small number of (presumably) well-behaved semantic classes is that identity or non-identity is usually straightforward to determine, whereas it may be very difficult to decide for abstract or vague objects. The disadvantage is that anaphoric resolvers trained on these data will not be very useful in different domains. For instance, artifacts other than vehicles and weapons are not annotated in the ACE corpora, but these turn out to be a key entity type in one of the GNOME [51] domains, namely museum objects.

Coding schemes may also choose to only mark NPs fulfilling certain semantic functions. As discussed in Chapter 2, nominal expressions can play at least four types of semantic function: **referring**, **quantificational**, **predicative**, or **expletive**. In many coding schemes, coders are instructed not to mark expletives (e.g., MUC [28]). In such schemes, predicative NPs are generally markables, but they are marked as coreferent with the referring NPs they are predicated about–i.e., referring and predicative mentions are treated as having the same function. More recent schemes generally make the distinction between coreference and predication. In some schemes (e.g., ANCORA, ONTONOTES), a different relation is used for marking attributive cases (e.g., appositive NPs are annotated as ATTRIBUTE of the encompassing NP). In other schemes (e.g., ARRAU), no relation is marked between the predicative NP and the referring NP of which it specifies a property. In some of these schemes (including ACE, GNOME, and ARRAU), special attributes are used to mark the semantic function of the markable. In ACE, the CLASS attribute was used to specify whether a markable is referential or attributive, and in the case of referential markables, whether it is generic or not [1]. In GNOME, the LF_TYPE attribute was used to mark the logical form interpretation of the markable: term, predicate, quantifier, or coordination, whereas the reference attribute specified terms as being directly referring, bound, or non-referring [51]. In ARRAU, these two attributes are merged in a single reference attribute.

4.2.2 Anaphoric relations

In the MUC coding scheme, annotators were asked to mark only the anaphoric relations involving entities introduced by NPs and mentioned using NPs or nominal modifiers, but none of the other anaphoric relations discussed in Chapter 2: associative relations, cases of identity of sense, and relations where the anaphor or the antecedent are not both explicitly introduced as part of a noun phrase. The reason was the difficulty in annotating such relations already discussed in Chapter 2. Annotation efforts that include associative anaphora are DRAMA [48], the UCREL scheme developed at the University of Lancaster [6], and a number of schemes implementing the MATE guidelines, in particular the GNOME annotation [51]. Discourse deixis was annotated in ARRAU [53].

As discussed in Chapter 2 and again in Section 4.2.1, NPs can perform different semantic functions but not all coding schemes distinguish between such functions. A famously controversial aspect of the definition of the coreference task in MUC was the proposal to annotate as coreferent appositive and copula constructions, which would normally be considered cases of predication. This drew criticism from researchers such as van Deemter and Kibble [15], since the inclusion of intensional descriptions leads to counter-intuitive effects in cases such as the following one:

(4.10) [Henry Higgins], who was formerly [sales director of Sudsy Soaps], became [president of Dreamy Detergents].

In this example, following the guidelines would lead to "*sales director of Sudsy Soaps*" and "*president of Dreamy Detergents*" being annotated as coreferent. This conflation of anaphoricity and predication has been abandoned in more recent coding schemes, following the guidelines proposed by the Discourse Resource Initiative [48] and the MATE project [51]. The coding schemes developed for the GNOME and ARRAU corpora [53] and for the corpora used in the 2010 SEMEVAL competition (ANCORA [66], COREA [25], TüBa-D/Z [27], LiveMemories [67], ONTONOTES [77, 63]), and for the CONLL-2011 and CONLL-2012 shared tasks (ONTONOTES), all distinguish between (transitive) coreference and (directed, non-transitive) predication. In some of these corpora (e.g., ARRAU), predication is simply not marked, whereas in other corpora (e.g., GNOME and ONTONOTES) it is marked as a different type of link.

A particularly difficult issue is **metonymy**, as in the following example.

(4.11) *Paris* rejected the "logic of ultimatums".

In this example, the NP *Paris* is not used to refer to a geographical entity (the city of Paris) but to a (political) entity linked to Paris by a systematic relation. This example could be interpreted roughly as meaning:

> A French government official made a statement to the effect that the official French position regarding the "logic of ultimatums" is of disapproval.

Such examples raise two types of issues. Semantically, the coder must decide what type of entity should be assigned to the markable. From the point of view of anaphoric annotation, the guidelines should specify whether the markable *Paris* in

(4.11) has to be annotated as coreferent to other mentions of any of the following entities:

1. the city of Paris;
2. the country of France (as a geographic entity);
3. the French government ;
4. the government official uttering the sentence

Different (partial) solutions have been adopted for this problem. The ACE guidelines resolve the ambiguity between 2 and 3 by assuming a semantic class of so-called **geopolitical entities** (GPEs), i.e., a conflation of a country, its government, and its inhabitants. In ONTONOTES, the diametrically opposite solution was chosen: metonymies are distinguished from other uses of an NP, e.g., coreferential ones. Thus, in a document that contains the sentences:

(4.12) $[_1$ South Korea] is a country in southeastern Asia. ... $[_2$ South Korea] has signed the agreement.

the annotation guidelines require to distinguish between "South Korea" mentioned as a country (1) and its metonymous use referring to the South Korean government (2).

4.2.3 Coordination and Plurals

The semantics of **coordination** and **plurals** is reasonably well understood, but it is not straightforward to annotate anaphoric relations involving coordinated or plural NPs, especially in a way that current anaphora resolution models could be trained to resolve them.

Coordinations like *John and Mary* in (4.13a) are generally considered NPs, and therefore treated in most coding schemes as markables. It is therefore possible in such cases to mark plural *they* as having the conjunction as antecedent. However, plurals can also have **split antecedents**–they can refer to a plural entity consisting of two entities introduced separately, but not previously mentioned (4.13b).

(4.13)a. [[John]$_i$ and [Mary]$_j$]$_k$ went to the movies. [They]$_k$ saw *Turtle Diary*.
b. [John]$_i$ went to the movies with [Mary]$_j$. [They]$_k$ saw *Turtle Diary*.

Clearly, there are many different ways in which to annotate anaphoricity information in these cases, and therefore different solutions have been adopted in the existing coding schemes. In MUC, ACE, and ONTONOTES, the coordinated NP is marked as the antecedent of *they* in (4.13a), but no antecedent is marked for *they* in (4.13b).

GNOME and ARRAU tried to treat the two cases of plural reference in a uniform way, but different solutions were adopted. In GNOME, the antecedents of plural pronouns are always marked using the associative relation **has-element**: both in (4.13a) and (4.13b), no identity relation is marked for *they*, but both *John* and *Mary* are marked as elements of the set denoted by the plural. In ANCORA and ARRAU, the possibility offered by the ANCORAPIPE and MMAX2 annotation tools (see Section

4.4.1) to annotate split antecedents was used: in both examples, plural *they* is marked as having John and Mary as antecedents.

4.3 Corpora Annotated with Anaphoric Information

4.3.1 The MUC *Corpora*

The sixth and seventh editions of the Message Understanding Conference (MUC-6 and MUC-7) introduced two ‘Semantic Evaluation’ (SEMEVAL) tasks in addition to the template-filling tasks evaluated at previous editions of the MUC competition: coreference and named entity disambiguation [22]. To this end, new datasets were created which, in the case of coreference, were the first corpora of any size available for training and evaluating coreference resolution systems. The dataset created for MUC-6 consists of 25 articles from the Wall Street Journal on negotiations of labor disputes and corporate management succession, for a total of around 30,000 words. The MUC-7 dataset consists of a similar amount of data on airplane crashes and rocket / missile launches. Now that larger resources exist, these two corpora are not widely used anymore except for comparison with older systems, but the task definition developed for their creation is still very influential.

4.3.1.1 Markup Scheme

The MUC corpora are annotated using inline SGML. Every markable that belongs to a coreference chain is identified with a `<COREF>` tag; `<COREF>` elements have three attributes: `ID` number, `TYPE` (always filled with `IDENT`) and `REF`. The first mention of a coref chain uses the attribute `id` to assign an ID to the coreference chain, and every subsequent mention uses the attribute `REF` to specify the coreference chain to which it belongs. There is an optional attribute, `STATUS`, that always takes the value `OPT` and marks optional links, like predications.

(4.14)
```
 <COREF ID="100">Lawson Mardon Group Ltd.</COREF>
said <COREF ID="101" TYPE="IDENT" REF="100">it</COREF>
...
```

4.3.1.2 Guidelines

The annotation scheme developed for MUC [28] virtually defined the focus for research on anaphora resolution and coreference for the fifteen years after. The scheme is focused on coreference between NPs. Only cases of nominal mention of discourse entities are considered; no other type of relation (no identity of sense or bridging relation, for instance). No relations where the anaphor or the antecedent are not both explicitly introduced as part of a NP are considered either (i.e., no ellipsis, and no reference to implicitly mentioned objects as in discourse deixis).

Markable Definition

Syntactically, annotators were asked to consider as markables NPs and nouns occurring in certain positions. Pronouns include both personal pronouns (including possessive pronouns) and demonstrative pronouns. Dates, percentages and currency expressions are considered nominal phrases.

Markables are defined as the maximal projections of the noun phrase, i.e., they include all pre-and post modifiers like non-restrictive relative clauses, prepositional phrases, etc. This definition of markable, while linguistically justified, could make system evaluation overly strict given that most mention extraction systems encounter difficulty at identifying all modifiers. Thus, in order to facilitate aligning the markables in the gold standard and the markables produced by a system, the MUC coding scheme introduced the solution discussed in Section 4.2.1– each markable is annotated with a `MIN` attribute containing the head of the NP (4.15).

(4.15) `But <COREF ID="42"` **MIN=”planes”**`>military training`
`planes</COREF> make up to ...`

If the head of the markable is a multi-word named entity, like *Julius Cesar* in 4.16, the entire named entity is specified as the value of `MIN`.

(4.16) `<COREF ID="1"` **MIN=”Julius Caesar”**`>Julius Caesar,`
`<COREF ID="2" REF="1" MIN="emperor" TYPE="IDENT">`
`the/a well-known emperor, </COREF></COREF>`

All and only mentions of entities which are introduced by an NP and are mentioned more than once are considered as markables: i.e., singletons are not annotated, and more entity types are considered than those specified in the guidelines for named entity annotation.[5] However, embedded named entities are not considered as markables: for example, the two occurrences of *Iowa* in (4.17) are not marked as coreferent, since the first one is a substring of a named entity.

(4.17) [Equitable of Iowa Cos.].... located in [Iowa]

In the case of conjoined NPs, both the individual NPs and the coordinated NP are potential markables, as shown in (4.18).

(4.18) [[the two Croatians] and [Brown]]

However, in the case of coordinated NPs, there isn't an obvious notion of 'head' other than perhaps the coordination itself (*and*). This is not a noun however, making the annotation of the `MIN` attribute problematic. Different solutions to this problem were adopted in MUC6 and MUC7. The MUC6 guidelines [19] prescribe not to treat

[5] These are persons, organizations, locations, temporal expressions, and numerical expressions– see, eg., [20].

as markables coordinated NPs that can have more than one head. The MUC7 guidelines [28], by contrast, propose to assign a coordinated head to such NPs: e.g., in example (4.18), the `MIN` should be the span "Croatians and Brown" as in (4.19).

```
(4.19)  <COREF ID="59" MIN="Croatians and Brown">
      <COREF ID="56" TYPE="IDENT" REF="14" MIN="Croatians">
      The two Croatians</COREF>
      and
      <COREF ID="57" TYPE="IDENT" REF="39">Brown</COREF>
```

Notice that the span of `MIN` in this example does not correspond to any linguistic category.

Range of relations

Apart from the simple examples mentioned above, the coders were also asked to consider the following as cases of coreference:

- Bound anaphora, as in
  [Most computational linguists] prefer [their] own parsers
  or
  [Every TV network] reported [its] profits yesterday. [They] plan to release full quarterly statements tomorrow.
- More controversially (see above and Chapter 2), the coders were asked to consider many cases of predication as cases of coreference. This includes most cases of appositions, as in
  [Julius Cesar], [the well known emperor]
  This identity of reference is to be represented by a coreference link between the appositional phrase, "the well-known emperor", and the ENTIRE NP, "Julius Caesar, the/a well-known emperor" 4.20:

  ```
  (4.20)  <COREF ID="1" MIN="Julius Caesar">Julius Caesar,
        <COREF ID="2" REF="1" MIN="emperor" TYPE="IDENT">
        the/a well-known emperor,</COREF>
        </COREF>
  ```

  Other predicative nominals, such as copular constructions, are also annotated as coreferent.
  [Bill Clinton] is [the President of the United States].
- Functions and values. Coders were required to link the most recent value to the function. In (4.21), coders were required to link [$3.85] and [The stock price]. (Again, see above why this is bound to cause problems in general.)

  (4.21) [The stock price] fell from [$4.02] to [$3.85];

4.3.1.3 Availability

Both MUC corpora are available from the Linguistic Data Consortium (LDC).

4.3.2 The ACE *Corpora*

The Automatic Content Extraction program (ACE),[6] was, like MUC (of which it forms the natural continuation), an initiative of the US government to promote content extraction technology, and in particular the identification of entities, relations, and events in text [16]. The program was articulated around evaluations of systems performing these tasks; many such evaluations took place from 2000 to 2008, supporting the annotation of data in three different languages–Arabic, Chinese and English. The ACE-2 and ACE-2005 Entity Detection and Tracking (EDT) English corpora, in particular, replaced the MUC corpora as the *de facto* standards for 'coreference.'

4.3.2.1 Markup Scheme

The corpora are marked up using the **ACE Pilot Format** (APF), a standoff XML markup format in which a base file contains the text with some inline SGML annotation; information about entities and their mentions is stored in a separate file with indices which refer to character positions in the base file. Anaphoric information in APF is organized around entities: all entities annotated in the document are identified with <ENTITY> elements, and each mention of entity *e* is then recorded as a child <ENTITY_MENTION> element of the <ENTITY> element for *e*.

Each mention is annotated with the attribute TYPE, with three possible values: NAM for named entities, NOM for NPs with a common noun as head, and PRO for pronouns. Each <ENTITY_MENTION> element has two children: the <EXTENT>, which specifies the character span in the base file realizing that mention, and also contains the string of characters; and the <HEAD> element, which specifies the span of characters and contains the string of the syntactic head of the NP. The markup for mentions is shown in (4.22).

(4.22)
```
<entity_mention ID="2-5" TYPE="NOM">
<extent>
<charseq START="1621" END="1671">an assistant director
at the Oregon Zoo in Portland</charseq>
</extent>
<head>
<charseq START="1634" END="1641">director</charseq>
```

[6] http://www.itl.nist.gov/iad/mig//tests/ace/

```
</head>
</entity_mention>
```

If the head is a named entity realized by more than one word, the full named entity is the head of the markable (4.23).

(4.23)
```
<entity_mention ID="1-2" TYPE="NAM">
<extent>
<charseq START="1573" END="1609">American Zoo and
Aquarium Association</charseq>
</extent>
<head>
<charseq START="1573" END="1609">American Zoo and
Aquarium Association</charseq>
</head>
</entity_mention>
```

4.3.2.2 Guidelines

In contrast to the MUC annotation scheme, the ACE annotation scheme for entity detection and tracking focuses on a small number of semantic classes considered particularly relevant for information extraction: persons, organizations, locations, geopolitical entities, weapons, and vehicles [39]. (See discussion of semantic restrictions in Section 4.2.1.) These classes have changed over the years: the first editions focused on five classes (facilities, geopolitical entities, locations, organizations, and persons), and the later editions on seven (facilities, geopolitical entities, locations, organizations, persons, vehicles, and weapons).

The ACE guidelines follow fairly closely the MUC guidelines, but include additional specifications as they were used for Arabic and Chinese as well as English data.

Markable Definition

One of the issues addressed in the ACE annotation guidelines is the problem of metonymy (see above). In (4.24), the mention *Iraq* refers to the country as a geographical entity, whereas in a further sentence of the same text (4.25), the mention *Iraq* refers to the political and economical institutions of the country.

(4.24) Russia's opposition to the use of force in **Iraq** is the latest in a series of foreign policy disputes with the United States.

(4.25) Russia, its economy in chaos, desperately needs the cash and also hopes for big new contracts with **Iraq** when sanctions end.

The solution proposed in the ACE guidelines to ensure consistency in the annotation is the creation of a **Geopolitical Entity** (GPE) category, which merges the meaning of the country as a physical place, the institution that governs the country, and the inhabitants.

Range of relations

Like the MUC guidelines, the ACE guidelines require annotating cases of nominal predication via apposition and copular clauses as cases of coreference. For instance, in (4.26), the mention *"an Asian power"* is marked as coreferent with *"China"*.

(4.26) Today , *China* is **an Asian power** and rightfully so.

Similarly, in the ACE annotation appositions are marked as coreferent with the main NP. For instance, in (4.27) the markable *deputy prosecutor of the war crimes tribunal* corefers with the full NP.

(4.27) Graham Blewitt , deputy prosecutor of the war crimes tribunal[7]

4.3.2.3 Availability

All ACE corpora are distributed through LDC. A useful summary of the available resources is at `https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications`.

4.3.3 The DRI *and* MATE *Guidelines*

The **Discourse Resource Initiative** [48] and the MATE project [51] started a reexamination of coding schemes for anaphora, leading to the schemes adopted in most of the more recent anaphoric annotation efforts, including GNOME [51], ARRAU [53], and ONTONOTES [77, 63, 64] for English, COREA [25] for Dutch, the Potsdam Commentary Corpus [37] and TüBa-D/Z corpus [27] for German, ANCORA for Catalan and Spanish [66], and LIVEMEMORIES [67] for Italian.

These schemes tend to be more linguistically inspired and less domain-oriented than the MUC and ACE schemes. All NPs are annotated, instead of only the mentions of a selected number of entity types, and markable boundaries tend to follow NP boundaries. From a semantic perspective, all of these annotation schemes distinguish between identity and predication, and some of these schemes attempt to mark a richer range of anaphoric relations, including associative relations (e.g., GNOME, ARRAU, COREA) or some types of discourse deixis–e.g., reference to events in ONTONOTES, or reference to abstract objects in ARRAU. Many such corpora also include annotations of other properties of mentions, such as agreement features. Also, agreement studies are generally carried out. The most recent evaluation campaigns for anaphora have used corpora of this type.

[7] npaper 9801.139

In this Section we discuss the MATE guidelines and the GNOME corpus; we will then discuss ARRAU and LIVEMEMORIES, the Prague Dependency Treebank, ANCORA, and ONTONOTES in separate sections.

4.3.3.1 The MATE Markup Scheme

The objective of the *Multilevel Annotation Tools Engineering* (MATE) project was to develop an annotation workbench supporting multilevel annotation in dialogue [42]. The project built on XML standoff technology developed in the MULTEXT project [30], and in particular on its application in the MapTask corpus [33]. The levels to be supported by the workbench included morphosyntax, prosody, dialogue acts, coreference, and disfluencies; for each of these levels a document was produced analyzing the needs of that type of annotation, and proposing a markup scheme that could support those needs.

The MATE proposals for coreference [36, 54] were based on an analysis of the best known coding schemes of the time, including MUC-style coreference, the more general notion of anaphoric reference and associative anaphora, supported by DRAMA (the scheme developed by Passonneau for the Discourse Resource Initiative) [48], and the MapTask reference scheme [4] supporting reference proper, i.e., mention of objects in the visual situation which may or may not have been linguistically introduced. The analysis also took into account the problems with the MUC scheme identified in work such as [15].

The markup scheme derived from this analysis incorporated not only devices to support MUC-style annotation, but also the annotation of an arbitrary number of anaphoric relations between a mention and previous entities through the use of linking elements derived from the `LINK` elements from the Text Encoding Initiative (TEI) [69], as well as the `UNIVERSE` device developed in the area of multimodal reference annotation to associate `IDs` to non-linguistic entities [7]. The markup also aimed at covering zero anaphora in languages other than English, and discourse deixis through the use of the `SEG` element, also developed by TEI.

The `coref` level for anaphora and coreference has two main elements: a `<coref:de>` tag for mentions, and a separate `<coref:link>` element to mark anaphoric relations. The use of these elements is illustrated in Figure 4.1. The MATE markup relied on so-called **token** standoff as in the MapTask, where the elements of the level file (`coref.xml` in Figure 1) point to tokens in the base file using hyperlinks (`words.xml` in Figure 1).

The form of `coref:link` proposed in MATE differed from that used in TEI by being structured–the `coref:link` only specifies the anaphor and the relation between anaphor and antecedent, the selected mention of the antecedent is marked using a separate `coref:anchor` element so as to allow coders to mark antecedent ambiguity (see discussion of the ARRAU coding scheme below).

The coreference markup scheme proposed in [36] was not implemented in the MATE toolkit, but using standoff for anaphoric annotation has become fairly standard. Aspects of the MATE markup scheme directly influenced the design of the

```
words.xml
...
<word ID="w1">we</word>
<word ID="w2">’re</word>
<word ID="w3">gonna</word>
<word ID="w4">take</word>
<word ID="w5">the</word>
<word ID="w6">engine</word>
<word ID="w7">E3</word>
<word ID="w8">and</word>
<word ID="w9">shove</word>
<word ID="w10">it/word>
<word ID="w11">over/word>
<word ID="w12">to/word>
<word ID="w13">Corning/word>
...

coref.xml:
...

<coref:de ID="de00" href="words.xml#id(w1)"/>
<coref:de ID="de01" href="words.xml#id(w5)..id(w7)"/>
<coref:de ID="de02" href="words.xml#id(w10)"/>
<coref:de ID="de03" href="words.xml#id(w13)"/>

<coref:link href="coref.xml#id(de02)"  type="ident">
      <coref:anchor href="coref.xml#id(de01)"/>
</coref:link>
```

Fig. 4.1 Mentions and links in the MATE markup scheme.

markup scheme supported by the MMAX2 annotation tool discussed below [44]. Other types of standoff are supported by CALLISTO and other annotation tools based on the ATLAS architecture [5].

4.3.3.2 The GNOME Corpus

The MATE proposals only identified a range of options without deciding among the alternatives. The GNOME corpus[8] [49, 51] was the first corpus annotated according to a coding scheme chosen among those options and using (a variant of) the markup scheme proposed in MATE. It was annotated to support research on the effect of local and global salience on the generation of referring expressions [57, 56].[9] The corpus consists of documents from three domains: the Museum Domain, including museum labels and material from museum catalogues; the Pharma Domain, consisting of

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

[9] The corpus was also subsequently used to study text structuring [35] and aggregation [12] as well as anaphora resolution [34].

several medicine leaflets; and the Sherlock domain, consisting of tutorial dialogues collected as part of the Sherlock project at the University of Pittsburgh and whose discourse structure was annotated according to Relational Discourse Analysis, or RDA [43]. The aim was to have around 5,000 markables for each domain; the total size of the corpus is around 40,000 tokens.

Markup Scheme

Due to the lack of availability of annotation tools supporting standoff (the MATE toolkit was only completed after the end of the GNOME annotation), an inline version of the MATE markup scheme was used. Attributes were marked for the elements `s` (sentences), `unit` (local update candidates), `ne` (the equivalent of the `coref:de` element of the MATE markup scheme), and `mod` (NP modifiers).

Anaphoric information was annotated through separate `ANTE` elements implementing the `COREF:LINK` elements of the MATE scheme. The `ANTE` elements had two attributes: `CURRENT` (the ID of the anaphor) and `REL` (the relation holding between the entity referred to by the anaphor and the antecedent entity in the discourse model). The embedded `ANCHOR` element coded the last mention of the antecedent. (See Figure 4.2.) Multiple `ANCHOR` elements indicated ambiguity.

```
<NE ID="ne07">Scottish-born, Canadian based jeweller, Alison
Bailey-Smith</NE>
....
<NE ID="ne08"> <NE ID="ne09">Her</NE> materials</NE>
...
<ANTE  CURRENT="ne09" REL="ident">
       <ANCHOR ANTECEDENT="ne07" />
</ANTE>
```

Fig. 4.2 Markup of anaphoric information in the GNOME corpus using separate and structured links.

Guidelines

As the corpus was annotated to study salience, a lot of information was annotated besides information about anaphoric relations, including information about document structure, potential local update units (the ‘utterances’ of Centering), and a variety of information about mentions. This includes morphosyntactic information (gender, number and person, grammatical function), semantic information (semantic function, semantic type–abstract or concrete, animate or inanimate, etc.–whether the object referred to is singular, mass or plural, functionality, genericity, etc.) and discourse information (e.g., whether the markable performed a deictic reference) [51, 50]. The information annotated for the `ne` element is shown in Figure 4.3.

```
<ne id="ne109"
    cat="this-np" per="per3" num="sing" gen="neut" gf="np-mod"
    lftype="term" reference="direct"
    onto="concrete" ani="inanimate" structure="atom"
    count="count-yes" generic="generic-no"
    deix="deix-yes" loeb="disc-function">
    this  monumental cabinet </ne>
```

Fig. 4.3 Morphosyntactic, semantic and discourse information about mentions in the GNOME corpus.

One of the key contributions of the work on GNOME was the decision to only annotate information that could be coded reliably [9, 3]. In particular, a systematic investigation was carried out of the types of associative ('bridging') relations that could be reliably annotated, building on the earlier work by Poesio and Vieira [59]. Separate reliability studies were carried out for all the attributes.

Availability

At present the GNOME corpus is available from the authors (see website at previous page); a MMAX2 version will soon become available through the Anaphoric Bank.

4.3.4 The ARRAU *and* LiveMemories *Corpora*

The objectives of the ARRAU project[10] were to further investigate 'difficult' cases of anaphoric reference, and in particular, ambiguous anaphoric expressions and cases of discourse deixis [58]. This required looking in greater detail than earlier work at agreement on anaphoric reference as κ was not appropriate [3]. These investigations led to the development of a coding scheme that was then employed for annotating the ARRAU corpus [53]. This corpus was also intended to include texts from genres not traditionally covered by anaphoric corpora, in particular dialogue and narrative, and therefore includes a full annotation of the task-oriented dialogues in the TRAINS-93 corpus,[11] and the complete collection of spoken narratives in the Pear Stories [11], often used to study salience. The corpus also includes news articles (the entire subset of the Penn Treebank that was annotated in the RST treebank [10]) and additional documents from the GNOME genres. The ARRAU guidelines were then adapted to annotate anaphora in Italian, and the LiveMemories corpus was created [67].

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

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

4.3.4.1 Markup Scheme

ARRAU and LIVEMEMORIES were annotated using the MMAX2 annotation tool discussed in Section 4.1. 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. The phrase layer also contains a number of attributes encoding semantic information.

4.3.4.2 Guidelines

The coding scheme inherits several aspects of the GNOME coding scheme, although with fewer attributes, but adding the ability to annotate discourse deixis, and more extensive provision for annotating ambiguity–for instance, the possibility of marking an ambiguity between a discourse-new and discourse-old reading, which was not possible with the GNOME scheme. The reliability of the coding scheme for ambiguity was also tested, with inconclusive results however [52].

Markable Definition

In ARRAU, all NPs are coded as markables at the `phrase` level. 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 (4.28a), and when the premodifier refers to a kind, like *exchange-rate* in (4.28b). Singletons are also marked as markables that are part of coreference chains.

(4.28)a. ... 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.

b. 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 ... ...

The full NP is marked with all its modifiers; in addition, a `min` attribute is marked, as in the MUC corpora.

All markables at this level are annotated for morphosyntactic agreement (gender, number and person), grammatical function (following the GNOME scheme), and `reference` (the values being non-referring, discourse-new, and discourse-old). Non-referring markables include expletives and predicative NPs (as standard), but also, more controversially, quantifiers and coordination. Referring mentions (mentions of discourse-new and discourse-old entities) also have a `category` attribute specifying the semantic type of the entity: `person`, `animate`, `concrete`, `organization`, `space`, `time`, `numerical`, `plan` (for actions), or `abstract`. Referring mentions also have a genericity attribute, also annotated following the GNOME guidelines.

Range of relations

All referring NPs are marked as either `new` or `old`. If marked as `old`, an antecedent can be identified, either of type `phrase` (already mentioned using an NP) or `segment` (not mentioned using an NP, in cases of discourse deixis). Referring NPs can be marked as ambiguous between a discourse-new and a discourse-old interpretation; discourse-old NPs 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. In addition, referring NPs can be marked as **related** to a previously mentioned discourse entity (associative or bridging anaphors). Associative descriptions were identified following the GNOME guidelines, but the type of relation was not explicitly marked.

4.3.4.3 Availability

The ARRAU corpus is available from LDC; it will also be made available through the Anaphoric Bank.[12]

4.3.4.4 The LIVEMEMORIES Corpus

The ARRAU guidelines were adapted to create the LIVEMEMORIES corpus of anaphora in Italian, containing texts from Wikipedia and blogs released through a Creative Commons license.

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, discussed below). In the LIVEMEMORIES corpus there are two types of markables: **nominal** markables, for nominal expressions and clitic particles, and **verbal** markables for zeros and incorporated clitics. The type of markable is specified by the `markable_type` attribute. In the case of a zero, the first element of the verbal complex following the position of the zero is identified as a verbal markable; in the case of an incorporated clitic, the verbal element is that to which the clitic is incorporated. Example (4.29) shows examples of nominal markables (with index $_n$) and verbal markables (with index $_v$).

(4.29) ...[Il giudice]$_n$ [gli]$_{n_i}$ nego' [questa richiesta]$_n$ e procedette invece ad acquistare [alcuni indumenti da [fargli]$_{v_i}$ indossare]$_n$

[12] 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.

> *The judge [to-him] rejected this request and proceeded instead to buy some clothes to make-[to-him] wear.*

The attribute verbal_type specifies the type of verbal markable: either clitic or empty_subject. In case multiple clitics are incorporated in the same verbal element (as in *darglielo*), multiple verbal markables are created. The annotation was used as the basis for the proposals concerning zero resolution in Italian and Japanese by [32].

An early annotation of about half of the Wikipedia subset of the LIVEMEMORIES corpus was used for the SEMEVAL-2010 coreference evaluation and is available in CONLL-style tabular format as part of that dataset. The entire corpus was used for the EVALITA-2011 evaluation of Italian resources. The entire corpus is available through the Anaphoric Bank.

4.3.5 The Prague Dependency Treebank

The Prague Dependency Treebank 2.0[13] (PDT 2.0) [23] is a corpus of samples from the Czech National Corpus (news and scientific articles) annotated according to the specifications of **Functional Generative Description**, a linguistic formalism developed by the Prague School since the 60s [68]. The annotation involves three levels:

m-layer The morphological layer contains POS and morphological information–Czech being a highly inflected language. This is available for over 2 million words.

a-layer The analytic layer specifies the surface syntactic structure of the sentence in the form of a dependency tree. This is available for around 1.5 million words.

t-layer The tectogrammatical layer specifies predicate-argument structure, topic-focus articulation, and coreference (pronouns only).

Until the recent release of ONTONOTES version 5.0, the PDT 2.0 was the largest anaphorically annotated corpus (although only anaphoric relations involving pronouns were annotated), and is still arguably the most advanced corpus from a linguistic and technologic perspective. We limit our discussion here to the anaphoric annotation as discussed in [38].

4.3.5.1 Markup Scheme

Each annotation layer builds on (and is linked to) the previous layer as shown in Figure 4.4, the PDT representation of the Czech sentence *Byl by s̃el dolesa*, *He-was would went to forest*.

[13] http://ufal.mff.cuni.cz/pdt2.0/

A variety of markup formats were used in the past for the layers, but the PDT 2.0 was standardized on PML, an XML format designed for linguistic annotation. The m-layer is annotated completely automatically; the a-layer and t-layer are annotated semi-automatically, by first running an automatic annotator and then having the coders correct mistakes and add information. The markup is however completely transparent to the coders, who annotate using a dedicated annotation tool called TRED.

4.3.5.2 Guidelines

Two types of anaphoric information is annotated: **grammatical** coreference (control verbs, reflexives, relative pronouns) and **textual** coreference. Only personal and demonstrative pronouns are annotated, but a very wide variety of types of (identity) anaphoric reference are annotated, including not just reference to antecedents introduced by nominals, but also discourse deixis and **exophoric** reference to entities that are part of common knowledge [38].

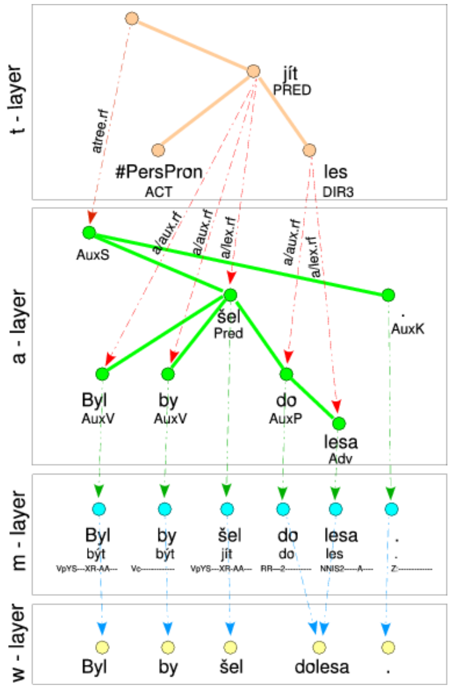

Fig. 4.4 The three annotation layers in the Prague Dependency Treebank.

4.3.5.3 Availability

The PDT is available through LDC.

4.3.6 The AnCora *Corpora*

The Annotated Corpora (AnCora)[14] of Spanish and Catalan are the result of years of annotation at different linguistic levels [71]. The corpora began as an initiative by the University of Barcelona, the Technical University of Catalonia, and the University of Alicante to create two half-million-word treebanks for Spanish and Catalan that could be used as training and test data for supervised machine learning, and as input for corpus-based linguistic studies. The initiative was continued by the University of Barcelona in an effort to further enrich the corpora with grammatical relations, argument structures, thematic roles, semantic verb classes, named entities, WordNet nominal senses, and, more recently, coreference relations [66].[15] AnCora are the first and largest corpora of Spanish and Catalan with coreference information including not only pronouns but all NPs. The two datasets, AnCora-CO-Es and AnCora-CO-Ca, consist of newspaper and newswire articles from *El Periódico* newspaper, the Spanish EFE news agency, and the Catalan ACN news agency.

Markup Scheme

The different layers of annotation, including coreference, are all marked up with inline XML tags. Unlike other corpora like MUC and ACE that began from scratch, markables in AnCora were identified based on the already existing syntactic annotations (see below for the list of syntactic nodes that were considered as markables). All referring mentions, including singletons, are annotated with an `entityref` attribute. If two or more mentions refer to the same entity, they all receive an `entity` attribute with the same ID value. The second and subsequent mentions in a coreference chain include a `coreftype` attribute that specifies the type of relation with the previous mention. The morphosyntactic and semantic markup of mentions is illustrated in (4.30) for the NP *el Consejo_de_Seguridad* 'the Security_Council'. The markup of coreference information is shown in (4.31).

(4.30)
```
<sn arg="arg0" entityref="ne" func="suj"
ne="organization" tem="agt"> <spec gen="m" num="s">
<d gen="m" lem="el" num="s" postype="article"
wd="el"/> </spec> <grup.nom gen="m" num="s">
```

[14] `http://clic.ub.edu/corpus/en`

[15] The portion of AnCora annotated with coreference information (AnCora-CO) amounts to a total of 400,000 words for each language.

```
<n lem="Consejo_de_Seguridad" ne="organization"
postype="proper" wd="Consejo_de_Seguridad"/>
</grup.nom> </sn>
```

(4.31)

```
<sn entity="entity5" entityref="ne"> el Consejo_de_
Seguridad </sn> no recogió en <d coreftype="ident"
entity="entity5" entityref="spec" wd="su"/>
declaración ...
```

[The Security_Council]$_i$ did-not include in [[their]$_i$ declaration]$_j$ ...

Alternatively, AnCora is also available in the CoNLL-style tabular format that was used for the SemEval-2010 task on coreference resolution [65]. See Section 4.3.8 below for further details.

Guidelines

The annotation scheme that was used for AnCora is inspired by the MATE guidelines (Section 4.3.3), as the resulting corpus was meant to be a comprehensive language resource rather than to serve the purpose of a specific evaluation campaign. Thus, the definition of both markables and coreference relations was linguistically motivated.

Markable Definition

As already mentioned, the coreference annotation in AnCora benefits from the existing syntactic annotation and asks annotators to consider as markables the following five syntactic nodes: (i) NPs (including elliptical subjects[16]), (ii) nominal groups in a conjoined NP, (iii) relative pronouns, (iv) possessive determiners, and (v) possessive pronouns. Additionally, non-nominal nodes (i.e., verbs, clauses, and sentences) are annotated if they are the antecedents in a discourse-deixis relation. A verb can also be annotated if it contains an incorporated clitic. Relying on the (manual) syntactic level ensures that markables include all premodifiers and postmodifiers; no `MIN` attribute is annotated.

To filter out the NP nodes that are not referential, the attribute `entityref` takes the values `ne`, `spec` or `nne` for referential mentions. The first value identifies named entities (e.g., *Barcelona*) belonging to six semantic types: person, organization, location, date, number, and others (publications, prizes, etc.). The second value identifies mentions that are not a named entity in form (e.g., pronouns, NP headed by a common noun), but that corefer with an NE. The third value indicates mentions that neither are a named entity in form nor refer to a named entity. The `entityref` attribute is included for both singletons and coreferent mentions, thus making it possible to extract singletons. Non-referential mentions (e.g., predicates)

[16] Elliptical subjects were manually inserted as part of the treebank.

either lack this attribute or receive the value `lex` if they are (within) a lexicalized expression, like *cats* and *dogs* in *to rain cats and dogs*.

Range of Relations

Of the range of relations proposed in MATE, ANCORA focused on three, which correspond to the three values that the attribute `coreftype` can take: `ident` (referential identity), `pred` (predication), and `dx` (discourse deixis). Following the MATE proposal, predication is separated from referential identity, and discourse deixis is also annotated, but bridging relations are not.

All the mentions with an `entityref` value of `ne`, `spec` or `nne` can participate in a relation of identity (4.32) or discourse deixis (4.33), whereas predicative relations (4.34) involve a non-referential mention, namely one lacking `entityref`. Identity relations that have a split antecedent (see (4.13b) above) are annotated by creating an entity that is the sum of two or more entities. In discourse deixis, the extent of the discourse segment is identified according to the syntactic annotation, thus it must correspond to one of the available phrasal nodes at the verbal, clausal or sentential level.

(4.32)a. [ES] Sobre la ausencia de [Argentina]$_i$ en la reunión, sólo se informó de que hubo una comunicación de los servicios sanitarios de [ese país]$_i$.
b. [EN] On the absence of [Argentina]$_i$ in the meeting, it was only reported that there was a communication from the health services of [that country]$_i$.

(4.33)a. [ES] ... algunos expertos calculan [que el precio del crudo ... llegará a 40 dólares a_finales_de este año]$_i$, pero que la OPEP hará "todo lo posible para que [eso]$_i$ no ocurra".
b. [EN] Some experts estimate [that oil prices will reach $40 by the end of this year]$_i$, but that OPEC will do "everything to ensure that [this]$_i$ does not happen".

(4.34)a. [ES] ... una posible fusión de la operadora española con [British_Telecom [(BT)]$_i$]$_i$
b. [EN] A possible merger of the Spanish operator with [British_Telecom [(BT)]$_i$]$_i$

Additionally, predicative relations and discourse deixis take the attribute `coref-subtype` that specifies further semantic information. Predicates are either `definite` (i.e., identifying) or `indefinite` (i.e., non-identifying). Discourse-deictic mentions can refer to the same `token` as the antecedent, the same event `type` as the antecedent, or the `proposition` (the actual words) of the antecedent, which is often the case with speech verbs (e.g., *He didn't say **this***).

Availability

The ANCORA corpora are freely available from `http://clic.ub.edu/corpus/en`. The column-based version that was used in SEMEVAL-2010 can be downloaded

at
http://stel.ub.edu/semeval2010-coref/download.

4.3.7 ONTONOTES

The ONTONOTES project [77, 64] created a multilingual corpus of large-scale, accurate, and integrated annotation of multiple levels of the shallow semantic structure in text. It spans multiple genres across three languages – English, Chinese and Arabic. The English and Chinese portions contain 1.6M words and 1M words, respectively, from newswire, broadcast news, broadcast conversation, web text, and telephone conversation. An English translation of the New Testament was also annotated as a pivot corpus to facilitate machine-translation research. The Arabic portion is relatively small, comprising 300K of newswire text. It is the largest corpus of English, Chinese and Arabic annotated with coreference. Such multi-layer annotations, with complex, cross-layer dependencies, demand a robust, efficient, scalable mechanism for storing them while providing efficient, convenient, integrated access to the underlying structure. To this effect, it uses a relational database representation that captures both the inter- and intra-layer dependencies and also provides an object-oriented API[17] for efficient, multi-tiered access to the data [64].

The coreference portion of ONTONOTES captures general anaphoric coreference that covers entities and events not limited to noun phrases, or a limited set of entity types [63, 61, 62]. The aim of the project was to annotate linguistic coreference using the most literal interpretation of the text at a very high degree of consistency, even if it meant departing from a particular linguistic theory. Two different types of coreference are distinguished: Identical (IDENT), and Appositive (APPOS). Appositives are treated separately because they function as attributions; the IDENT type is used for anaphoric coreference, meaning links between pronominal, nominal, and named mentions of specific referents. It does not include mentions of generic, underspecified, or abstract entities. All the data was double blind annotated and adjudicated.

4.3.7.1 Markup Scheme

The corpus is annotated using inline SGML, similar to the MUC corpus except that the MIN mention span is not identified as there is gold treebank infomation from which one can derive the syntactic head. Every markable that belongs to a coreference chain is identified with a <COREF> tag; <COREF> elements have three attributes: i) ID, the identifier for a mention; ii) TYPE, which can be IDENT or APPOS; and iii) SUBTYPE, which is only for the APPOS types, and can be either HEAD or ATTRIB. The first mention of a coreference chain uses the attribute ID to assign an ID to

[17] http://cemantix.org/software/ontonotes-db-tool.html

the coreference chain, and every subsequent mention uses the same ID to specify the coreference chain to which it belongs. In case of conversational data and web data, where speaker or writer could be identified, it was captured in the SPEAKER attribute.

The majority of the ONTONOTES annotation is based on the tokens in the treebank. However, a solution was needed for identifying partial-token mentions, such as *Walmart* in tokens such as *Pro-Walmart*; or *India* and *Japan* in a token such as *India/Japan*, which are not separated into distinct tokens during treebanking. This was not a problem for CALLISTO, the annotation tool, but reconciling sub-token spans with the SGML markup needed to be addressed. This was done by using two optional attributes, S_OFF and E_OFF, that identified the start and end offset of the string. Many a times, the partial token is either a prefix or a suffix, and so usually only one of these two attributes need to be specified, and the other attribute defaults to either zero (for S_OFF) or the length of the mention in characters (for E_OFF). For example, in the case of *Pro-Walmart*, the mention *Walmart* is identified with a S_OFF of 4, and the E_OFF is absent. And, for *India* in *India/Japan*, the S_OFF is absent, and the E_OFF is 5, whereas for *Japan*, the S_OFF is 6, and E_OFF is absent.

Some of the broadcast and telephone conversation documents were very long as they typically include transcriptions of recordings of entire shows that cover various topics. Full-document coreference annotation was not an option. Therefore, the documents were manually segmented into multiple parts, breaking along story boundaries as much as possible, and these were annotated independently of each other, and therefore the coreference chains do not carry information across parts. Each part is encoded in a separate TEXT segment with a PARTNO attribute.

Example (4.35) shows a sample markup of an ONTONOTES document.

(4.35)
```
<DOC DOCNO="bc/cnn/00/cnn_0003@0003@cnn@bc@en@on">
<TEXT PARTNO="000">
...
<COREF ID="26" TYPE="IDENT" E_OFF="1" SPEAKER="Linda_Hamilton">
I-</COREF> <COREF ID="26" TYPE="IDENT" SPEAKER="Linda_Hamilton">I
</COREF> 'm sure 0 there is *?* .
```

*I- I 'm sure 0 there is *?* .*

```
Um if <COREF ID="26" TYPE="IDENT" SPEAKER="Linda_Hamilton">I
</COREF> were <COREF ID="14" TYPE="IDENT" SPEAKER=
"caller_7">you </COREF> , because <COREF ID="26" TYPE="IDENT"
SPEAKER="Linda_Hamilton">I</COREF> do n't know <COREF ID="43"
TYPE="IDENT">that number</COREF> off hand um <COREF ID="14"
TYPE="IDENT" SPEAKER="caller_7">you</COREF> can call
<COREF ID="70" TYPE="IDENT">the University of Medicine
and Dentistry in <COREF ID="50" TYPE="IDENT">New Jersey</COREF>
</COREF> .
```

Um if I were you , because , I do n't know that number off hand um you can call the University of Medicine and Dentistry in New Jersey .

```
Um oh <COREF ID="74" TYPE="IDENT">they</COREF> would have *-1
```

```
to know where in <COREF ID="50" TYPE="IDENT">New Jersey</COREF>
then .
```

*Um oh they would have *-1 to know where in New Jersey then .*

```
...
</TEXT>
</DOC>
```

4.3.7.2 Guidelines

The ONTONOTES coreference guidelines are mostly inspired by the MUC and ACE tasks, and are consistent with the DRAMA / MATE ideas. As in MUC, all NPs – irrespective of their semantic type – are linked with coreferent NPs, and *singleton* entities are left out. We look now at some salient aspects of the guidelines.

Generics are not considered as markables unless they are referred to by neighboring pronouns. Generic nominal mentions can be linked with referring pronouns and other definite mentions, but are not linked to other generic nominal mentions. This allows coreference between the bolded mentions in (4.36) and (4.37), but not in (4.38).

(4.36) **Officials** said **they** are tired of making the same statements.

(4.37) **Meetings** are most productive when **they** are held in the morning. **Those meetings**, however, generally have the worst attendance.

(4.38) Allergan Inc. said it received approval to sell the PhacoFlex intraocular lens, the first foldable silicone lens available for ***cataract surgery**. The lens foldability enables it to be inserted in smaller incisions than are now possible for ***cataract surgery**.

Pronouns Pleonastic pronouns and generic *you* are not treated as markables.

Premodifiers Only non-adjectival premodifiers can be markables. Proper nouns that are morphologically adjectival are treated as adjectives. For example, adjectival forms of GPEs such as *Chinese* in *the Chinese leader*, are not linked. Thus, *United States* in *the United States policy* can be linked with another mention of the same entity, but not *American* in *the American policy.* GPEs and nationality acronyms (e.g., *U.S.S.R.* or *U.S.*) are also considered as adjectival. Premodifier acronyms are marbles unless they refer to a nationality. Thus, *FBI* is a markable in (4.39), but not *U.S.* in (4.40). cannot.

(4.39) **FBI** spokesman

(4.40) ***U.S.** spokesman

Events In addition to NP entities, events described by NPs and verbs are annotated as well. Only events that are (usually) introduced by a verb and then coreferred using an NP were annotated in order to keep the task manageable. This includes morphologically related nominalizations, *grew* and *the strong growth* in (4.41), and NPs that refer to the same event, even if they are lexically distinct from the verb (4.42).

(4.41) Sales of passenger cars **grew** 22%. **The strong growth** followed year-to-year increases.

(4.42) Japans domestic sales of cars, trucks and buses in October **rose** 18% from a year earlier to 500,004 units. **The strong growth** followed year-to-year increases of 21% in August and 12% in September.

Copular and Predicative Copular and predicative constructions as well as small clause constructions are not markables: a separate attributive link is used for them.
Like copulas, small clause constructions are not marked. Example (4.43) is treated as if the copula were present (*John considers Fred to be an idiot.*)

(4.43) John considers *__Fred *an idiot__.

Appositives are not marbles, but marked with special labels. For example, in (4.44), an APPOS(itive) link is annotated between *Washington* (marked as HEAD) and *the capital city* (marked as ATTRIB (ute)). The intended semantic connection is then filled by supplying the implicit copula. An APPOS chain contains at least one HEAD mention and one or more ATTRIB mentions.

(4.44) **Washington** $_{\text{HEAD}}$, **the capital city**$_{\text{ATTRIB}}$, is on the East coast.

When the entity to which an appositive refers is also mentioned elsewhere, only the single span containing the entire appositive construction is included in the larger IDENT chain. None of the nested NP spans are linked. In example (4.45), the entire span can be linked to later mentions of Richard Godown.

(4.45) Richard Godown, president of the Industrial Biotechnology Association

Metonymy As mentioned in Section 4.2.2, metonymic referents were treated as separate entities to meet the required level of annotation consistency.
Part/Whole and other associative relations were not annotated.
Zero Anaphora For the most part the guidelines are language independent. However, unlike English, Chinese and Arabic are pro-drop languages, in which pronouns may be omitted and filled from the context. The treebank introduces and tags all these constituents. All these (i.e., ⋆ and ⋆pro⋆) were considered as markables.

Markable Definition

Since all the text in ONTONOTES had been treebanked prior to coreference annotation, hand-tagged NPs were available. From the point of view of consistency and completeness, the starting set of markables was based on the hand-tagged NPs. In addition, all relative pronouns (PRP$), which do not usually constitute an NP by themselves, were considered as markables. There were two type of markables that

Language	Genre	A1-A2	A1-ADJ	A2-ADJ
English	Newswire [NW]	80.9	85.2	88.3
	Broadcast News [BN]	78.6	83.5	89.4
	Broadcast Conversation [BC]	86.7	91.6	93.7
	Magazine [MZ]	78.4	83.2	88.8
	Weblogs and Newsgroups [WB]	85.9	92.2	91.2
	Telephone Conversation [TC]	81.3	94.1	84.7
	Pivot Text [PT] (New Testament)	89.4	96.0	92.0
Chinese	Newswire [NW]	73.6	84.8	75.1
	Broadcast News [BN]	80.5	86.4	91.6
	Broadcast Conversation [BC]	84.1	90.7	91.2
	Magazine [MZ]	74.9	81.2	80.0
	Weblogs and Newsgroups [WB]	87.6	92.3	93.5
	Telephone Conversation [TC]	65.6	86.6	77.1
Arabic	Newswire	73.8	88.1	75.6

Table 4.1 [SP: updated this table]Inter Annotator (A1 and A2) and Adjudicator (ADJ) agreement for the Coreference Layer in ONTONOTES, measured in terms of the MUC score.

were later added by the annotators: verbs triggering an eventive chain (the head verb was annotated as a markable), and portions of flat non-NP constituents (usually names) in an NP: these were marked and tagged, but constitute a very small portion (˜2%) of the total markables.

Since there was a hand-tagged treebank underlying the annotations, the syntactic heads of the markables could be determined with high enough accuracy, and so the MIN attribute from MUC was not added. Similarly to MUC, it was difficult to identify the head in case of conjunctive constructions. In the spoken genre, there are often pronominal references to the speaker(s) and, given that the speaker metadata was available, this was tagged alongside the sentence during annotation, which made it easier for the annotators to disambiguate the pronouns. One of the pronouns is connected to the speaker metadata markable, thus speaker information is propagated throughout the coreference chain.

Agreement

Table 4.1 shows the inter-annotator and annotator-adjudicator agreement on all the genres in ONTONOTES.

A set of 15K disagreements in various parts of the data were classified into one of the categories shown in Figure 4.5. Genuine ambiguity and annotator error were the biggest contributors – the latter of which is usually captured during adjudication, thus showing the increased agreement between the adjudicated version and the individual version.

4.3.7.3 Availability

ONTONOTES is available free of charge for research purposes from LDC.

4.3.8 The SemEval*-2010 Task 1 Corpus*

The 2010 edition of the SemEval evaluation campaign included a multilingual coreference resolution task[18] [65]. The datasets used for the task included subsets of the COREA corpus for Dutch (104,000 words), the AnCora corpora for Spanish (380,000 words) and Catalan (345,000 words), the LiveMemories corpus for Italian (140,000 words), the Tüba/DZ corpus for German (455,000 words), and the OntoNotes corpus for English (120,000 words).

The most valuable contribution of the task was to convert all the datasets to a common format and annotate them in the most similar and consistent manner,[19] thus providing a multilingual corpus of coreference that can be easily used to train and test coreference resolution systems for different languages, and to compare their results. Unlike the corpora used in MUC and ACE, all NPs are considered, single-

Type	Description
Annotator Error	An annotator error. This is a catch-all category for cases of errors that do not fit in the other categories.
Genuine Ambiguity	This is just genuinely ambiguous. Often the case with pronouns that have no clear antecedent (especially this & that)
Generics	One person thought this was a generic mention, and the other person didn't
Guidelines	The guidelines need to be clear about this example
Callisto Layout	Something to do with the usage/design of Callisto
Referents	Each annotator thought this was referring to two completely different things
Possessives	One person did not mark this possessive
Verb	One person did not mark this verb
Pre Modifiers	One person did not mark this Pre Modifier
Appositive	One person did not mark this appositive
Extent	Both people marked the same entity, but one person's mention was longer
Copula	Disagreement arose because this mention is part of a copular structure a) Either each annotator marked a different half of the copula b) Or one annotator unnecessarily marked both

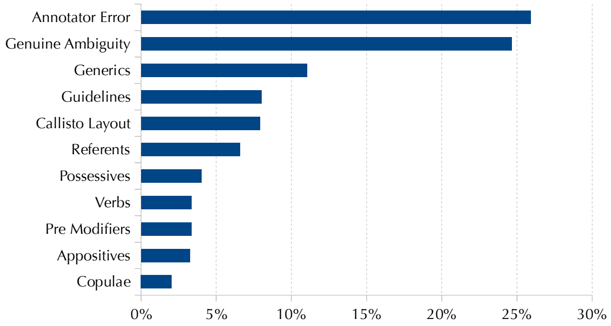

Fig. 4.5 Frequency of each type of disagreement.

[18] http://stel.ub.edu/semeval2010-coref

[19] The morphosyntactic and semantic tag sets differ between languages.

tons[20] are included, and predicative relations are not annotated. A further asset of the corpus is that it contains both gold-standard and automatically predicted morphosyntactic and semantic information.

4.3.8.1 Markup Scheme

The SEMEVAL-2010 Task 1 datasets are formatted following the CONLL-style tabular format based on dependency relations. There is one line per token, and the different layers of annotation for each token are displayed across multiple tabular-separated columns. Although not all the datasets include every layer of linguistic annotation, they usually contain the token ID in the sentence, the actual token, lemma, part of speech, morphological features (e.g., number, gender, tense), head, dependency relation, named entity type, predicate semantic class, semantic dependency, and coreference information. Apart from the first two columns and the last column (containing coreference relations), columns are repeated for each level of linguistic information to provide the gold-standard and automatically predicted information.

Coreference relations are represented in open-close notation with the entity number in parentheses. Every entity has an ID, and every mention is marked with the ID of the entity it refers to: an opening parenthesis indicates the first token of the mention, whereas a closing parenthesis indicates the last token of the mention. If a mention consists of one single token, the opening and closing parentheses appear in the same line separated by the entity ID. If a token belongs to more than one mention, a pipe symbol separates the multiple entity IDs. Figure 4.6 illustrates the markup. Note the coreference relation between *high standards* and *the standards* with entity ID: 38.

4.3.8.2 Guidelines

None of the datasets was developed explicitly for the SEMEVAL task, thus the guidelines largely correspond to those of the respective source corpora. However, to make the evaluation as fair as possible between the different languages, the task organizers laid down a few principles that are summarized in this section. In some cases, the annotation of the source corpora had to be partially adapted or modified.

Markable Definition

Markables include all NPs and possessive determiners. Singletons also receive an entity ID. Non-referential NPs (e.g., predicates, appositions, expletive pronouns, etc.) are not annotated. Although an effort was made to ensure consistency between the different annotation schemes, datasets differ slightly. For instance, the Dutch dataset

[20] In the case of ONTONOTES, the singletons were heuristically added.

```
1  Inherent  _ _ JJ  _ JJ  _ 5  _ PRD      _ _ _ _         _ arg2 _      _ _ _
2  in        _ _ IN  _ IN  _ 1  _ LOC      _ _ _ _         _ _    _      _ _ _
3  the       _ _ DT  _ DT  _ 4  _ NMOD     _ _ _ _         _ _    _      _ _ (18
4  law       _ _ NN  _ NN  _ 2  _ PMOD     _ _ _ _         _ _    _      _ _ 18)
5  is        _ _ VBZ _ VBZ _ 0  _ sentence _ _ _ be.01     _ _    _      _ _ _
6  the       _ _ DT  _ DT  _ 7  _ NMOD     _ _ _ _         _ _    _      _ _ (423|(380
7  vision    _ _ NN  _ NN  _ 5  _ SBJ      _ _ _ _         _ arg1 _      _ _ _
8  of        _ _ IN  _ IN  _ 7  _ NMOD     _ _ _ _         _ _    _      _ _ _
9  high      _ _ JJ  _ JJ  _ 10 _ NMOD     _ _ _ _         _ _    _      _ _ (38
10 standards _ _ NNS _ NNS _ 8  _ PMOD     _ _ _ _         _ _    _      _ _ 38)
11 ,         _ _ ,   _ ,   _ 7  _ P        _ _ _ _         _ _    _      _ _ 380)
12 and       _ _ CC  _ CC  _ 7  _ COORD    _ _ _ _         _ _    _      _ _ _
13 money     _ _ NN  _ NN  _ 12 _ CONJ     _ _ _ _         _ _    _      _ _ (421
14 to        _ _ TO  _ TO  _ 13 _ NMOD     _ _ _ _         _ _    _      _ _ _
15 meet      _ _ VB  _ VB  _ 14 _ IM       _ _ _ meet.01   _ _    _      _ _ _
16 the       _ _ DT  _ DT  _ 17 _ NMOD     _ _ _ _         _ _    _      _ _ (38
17 standards _ _ NNS _ NNS _ 15 _ OBJ      _ _ _ _         _ _    arg1   _ _ 38)|421)|423)
18 .         _ _ .   _ .   _ 5  _ P        _ _ _ _         _ _    _      _ _ _
```

Fig. 4.6 Markup of morphosyntactic, semantic and coreferential information in the SemEval-2010 Task 1 corpus.

only contains singletons for named entities, and expletive pronouns are annotated as singletons in the English dataset.

Range of Relations

The goal of the task was the development of systems that would solve the relations of referential identity between NPs. As a result, the only relation annotated in the datasets is referential identity, excluding predicates, discourse deixis, event anaphora, and bridging relations.

4.3.8.3 Availability

After the evaluation campaign, the organizers made freely available the development, training and test datasets of Catalan, Dutch, Italian, and Spanish at `http://stel.ub.edu/semeval2010-coref/download`. To acquire the German dataset, an Archiv-DVD from the tageszeitung must be purchased; detailed instructions are provided in the package with the rest of datasets. The English dataset is distributed by LDC.

4.3.9 Other Genres and Domain-Specific Corpora

Like most areas of Computational Linguistics, anaphora resolution is mainly focused on the genre of written news. All the corpora discussed or mentioned so far are collections of either news or broadcast data, and tend to focus on written language, with a few exceptions. Among the English corpora mentioned, GNOME is the one not focused on news: its three subcollections consist of pharmaceutical leaflets, museum catalogues and tutorial dialogues. ARRAU includes, in addition to a portion of the Wall Street Journal section of the Penn Treebank, the GNOME corpus as a subset, and also contains the TRAINS corpus and other dialogue material; and ONTONOTES also includes material from telephone conversations. Some substantial resources have however been created for other genres. They are briefly reviewed In this Section.

Spoken dialogue and online conversations

There are few corpora of anaphora in dialogue apart from those just mentioned, and they have generally been created for comparative studies of *it* vs. demonstratives *this* and LINGEXthat. Müller annotated the ICSI meeting corpus for a study of this type [45]. Navarretta [**?**] created the DAD corpora of abstract anaphora in Danish and Italian[21] that also focus on the study of demonstratives and pronouns.

More recently, more and more attention has been paid to online forums and other types of social media that can be seen as forming a type of 'textual conversation'. The LIVEMEMORIES ANAPHORA corpus discussed above [67] includes annotations of blogs in Italian as well as of Wikipedia pages. The SENSEI corpus, under construction, consists of annotations of online forums in English (from *The Guardian* newspaper) and Italian (from *La Repubblica* newspaper).

Technical and Scientifical Domains

Finally, there are anaphorically annotated corpora of technical and scientifical text. The NLP4EVENTS) corpus from the University of Wolverhampton is a collection of computer manuals. The domain with the most substantially anaphorically annotated corpora is Bio NLP. The best-known resource in this area is the GENIA corpus [22], that was also annotated for coreference in the GENIA-MEDCO project[23] [**?**]. This annotation was used for the 2011 BioNLP Shared Task on Coreference. Other anaphoric annotations of biomedical corpora have been carried out by Gasperin *et al.* [**?**] and as part of the creation of the Colorado Richly Annotated Full Text (CRAFT) corpus [**?**].

[21] `http://www.cst.dk/dad/`

[22] `http://www.nactem.ac.uk/genia/`

[23] `http://nlp.i2r.a-star.edu.sg/medco.html`

4.3.10 A Summary of Available Resources

Table 4.2 summarizes the corpora annotated with anaphora / coreference we are aware of, with references to the main publications and sites with information. Ongoing efforts as part of the Anaphoric Bank initiative[24] aim at making some of these anaphorically annotated corpora available in compatible markup formats. Some data are also available from the SemEval-2010 site.[25]

Language	Name	Reference	Size (words)
Arabic	ACE-2005[26]	[75]	100k
	OntoNotes 5.0[27]	[78]	300k
Bengali	ICON	[?]	
Catalan	AnCora-CO-Ca[28]	[66]	400k
Chinese	ACE-2005	[75]	≈200k
	OntoNotes 5.0	[78]	1200k
Czech	Prague Dependency Treebank 2.0[29]	[23]	≈800k
Dutch	COREA[30]	[25]	325k
English	MUC-6[31]	[21]	30k
	MUC-7[32]	[13]	30k
	GNOME[33]	[51]	40k
	ACE-2[34]		180K
	ACE-2005[35]	[75]	400k
	NP4Events[36]	[24]	50k
	ARRAU 2.0[37]	[53]	300k
	ICSI Meeting Corpus (dialogue)	[45]	
	GENIA-MEDCO (pronouns)[38]	[46]	800 documents
	OntoNotes 5.0	[78]	1450k
	Phrase Detectives	[?]	320k
French	CRISTAL-GRESEC / XRCE corpus (pronouns)[39]	[72]	1000k
	DEDE (definite descriptions)[40]	[17]	50k
German	Potsdam Commentary Corpus[41]	[70]	33k
	TüBa-D/Z [42]	[27]	600k
Hindi	ICON	[?]	
Italian	VENEX	[55]	40k
	i-Cab[43]	[40]	250k
	LiveMemories 1.0[44]	[67]	250k
Japanese	NAIST Text Corpus[45]	[31]	38k sentences
Portuguese	Summ-It[46]	[14]	50 documents
Russian	RU-EVAL		
Spanish	AnCora-CO-Es	[66]	400k
Tamil	ICON	[?]	
Tibetan	Tusnelda (B11)	[74]	<15k

Table 4.2 Anaphorically annotated corpora in different languages.

[24] http://www.anaphoricbank.org

[25] http://stel.ub.edu/semeval2010-coref/

4.4 Annotating Anaphora

As shown in Table 4.2, there are today quite a few corpora annotated with anaphoric information, and for many different languages, so researchers whose only interest is to develop and test domain-independent anaphoric resolvers, especially for English but also for many other languages including Arabic, Bengali, Catalan, Chinese, Czech, Danish, Dutch, German, Hindi, Italian and Spanish, have the resources to do so. However, there are many languages, genres, and domains for which resources are still lacking. Those interested in non-NP anaphora (e.g., ellipsis) and/or in these other languages, genres and domains, will therefore need to annotate their own data. This Section briefly discusses what this involves, beginning with a discussion of what tools are available, then discussing coding schemes for anaphora and agreement, and markable identification; for a more extensive discussion of annotation practice in general and anaphoric annotation in general, we recommend [**?**].

[26] https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications

[27] https://catalog.ldc.upenn.edu/LDC2013T19

[28] http://clic.ub.edu/ancora/

[29] http://ufal.mff.cuni.cz/pdt2.0/

[30] http://www.clips.ua.ac.be/~iris/corea.html

[31] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T13

[32] http://ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001T02

[33] http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/

[34] https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications

[35] https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications

[36] http://clg.wlv.ac.uk/projects/NP4E/\#corpus

[37] https://catalog.ldc.upenn.edu/LDC2013T22

[38] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Coreference+Annotation

[39] http://catalog.elra.info/product_info.php?products_id=634\&language=en

[40] http://www.cnrtl.fr/corpus/dede/

[41] http://www-old.ling.uni-potsdam.de/cl/cl/res/forsch_pcc.en.html

[42] http://www.sfs.uni-tuebingen.de/tuebadz.shtml

[43] http://www.celct.it/projects/icab.php

[44] http://www.anaphoricbank.org

[45] http://cl.naist.jp/nldata/corpus/

[46] http://www.inf.pucrs.br/~linatural/procacosa.html

4.4.1 Annotation Tools

The annotation of anaphora is a complex task. Related spans in the text need to be marked and set in relation to each other. Since a whole span of text has to be considered at once when looking for an antecedent, and visualization of (all) coreference chains is not always possible or desirable, care has to be taken to ensure the consistency of the annotated data and to support the annotation with adequate tools, which can help to ease common tasks (e.g., going back in the text to look for a same-head antecedent), lighten the cognitive load necessary for the annotation, and also help maintain consistency with formal specifications. The choice of tools also crucially affects the type of markup that can be used.

The organization of our discussion of the available corpora in Section 4.3 reflects the fact that anaphoric annotation can be divided into two more or less distinct phases: (i) the identification of markables in the text, and (ii) the identification of anaphoric relations between the entities realized by these mentions. For the latter task, two models can be used. One is *link-based*, the annotator marks the antecedent of a given NP by linking the anaphor and antecedent; the other model is *set-based*, where the annotator puts the elements of a coreference chain together into one group of markables. Both link-based and set-based annotation models have their quirks: in link-based annotation models, it is necessary to specify which antecedent the annotators should mark (either the closest, or the first one in the coreference chain, or –for definite NPs– the closest non-pronominal antecedent). Set-based annotation, on the other hand, does not easily allow marking uncertainty on the links.

Many anaphora annotation projects have been carried out using purpose-developed tools, such as TRED for the Prague Dependency Treebank.[47] In addition, there are a number of tools for 'generic' anaphoric annotation. Given this abundance of freely downloadable tools, which support the most typical coding schemes and UNICODE, developing one's own tool should only be considered as a last option. (There is always the risk of spending most of the time in the project creating the annotation tool.)

In the following, we present some of the best known freely downloadable annotation tools. (See also the annotation wiki at `http://annotation.exmaralda.org/index.php/Linguistic_Annotation` for links to additional information.)

CALLISTO

CALLISTO[48] was the tool used for all ACE annotation tasks and for the ONTONOTES annotation project. It uses a form of character standoff based on the ATLAS architecture, jointly developed by LDC, MITRE and NIST [5], which in turn is based on the idea of **annotation graphs**. The basic annotation procedure involves selecting

[47] `http://ufal.mff.cuni.cz/~pajas/tred/`

[48] `http://callisto.mitre.org/`

(‘swiping’) two markables in the main pane, and then specifying the relation between the two markables. CALLISTO is highly customizable, allowing for instance to specify whether coders can select words or characters, and a variety of export formats, such as APF used in ACE (see Section 4.3.2).

MMAX2

MMAX2[49] [44] uses token standoff, i.e., a standoff file format where one file (the *words* file) contains a list of the tokens, while other files (the *markable* files) contain one or multiple annotation layers, where an annotation layer contains exactly one type of markable - for example, it is possible to use one annotation layer for coreference annotation while using another annotation layer for annotation of discourse connectives. MMAX2 allows the user to specify the attributes of markables in a schema file where the kind of attribute (nominal, freetext) and the possible attribute values (for nominal attributes) can be described.

PALinkA

PALinkA [47] is specifically geared towards coreference annotation. It uses an inline XML format that does not allow overlapping markables, and at the same time offers an interaction mode that is challenging for novices, but that offers significant efficiency gains for expert annotators through avoidance of drag gestures (it is possible to mark a markable span through multiple clicks, which is significantly faster, but less intuitive, than marking the span by a click-and-drag gesture), and efficient keyboard shortcuts. PALinkA also allows the user to specify attributes for markables and markable relations.

AnCoraPipe

AnCoraPipe[50] is the tool used for annotating AnCora with different layers of annotation, including coreference information. It is a Java-based plug-in for Eclipse. It can be combined with Eclipse’s version-control plug-in to make it possible for several annotators to work simultaneously and easily synchronize their work. AnCoraPipe takes XML documents as input. For coreference, it follows the set-based annotation model, showing a list of all the entities in a document and the (coreferent) mentions in each of them. Annotations can be added by inserting mentions into entities, or merging, splitting and deleting entities. AnCoraPipe also includes a generic search tool that uses XPath expressions, and supports exporting the data into different formats for analysis such as Excel and CSV.

[49] `http://mmax2.sourceforge.net/`

[50] `http://clic.ub.edu/ancorapipe/`

Summary

All the annotation tools discussed here (and others) offer the flexibility for adapting to specific annotation goals, as well as the required interaction for making annotation efficient, including visual display of markable chains (all tools offer a list view for markable chains that allows grasping quickly all the markables from a set) and a search function for text (which is especially helpful for name coreference where mention strings are similar or identical, but the mentions in a chain are far apart). Since they are available as freeware or open source and run on all major platforms, they should definitely be taken into consideration before taking on the risk of developing another annotation tool or using a simple XML or plaintext editor.

4.4.2 Markup and Coding Scheme

Markup

Most corpora with anaphoric information are stored in XML format.[51] The corpora in XML format are generally stored using a standoff representation –either character-based, as in the corpora created using ATLAS-based tools like CALLISTO, or token standoff, as in the corpora created using MMAX2.

As discussed above, for languages with phonetically unrealized anaphoric expressions (**zeros**) like Italian and Japanese, if argument structure is being annotated at the same time and appropriate annotation tools are available, the corpus creators should consider using the argument structure annotation as the base level, as done, e.g., in the Prague Dependency Treebank or ANCORA.

Coding scheme

The most common options for anaphoric annotation have been discussed in Section 4.2. There are two basic options in terms of mention selection: either annotating only the entities that are most relevant for a given domain, ACE style, or annotating all NPs, as done in most other corpora. The MUC7 guidelines [28] still provide a very useful analysis of potential difficulties in mention identification. Most corpora require coders to mark the entire NP boundary.

In terms of anaphoric relations, most corpora focus only on NP anaphora with antecedents introduced by NPs; most coding guidelines provide detailed examples for this type of annotation. Guidelines for bridging were produced by, e.g., DRAMA, GNOME,[52] ARRAU, and DEDE [17].

[51] A notable exception is the ONTONOTES corpus, where all semantic levels are stored in a unified format in a database [64].

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

Markable attributes useful for anaphora include grammatical function, agreement features, and semantic features, ontological category first of all. Most modern anaphoric annotations, and most notably, ANCORA, ONTONOTES and the Prague Dependency Treebank, are carried out in combination with the annotation of other levels, which provide some of this additional information about markables. Virtually all annotation efforts rely at least on (semi-)automatic constituency or dependency annotation. Including named entity type information is also a very good idea, especially if automatic tools to do so are available for the particular language and domain. Anaphora annotation projects that have to create all information by hand could look at the GNOME guidelines for suggestions.

Agreement

The design of the coding scheme should be informed by awareness of what can be reliably annotated. While some of the initial efforts, such as MUC, reported agreement scores,[53] many of the more recent ones–most notably, the ACE campaign– do not [3] . In-depth studies of agreement on anaphoric annotations have been carried out by [59] and as part of the development of the GNOME and ARRAU [53] corpora. The results suggest that reasonable agreement can be obtained on the distinction between discourse-old and discourse-new, but that annotating bridging reference requires identifying very clearly the subset of bridging relations of interest. Any attempt at marking more complex types of anaphoric information should be accompanied by a study of the agreement between annotators.

The GNOME annotation effort also involved an extensive evaluation of the reliability of other types of information (grammatical function, agreement, semantic features, etc.) [51].

4.4.3 Annotation Procedure

(Semi-)Automatic Steps

Carrying out as much of the work automatically is essential to create a resource of adequate size given the constraints most efforts work under. The aspect of the process that can be automated to a greater extent is the identification of markables,[54] but the accuracy of parsers still typically requires that coders be able to correct markable boundaries by hand.

[53] In MUC and other projects, the MUC scoring metric was used (see Chapter 6). The MUC-6 annotators reached an agreement level of F_1=0.83 [29], comparable with later efforts such as the German TüBa-D/Z corpus (F_1=0.83, [73]), or the Dutch COREA corpus (F_1=0.76, [26]), which relied on more refined annotation guidelines.

[54] Researchers working on languages for which not even chunkers exist need to be aware that the corpora they create will probably only be usable for linguistic studies.

Named entity taggers of reasonable quality also exist for many languages, at least for unrestricted domains. Last but not least, the accuracy of dependency parsing is now such that grammatical function identification can also be by and large carried out automatically.

Guidelines

Not all annotation projects produce written guidelines, but experience suggests it is very useful to do so both to carry out agreement studies and to help coders. Most large scale annotation efforts have provided useful examples that could be adapted.

Multiple Coding and Checking

Checking the output from coders is essential for quality checking. In projects with substantial financial support like ONTONOTES, all documents are coded twice and the annotations reconciled. This is unlikely to be possible for most projects, but we would recommend that the researchers leading the effort check at least 10% of the annotation produced by their coders, and have at least 10% of the documents doubly coded.

4.5 Conclusions

The availability of resources for studying anaphora resolution has greatly improved in recent years, to the extent that researchers interested in the development of computational models of anaphora resolution have now resources comparable to those available to the developers of parsers and predicate argument structure analyzers, and not just for English but also for a variety of other languages including at least Arabic, Bengali, Chinese, Catalan, Czech, Danish, Dutch, French, German, Hindi, Italian, Japanese, Spanish, and Tamil. This effort, however, has also revealed that many aspects of anaphora are still poorly understood from a theoretical perspective, and that the situation is not so good for genres other than news.

Acknowledgements This work was supported in part by a PhD studentship offered by Cogito / Expert Systems (Kepa Rodriguez), in part by the LIVEMEMORIES project (Poesio), and in part by the SENSEI project (Poesio).

References

[1] ACE: Annotation guidelines for entity detection and tracking (edt) (2004). Version 4.2.6

[2] Aone, C., Bennett, S.: Evaluating automated and manual acquisition of anaphora resolution strategies. In: Proc. ACL 1995 (1995)

[3] Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4), 555–596 (2008). An early version of this paper has been circulating since 2005 as "Kappa^3 = Alpha (or Beta)". This version is still available from the ARRAU website.

[4] Bard, E.G., Anderson, A.H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., Newlands, A.: Controlling the intelligibility of referring expressions. Journal of Memory and Language **42**, 1–22 (2000)

[5] Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., Liberman, M.: Atlas: A flexible and extensible architecture for linguistic annotation (2000). URL `http://arxiv.org/abs/cs/0007022`

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

[7] Bruneseaux, F., Romary, L.: Codage des références et coréférences dans le dialogues homme-machine. In: Proc. of ACH-ALLC. Kingston (1997)

[8] Byron, D.: Resolving pronominal references to abstract entities. In: Proc. of the ACL, pp. 80–87 (2002)

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

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

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

[12] Cheng, H.: Modelling aggregation motivated interactions in descriptive text generation. Ph.D. thesis, Division of Informatics, the University of Edinburgh, Edinburgh (2001)

[13] Chinchor, N.A.: Overview of MUC-7/MET-2. In: Proceedings of the Seventh Message Understanding Conference (MUC-7) (1998)

[14] Collovini, S., Carbonel, T., Thielsen Fuchs, J., Coelho, J.C., Rino, L., Vieira, R.: Summit: Um corpus anotado com informa cões discursivas visando à sumariza cão automática. In: 52nd Workshop em Tecnologia da Informa cão e da Linguagem Humana (TIL'2007). Rio de Janeiro (2007)

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

[16] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassell, S., Weischedel, R.: The automatic content extraction (ACE) program–tasks, data, and evaluation. In: Proc. of LREC (2000)

[17] Gardent, C., Manuélian, H.: Création d'un corpus annoté pour le traitement des déscriptions d éfinies. Traitement Automatique des Langues **46**(1), 115–140 (2005). URL http://www.loria.fr/~gardent/publis/tal2005.pdf

[18] Ge, N., Hale, J., Charniak, E.: A statistical approach to anaphora resolution. In: Proc. WVLC/EMNLP 1998 (1998)

[19] Grishman, R.: Coreference task definition. Tech. rep., NYU (1995). URL http://www.cs.nyu.edu/cs/faculty/grishman/COtask21.book_1.html

[20] Grishman, R.: Named entity task definition. Tech. rep., NYU (1995). URL http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book_1.html

[21] Grishman, R., Sundheim, B.: Design of the MUC-6 evalutation. In: Proceedings of the Sixth Message Understanding Conference (MUC-6) (1995)

[22] Grishman, R., Sundheim, B.: Message understanding conference-6: a brief history. In: Proceedings of the 16th COLING, COLING '96, pp. 466–471. Association for Computational Linguistics, Stroudsburg, PA, USA (1996). DOI http://dx.doi.org/10.3115/992628.992709. URL http://dx.doi.org/10.3115/992628.992709

[23] Hajič, J., Böhmová, A., Hajičová, E., Vidová-Hladká, B.: The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: A. Abeillé (ed.) Treebanks: Building and Using Parsed Corpora, pp. 103–127. Amsterdam:Kluwer (2000)

[24] Hasler, L., Orasan, C., Naumann, K.: NPs for events: Experiments in coreference annotation. In: LREC 2006 (2006)

[25] Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.M., Van Der Vloet, J., Verschelde, J.L.: A coreference corpus and resolution system for dutch. In: Proc. LREC (2008)

[26] Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.M., Vloet, J.V.D., Verschelde, J.L.: A coreference corpus and resolution system for dutch. In: LREC 2008 (2008)

[27] Hinrichs, E., Kübler, S., Naumann, K.: 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)

[28] Hirschman, L.: MUC-7 coreference task definition, version 3.0. In: N. Chinchor (ed.) In Proc. of the 7th Message Understanding Conference (1998). Available at http://www.muc.saic.com/proceedings/muc_7_toc.html

[29] Hirschman, L., Robinson, P., Burger, J., Vilain, M.: Automating coreference: The role of automated training data. In: Proc. of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing (1997). URL http://arxiv.org/pdf/cmp-lg/9803001

[30] Ide, N.: Corpus encoding standard: Sgml guidelines for encoding linguistic corpora. In: Proc. of LREC. Granada (1998)

[31] Iida, R., Komachi, M., Inui, K., Matsumoto, Y.: Annotating a Japanese text corpus with predicate-argument and coreference relations. In: Proceeding of the ACL Linguistic Annotation Workshop (LAW), pp. 132–139 (2007)

[32] Iida, R., Poesio, M.: A cross-lingual ilp solution to zero anaphora resolution. In: Proc. of ACL. ACL, Boulder, Colorado (2011)

[33] Isard, A.: An XML architecture for the HCRC map task corpus. In: P. Kühnlein, H. Rieser, H. Zeevat (eds.) Proc. of BI-DIALOG (2001)

[34] Kabadjov, M.A.: Task-oriented evaluation of anaphora resolution. Ph.D. thesis, University of Essex, Dept. of Computing and Electronic Systems, Colchester, UK (2007)

[35] Karamanis, N.: Entity coherence for descriptive text structuring. Ph.D. thesis, University of Edinburgh, Informatics (2003)

[36] Klein, M., Bernsen, N.O., Davies, S., Dybkjaer, L., Garrido, J., Kasch, H., Mengel, A., Pirelli, V., Poesio, M., Quazza, S., Soria, C.: Supported coding schemes. Deliverable 1.1, The MATE Consortium (1998). URL \url{mate.nis.sdu.dk/about/deliverables.html}

[37] Krasavina, O., Chiarcos, C.: The potsdam coreference scheme. In: Proc. of the 1st Linguistic Annotation Workshop, pp. 156–163 (2007)

[38] Kučová, L., Hajičová, E.: Coreferential relations in the prague dependency treebank. In: Proc. of DAARC, pp. 94–102 (2004)

[39] LDC: ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, version 5.6.1 (2004)

[40] Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Lenzi, V.B., Sprugnoli, R.: I-cab: the italian content annotation bank. In: LREC 2006 (2006)

[41] McCarthy, J.F., Lehnert, W.G.: Using decision trees for coreference resolution. In: Proc. IJCAI 1995 (1995)

[42] McKelvie, D., Isard, A., Mengel, A., Moeller, M.B., Grosse, M., Klein, M.: The MATE workbench - an annotation tool for XML corpora. Speech Communication **33**(1-2), 97–112 (2001)

[43] Moser, M., Moore, J.D.: Toward a synthesis of two accounts of discourse structure. Computational Linguistics **22**(3), 409–419 (1996)

[44] Müller, C., Strube, M.: 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, *English Corpus Linguistics*, vol. 3, pp. 197–214. Peter Lang (2006)

[45] Müller, M.C.: Fully automatic resolution of it, this and that in unrestricted multy-party dialog. Ph.D. thesis, Universität Tübingen (2008)

[46] Nguyen, N.L.T., Kim, J.D., Tsujii, J.: Challenges in pronoun resolution system for biomedical text. In: Proc. of LREC (2008)

[47] Orasan, C.: Palinka: a highly customizable tool for discourse annotation. In: Proceedings of the 4th SIGdial Workshop on Discourse and Dialog (2003)

[48] Passonneau, R.J.: Instructions for applying discourse reference annotation for multiple applications (DRAMA) (1997). Unpublished manuscript.

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

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

[51] Poesio, M.: The MATE/GNOME scheme for anaphoric annotation, revisited. In: Proc. of SIGDIAL. Boston (2004)

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

[53] Poesio, M., Artstein, R.: Anaphoric annotation in the arrau corpus. In: Proc. of LREC. Marrakesh (2008)

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

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

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

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

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

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

[60] Postal, P.M.: Anaphoric islands. In: R.I.B. et al. (ed.) Papers from the Fifth Regional Meeting of the Chicago Linguistic Society, pp. 205–235. University of Chicago (1969)

[61] Pradhan, S., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R., Xue, N.: 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)

[62] Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., Zhang, Y.: Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In: Joint Conference on EMNLP and CoNLL - Shared Task, pp. 1–40. Associa-